VRIJE
UNIVERSITEIT
BRUSSEL

Master thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in Applied Sciences and Engineering: Computer Science

MUSIC ANALYSIS USING SPECTRAL KNOWLEDGE REPRESENTATION AND REASONING

Density-based Clustering and Representation of Perceived Structure in Audio Signals

Anastasia Nastya Krouglova

2022-2023

Promotor(s): Prof. Dr. Dr. Geraint Wiggins

Advisor(s): Dr. Nicholas Harley, Steven T. Homer

Science and Bio-Engineering Sciences

VRIJE
UNIVERSITEIT
BRUSSEL

Proefschrift ingediend met het oog op het behalen van de graad van Master of Science in Applied Sciences and Engineering: Computer Science

MUZIEKANALYSE DOOR SPECTRALE KENNISREPRESENTATIE EN REDENERING

Dichtheidsgebaseerde clustering en representatie van verkregen structuren in audio signalen

Anastasia Nastya Krouglova

2022-2023

Promotor(s): Prof. Dr. Dr. Geraint Wiggins

Advisor(s): Dr. Nicholas Harley, Steven T. Homer

Wetenschappen en Bio-ingenieurswetenschappen

Abstract

The extraction and formation of musical structures through the analysis of complex auditory scenes is a challenging task in signal processing and machine learning. Musical analysis includes multiple open subtasks to be resolved, such as multi-pitch estimation, musical note tracking and multi-pitch streaming. The main goal of this thesis is to create a framework for the multipurpose description and evaluation of music, allowing inference from different subtasks and a general improvement in the learnability of machine learning models. This was achieved by investigating into the implementation of a coherent structure between a spectral analysis of resonances and a type-based knowledge representation in the musical domain, forming an analogy to the perception, cognition and knowledge representation of human intelligence. We created pitch-based hierarchies formed through density-based clustering techniques in our self-defined hierarchical structure for the definition of musical objects perceived from audio signals. Our multipurpose framework for musical analysis has a methodological contribution to various practical applications due to its precision and ability to deal with overlapping sound events, which is one of the key challenges in music signal processing. Approaching this problem through a cognitive perspective has a significant impact on the way machine learning is performed nowadays, due to the possibility of model inference for various subtasks in machine learning. Our software also contributes to long-term prospective of explainable modelling and can be used in other early related fields, including speech recognition. Overall, this thesis bridges the gap between human intelligence and machine learning through the development of a framework for knowledge representation and the recognition of musical objects in a resonance spectrum.

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my supervisors Geraint Wiggins, Steven Homer, and Nicholas Harley for their extensive guidance and unwavering support. Geraint, thank you for proposing this subject and guiding me throughout this incredible learning journey. Your constant encouragement to ask questions, even during your busy schedule, has been invaluable. Steven, thank you for introducing me to the intricate world of resonances, and Nick, I appreciate your extensive explanations of the CHAKRA framework and the insights you gave during our whiteboarding sessions to solve the problems we faced. I have been fortunate to experience an exceptionally supportive and positive research environment during the process of writing this master's thesis, for which I am immensely grateful.

I am also indebted to my dear friend Ru Chen at Chalmers University and my colleagues at ETH Zürich for providing an immersive positive intellectual influence on me and my attitude towards research. Additionally, James, I would like to thank you for always standing beside me, no matter what happens. Thank you for always carefully listening to my enthusiastic raves about this thesis, the "seally" jokes for cheering me up and your thoughtful advice and feedback.

Lastly, I would like to express my profound gratitude to Gilles Castel. Though he is no longer with us, his lasting influence on me and other students will endure. Throughout the past years, he has been a tremendous source of inspiration, and I wish to thank him for that. The layout of this thesis includes a substantial portion of his thesis layout, serving as a tribute to his memory.

Anastasia N. Krouglova
19 May 2023

Contents

Abstract i

Acknowledgement ii

0 Introduction 2

I Background on Musical Audio Analysis 5

1 Fourier Analysis of Auditory Signals 6
- 1.1 Theoretical background . . . 6
  - 1.1.1 Fourier Series . . . 6
  - 1.1.2 Continuous Fourier Transform . . . 8
  - 1.1.3 Discrete Fourier Transform . . . 8
- 1.2 Practical Background . . . 9
  - 1.2.1 Fast Fourier Transform . . . 9
  - 1.2.2 Short-Time Fourier Transform . . . 9
- 1.3 Summary . . . 9

2 Hilbert Spaces 10
- 2.1 Completeness, Integration and Infinity in Hilbert Spaces . . . 11
- 2.2 Bases in a Hilbert Space . . . 11
  - 2.2.1 Hilbert Space with Orthogonal Bases . . . 11
- 2.3 Real-valued signals in a Hilbert space . . . 13
- 2.4 Summary . . . 13

3 Psychoacoustics 14
- 3.1 Sensing and Perceiving . . . 14
- 3.2 Perception of Sound . . . 14
  - 3.2.1 Missing Fundamental . . . 14
  - 3.2.2 Combination Tone . . . 15
- 3.3 The Fundamental and (Non)harmonic Overtones . . . 15
  - 3.3.1 Structualism . . . 16
- 3.4 Summary . . . 16

II A Cognitive Approach to Musical Audio Analysis 17

4 The Mechanisms of Hearing 18

4.1 Cochlea as a Fourier Analyzer . . . 18
4.2 Modelling the Mechanisms of Hearing . . . 18
4.2.1 The Fourier Uncertainty Principle and Hearing . . . 19
4.3 Summary . . . 19

5 Discrete Resonance Spectrum 20
5.1 Discrete Resonances in Time Domain . . . 20
5.2 Discrete Resonances in Frequency Domain . . . 21
5.3 The Fast Padé Transform (FPT) . . . 22
5.3.1 Preliminary Knowledge . . . 22
5.3.2 Mathematical Derivation . . . 23
5.4 Towards higher Precision with Non-Orthogonal Bases in a Hilbert Space . . . 25
5.5 Extracting Attributes from the Discrete Resonance Spectrum 25
5.5.1 Dynamic Resonances . . . 25
5.6 Summary . . . 26

6 Cluster Analysis of Resonances 27
6.1 Density-based Cluster Algorithms . . . 27
6.2 DBSCAN clustering algorithm . . . 28
6.2.1 Complexity and Data structure . . . 28
6.2.2 DMBSDSCAN . . . 28
6.3 Summary . . . 28

7 Type-Based Knowledge Representation 29
7.1 The Three Perspectives on Computation . . . 29
7.1.1 Type Theory . . . 29
7.1.2 Category Theory . . . 30
7.1.3 Typed logic . . . 30
7.2 The CHAKRA System . . . 31
7.2.1 The CHARM System . . . 31
7.3 Summary . . . 32

III Software Architecture 33

8 Knowledge Representation Applied to Audio Files 34
8.1 Specification of the CHAKRA abstraction . . . 34
8.1.1 Constituent . . . 34
8.1.2 Identifier . . . 35
8.1.3 Hierarchy . . . 35
8.1.4 Operations . . . 36
8.2 Engineering Considerations . . . 36
8.3 Summary . . . 37

9 Note-level description 38
9.1 Fundamental Frequency Detection . . . 38
9.2 Clustering . . . 39
9.2.1 Parameter estimation . . . 39
9.2.2 Clustering Performance Evaluation . . . 40
9.2.3 Clusters of noise . . . 41

9.3 Attributing a Note Constituent 41
9.4 Towards a Frame-level description 42
9.4.1 Attributing the Harmonic Constituent 42

Conclusion **43**

Further Work **44**

Nomenclature **45**

CHAPTER ZERO

Introduction

Human intelligence has the fascinating capability of recognizing musical instruments, rhythm, pitches and other structures in music. It can focus the auditory attention on a particular task and filter out a range of other stimuli. This part of our intelligence involves four key components: *perception*, *cognition*, *knowledge representation* and *inference* (Benetos et al., 2019). *Perception* refers to the ability of analyzing input audio, and can significantly vary based on what it has learned over different occasions, unlike sensation, which remains relatively constant over time (O'Brien, 2023). *Cognition* is the ability of recognizing musical objects, and can be seen as a resolution of a system (Oppenheim and Magnasco, 2013). *Knowledge representation* is the formation of musical structures from the obtained cognition, and *inference* refers to the ability of learning from musical structures.

However, extracting musical structures with *digital* equipment is a challenging task in signal processing and machine learning due to the complicated nature of music. Interference between sound waves, noise (unwanted sound), and reverberation can result in a complex cocktail of stimuli. Recent approaches in the literature have made several attempts to extract musical structures, including non-negative matrix factorization (NMF) (López-Serrano et al., 2019; Holzapfel and Stylianou, 2008), Bayesian approaches (Donnelly, 2012; Temperley, 2004) and neural networks (Draguns et al., 2021; Sleep, 2017). Nonetheless, many aspects of musical analysis are still considered as open problems in the literature.

Although the power of (deep) neural networks should not be questioned, keeping the progress in the past years in mind, they do not *really* simulate the functioning of the brain. Therefore, Wiggins (2020) proposed to construct an explanatory model with a level of abstraction that describes the hypothetical mechanisms (and not necessarily the effect) of the brain treating auditive signals. Since music analysis has a close relation to other signal processing problems, including speech recognition, the acquired solutions and insights throughout the process can help to solve similar problems from a cognitive perspective.

To advance towards the human-like **perception**, Homer, Harley, and Wiggins (2023) proposed an idealized model of auditory receptive fields. The input information for this model are the so-called *discrete resonances*, which are

directly inspired from cochlear mechanics: the vibration of the eardrum in the outer ear can be described as a damped harmonic oscillator (Chung, Pettigrew, and Anson, 1981), from which the middle ear translates pressure waves to mechanical energy. This awakens a resonance of the basilar membrane in the inner ear that can be described as a mechanical resonance. **Knowledge representation**, on the other hand, has been modeled by Harley (2020). He developed a Common Hierarchical Abstract Knowledge Representation for Anything (Harley, 2022a, CHAKRA). This type-based framework for knowledge representation supporting the idea of life-long learning and can be used for the representation of musical knowledge (Wiggins, Harris, and Smaill, 1989).

The goal of this thesis is to tie the model of perception and knowledge representation together through a model for **cognition**. Although audio is in general expressive complete, containing a lot of information in a signal (unlike MIDI files), it lacks in structural generality, making it not evident to extract structures from audio files (Wiggins, 2020). We apply a clustering-based approach for the extraction of musical objects in a discrete resonance spectra and create a type-based knowledge representation specifically for audio signals. Our aim is to create a bidirectional system which scores well on both expressive completeness (e.g., audio waveforms) and structural generality (e.g., sheet music) (Wiggins, Harris, and Smaill, 1989; Collins, 2018). Finally, a demonstration will be given of the possibilities with the developed software by showcasing the extraction of pitch and overtones and how they are structured in a knowledge representation.

In the first part of this thesis, the reader will be guided through a few essential concepts in signal processing and psychoacoustics. These concepts play a crucial role in the understanding of our cognitive model and approach. Therefore, **Chapter 1** gives an introduction to the different categories in Fourier analysis and discusses the decomposition of a signal into complex exponentials applies to audio signals. **Chapter 2** delves deeper into this subject and provides a definition of the Fourier Transform within the context of a L^2 Hilbert Space. In **Chapter 3**, several aspects of psychoacoustics are briefly discussed, including the excessive explanation of our terminology (e.g., perception, constituent elements).

The second part provides the theoretical background behind our state-of-the-art cognitive model. **Chapter 4** lays out the fundamentals of cochlear mechanics and discusses several observations about the transfer of perceptual information between the basilar membrane in the cochlea and auditory cortex. Using the presumption about cochlear mechanics, **Chapter 5** introduces the intricate world of discrete resonances. We also define resonances in a Hilbert space, but this time, the basis in a Hilbert Space is not necessarily orthogonal anymore. This entails interesting features for musical analysis, including precision. This level of precision will play a crucial role in grouping the resonances. Therefore, **Chapter 6** describes the background behind a density-based clustering algorithm (Ester et al., 1996, DBSCAN). This method will be applied to the resonance spectra for the generation of a stronger structural generality in audio files. The acquired knowledge will be structured in a type-based knowledge representation. **Chapter 7** finally

presents the theory behind the framework (Harley, 2020, CHAKRA).

The last part outlines my own contribution to this research field. **Chapter 8** describes our particular implementation of the CHAKRA framework, and **Chapter 9** discusses the implementation of our model that simulates cognition in human intelligence. We simulate the connections made in the brain to perceive music through a density-based clustering approach. We use the DBSCAN algorithm, which clusters data as a human would do, and compare the performance of two hyperparameter estimations, namely the silhouette score and the *kneedle* method. Finally, we demonstrate the performance of the cognition of pitch and overtones.

I

Background on Musical Audio Analysis

In order to understand our cognitive approach to musical audio analysis, we first need to discuss the main parts of the Fourier Analysis and its classic definition in a Hilbert space with orthogonal bases. However, in the next part, this concept of orthogonality of bases for a decomposition into resonances will not hold anymore. Finally, a concise yet essential introduction to psychoacoustics will be given, which plays an important role in the understanding of the cognitive clusters, formed from resonance information.

CHAPTER ONE

Fourier Analysis of Auditory Signals

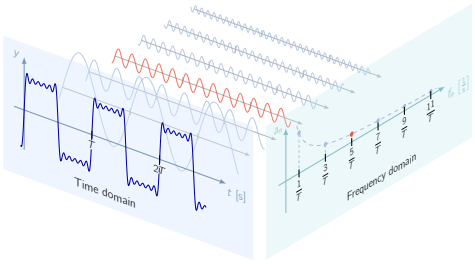

Figure 1.1: A visualization showcasing Fourier analysis, where a square wave in the time domain is decomposed into sinusoidal components. This decomposition allows a transformation into the frequency domain representation and vice versa. Edited from Neutelings (2021a). The modifications made to the figure include coloring and change of axis labels.

[1] The Nyquist sampling theorem states that in order to sample a signal at a certain frequency without significant lost, it should be sampled at twice that frequency. And since the limit of human hearing is approximately 20kHz, it requires a sample rate of 44 kHz.

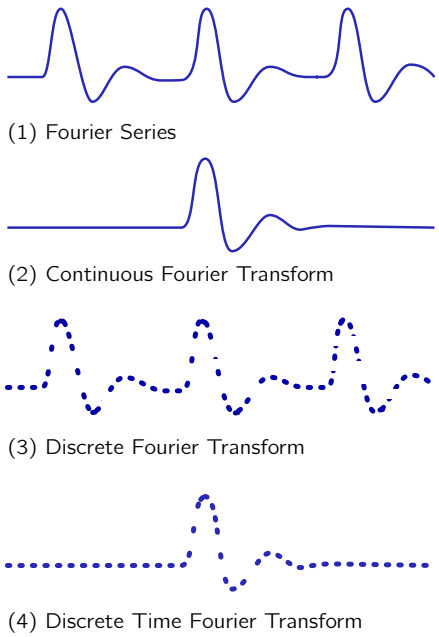

Figure 1.2: An illustration providing the behavioral nature of the four fundamental Fourier Transform Categories, namely Fourier Series, Continuous Fourier Transform, Discrete Fourier Transform, and Discrete Time Fourier Transform.

An auditory signal, such as a tone, sound, or spoken message, is a function that carries information. Machines are not able to store values up to infinity, and thus audio files are a discretization of the actual signal, sampled at a certain rate. The specifics of sampling rates exceed the scope of this topic, but as a general guideline, recording at 44.1 kHz typically yields high-quality results[1].

For many applications, such as the analysis of pitch intensities, it is interesting to transform the time-domain signal into a frequency-domain representation with the Fourier transform (Figure 1.1). The Fourier transform is a general term that can be split into four categories, depending on a combination of four characteristics of the input signal: periodic or aperiodic and continuous or discrete. For each category, a specific name is given to refer to a type of signal (Smith, 1997):

1. Periodic-Continuous: "Fourier Series"
2. Aperiodic-Continuous: "(Continuous) Fourier Transform"
3. Periodic-Discrete: "Discrete Fourier Transform"
4. Aperiodic-Discrete: "Discrete Time Fourier Transform"

Sometimes, the Fourier transform is regarded as an extension of Fourier series because it can handle both periodic and aperiodic signals. We will begin by introducing some fundamental theoretical concepts using Fourier series and then delve into the Continuous and Discrete Fourier Transform. The discrete-time Fourier transform represents a fourth category of the Fourier Transform; however, due to its limited relevance to the subsequent discussions, we will not explore its intricacies in detail. Lastly, the Fast Fourier Transform and Short-Time Fourier Transform will be introduced for a more practical background.

1.1 Theoretical background

1.1.1 Fourier Series

Consider an ideal string vibrating solely at the fundamental harmonic A4. This repeating pattern corresponds to a single sinusoidal wave with a frequency of

440 Hz (cycles per second), as roughly depicted in Figure 1.3(a). However, when an instrument plays the A4 note, the resulting sound differs from the sound of an ideal string. This variation is caused by differences in the combination of multiple harmonics or soundwaves, which makes determining the precise real-valued function of a violin or piano a more challenging task, as illustrated in Figures 1.3 and 1.4. Nonetheless, it can be approximated by summing weighted sines and cosines using the Fourier series:

$$f(t) = \frac{a_0}{2} + \sum_{n=-\infty}^{\infty} [A_n \cos(\omega n t) + i A_n \sin(\omega n t)]. \tag{1.1}$$

The sine and cosine waves serve as the fundamental basis functions in the Fourier transform. Frequency ω is a compact representation of the natural frequency $2\pi\phi = \frac{2\pi}{T}$, where T represents the period (the time required for a complete rotation around the unit circle). In this context, $\phi = \frac{1}{T}$. ω can be interpreted as angular speed[2]. Note that the sine and cosine notation can be expressed in exponential form using Euler's formula: $e^{i\omega t} = \cos(\omega t) + i \sin(\omega t)$:

$$f(t) = \sum_{n=-\infty}^{\infty} A_n e^{i\omega n t}. \tag{1.2}$$

The exponential form using Euler's formula is a more convenient way to represent the sine and cosine waves due to its compactness. It can also be visualized effectively in the complex plane, as depicted in Figure 1.5.

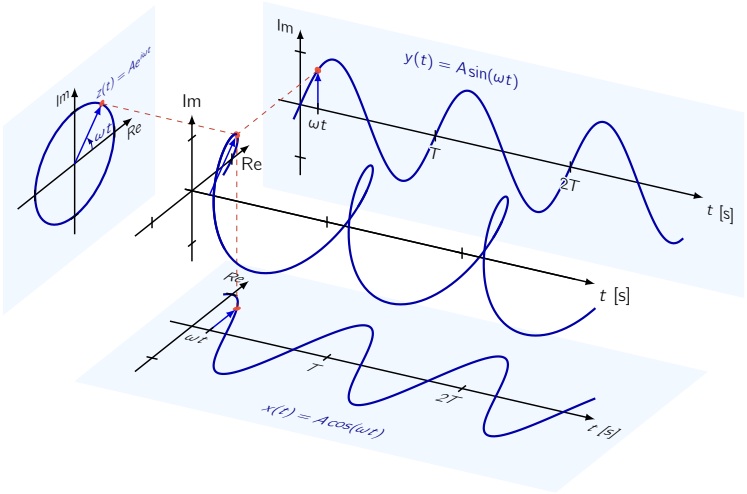

Figure 1.5: A simple harmonic oscillator projected in 3D with sinusoidal waves in the Cartesian plane and their compact representation in the complex plane. The real part and the imaginary part of this analytic signal are related through the Hilbert transform. Edited from Neutelings (2021b).

Note that in the description above, the notion of phase shift ϕ was omitted for simplicity. In reality, digital encoding of audio signals can introduce (unwanted) phase shifts ϕ (Figure 1.6):

$$f(t) = \sum_{n=-\infty}^{\infty} A_n e^{i(\omega n t + \phi)}. \tag{1.3}$$

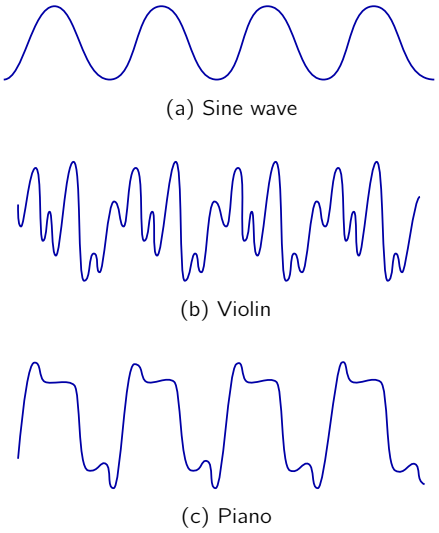

Figure 1.3: Three different time-domain signals of similar notes played with a pure tone, violin and piano.

[2] The choice of notation (ω or $2\pi\phi$) depends on the specific field of study, here, ω is measured in radians per second, while $2\pi\phi$ is measured in cycles per second.

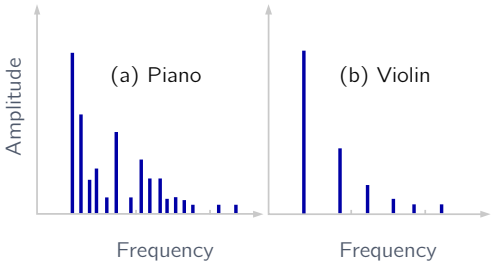

Figure 1.4: An example of the harmonics of similar notes in the frequency-domain played with piano and violin (Arvin and Doraisamy, 2009).

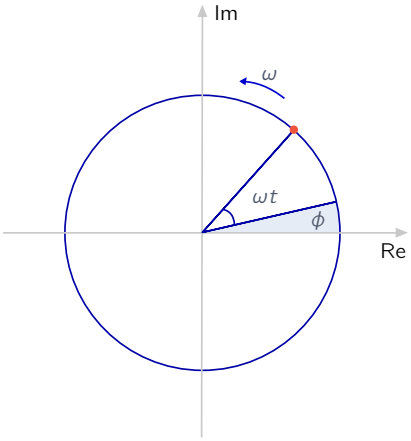

Figure 1.6: Simple harmonic oscillator with phase shift ϕ in the polar plane.

1.1.2 Continuous Fourier Transform

Fourier series is sufficient to describe periodic signals, but for more complicated aperiodic functions, the (continuous) Fourier transform is required. This transform allows the conversion of a continuous time-domain signal into the frequency domain through an invertible linear transformation, defined as follows:

$$\mathcal{F}[f(t)] = \hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{i\omega t} dt \qquad t, \omega \in \mathbb{R}, \tag{1.4}$$

and the inverse Fourier transform as

$$\mathcal{F}^{-1}[\hat{f}(\omega)] = f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{-i\omega t} d\omega \qquad t, \omega \in \mathbb{R}. \tag{1.5}$$

Notice the similarity between the Fourier transform and its inverse. In this case, a symmetric notation is used for both transforms. However, in other literature, the term $\frac{1}{\sqrt{2\pi}}$ may be excluded from the Fourier transform and only included in the inverse Fourier transform as $\frac{1}{2\pi}$ (Baraniuk, 2020). Here, we evenly distribute this term across both transforms, acting as a normalization factor.

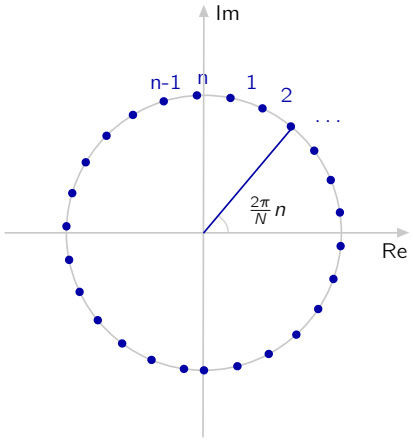

Figure 1.7: Polar representation of phasor in a discrete unit circle.

1.1.3 Discrete Fourier Transform

Digital computers can only process discrete and finite data, and since audio data is aperiodic and discrete, the Discrete Time Fourier Transform (DTFT) is intuitively applicable. However, in practice, the "Discrete Fourier Transform" (DFT) is mostly used, often extended with various techniques to enhance its performance, such as improving speed. In this equation, the continuous-time signal is represented by discrete values:

$$\mathcal{F}[x[n]] = X[k] = \sum_{n=0}^{N-1} x[n] e^{-i\frac{2\pi}{N}kn} \qquad k = 0, 1, \ldots N-1, \tag{1.6}$$

and the inverse Discrete Fourier transform:

$$\mathcal{F}^{-1}[X[k]] = x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{i\frac{2\pi}{N}kn} \qquad k = 0, 1, \ldots N-1. \tag{1.7}$$

N represents the number of samples, often chosen as a power of two, indicating that only specific points n can be captured by the Discrete Fourier Transform (Figure 1.7).

In the frequency domain, the sampled frequency k corresponds to the number of rotations around the unit circle in N points. It is worth noting that in this notation, we avoid using the abstraction of ω to highlight the discrete nature of the formulation.

1.2 Practical Background

The Fast Fourier Transform (FFT) and Short-Time Fourier Transform (STFT) are two commonly used techniques for the calculation of a signal's DFT. Due to significant speed enhancements of the implemented algorithms, the FFT and STFT became valuable tools across a range of signal processing applications.

1.2.1 Fast Fourier Transform

The FFT is an efficient algorithm used to compute the Discrete Fourier Transform (DFT) of a signal. The main practical problem with the DFT (Discrete Fourier Transform), is that it requires large matrix multiplications and summations over all its elements, which are computationally expensive operations. A viable solution is the Cooley-Tukey algorithm, which is one of the most common algorithms used in the FFT. It reduces the computational time of the DFT from $\mathcal{O}(n^2)$ to $\mathcal{O}(n\log(n))$ by dividing its elements recursively into two groups based on parity (odd or even indices) until the elements are computationally manageable with the DFT (Cooley and Tukey, 1965).

An application of the Fast Fourier Transform is the estimation of the power spectral density function, as shown in Figure 1.8. The power spectral density (PSD) represents the distribution of signal power across different frequencies.

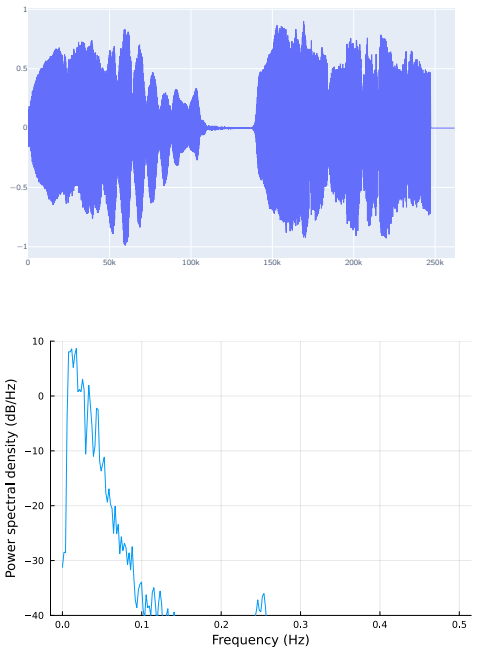

Figure 1.8: The power spectral density function estimated through the Fast Fourier Transform by calculating the squared magnitude of the Fourier coefficients.

1.2.2 Short-Time Fourier Transform

The Short-Time Fourier Transform (STFT) is performed by calculating the FFT over shorter time segments that might overlap. These time segments, referred to as the window size, determine the resolution of the STFT. Similarly, the frequency is divided into frequency bins. It is crucial to note that the window size and frequency bin count collectively determine the precision of the STFT. In Chapter 4, we will discuss the Fourier uncertainty principle in more detail, which sets a limit on the precision of this transform. Consequently, when applying the STFT, there is a trade-off between time and frequency that needs to be considered. For instance, a spectrogram (Figure 1.9) illustrates the impact of this uncertainty bound on the precision of the transform.

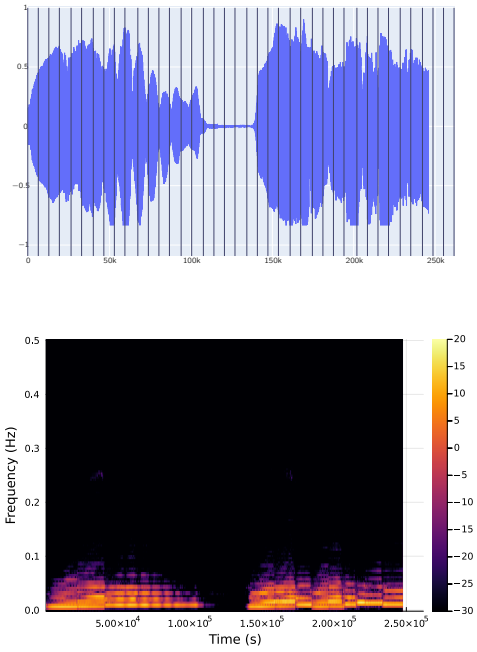

Figure 1.9: The spectrogram of a fragment from Debussy's Syrinx, obtained by calculating the squared magnitude of the signal's power spectral density in each time segment.

1.3 Summary

We discussed four categories of the well-known Fourier Transform and highlighted the Discrete Fourier Transform, which is used the most in practice. We also gave a practical introduction to the Fast Fourier Transform and Short Time Fourier Transform. The understanding of the main idea behind the DFT and FFT will play an important role in Chapter 5, with the introduction of the so-called Fast Padé Transform. The next chapter takes a deeper dive into the underlying mathematics behind Fourier Analysis. We will define the Fourier Transform in a L^2 Hilbert Space, which is a complete inner product space.

CHAPTER TWO

Hilbert Spaces

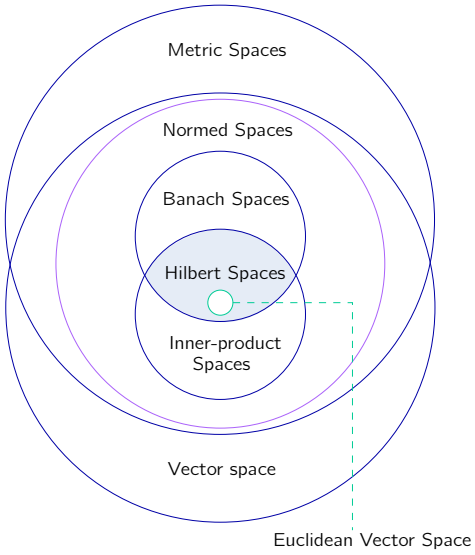

Figure 2.1: Relations among different spaces in functional analysis.

A Hilbert space $\mathcal{H}$ is a complete inner product space, and its theory can be considered as a generalization of the familiar Euclidean vector space (e.g. $\mathbb{R}^n$). It is a vector space where vectors (usually described using finite real or complex numbers), functions, and even more general objects can be represented. It allows us to define us a complete set of basis functions, which are, in case of the Fourier Transform, complex exponentials (e.g., sinusoidals). For this chapter, the crucial idea to understand is that functions can be thought of as vectors with an infinite number of dimensions with certain basis properties.

Theorem 2.1 (Complete normed spaces (Banach Spaces))**.** A normed space $\mathcal{H}$ is called complete if every Cauchy sequence of vectors in $\mathcal{H}$ converges to a vector in $\mathcal{H}$. A complete normed space is called a Banach space (Kennedy and Sadeghi, 2013).

As illustrated in Figure 2.1, all (finite-dimensional) Hilbert spaces are Banach spaces, which means that it is both normed and complete. However, not all Banach spaces are Hilbert spaces. In order for a Banach space to be considered a Hilbert space, the norm (or distance) must be induced by the inner product. The inner product of two signals is a scaled projection, i.e., a scalar which may contain complex values and can be defined as

$$\langle \alpha | \beta \rangle = \alpha_1 \beta_1 + \alpha_2 \beta_2 + ... + \alpha_N \beta_N. \tag{2.1}$$

This means that given an inner product $\langle \cdot \rangle$ in a vector space $\mathcal{H}$, the norm is defined as

$$||f|| = \sqrt{\langle f | f \rangle}. \tag{2.2}$$

Examples of relevant Hilbert Spaces are $\mathbb{C}^n$, the ℓ^2- and L^2-spaces. The ℓ^2 is the space of absolutely square-summable *sequences* and L^2-space of absolutely square-summable *functions*. It should be noted that the bases of a normed space are not necessarily orthonormal. A normed space is a type of vector space that has a norm, which is a mathematical function that gives each vector in the space a non-negative size or magnitude, and satisfies certain properties such as non-negativity, homogeneity, and the

triangle inequality. A second important point to be aware of, is that the concept of completeness is closely linked to the norm. In a normed space, a sequence of vectors should converge, in terms of their norm, to a vector that is within the original space, in order for the space to be considered complete.

2.1 Completeness, Integration and Infinity in Hilbert Spaces

The usage of Hilbert spaces is in general interesting for dealing with infinity, the meaning of Fourier series, and the definition of an inner product in terms of integrals (Kennedy and Sadeghi, 2013). When talking about infinity and Hilbert spaces, one refers mostly to its dimensionality $\dim_F(\mathcal{H})$. The dimension of a Hilbert space is the number of vectors (i.e., cardinality) in its basis. Finite-dimensional Hilbert spaces are defined as complete. This means that they are suitable for including the natural limits of converging vector sequences. In contrast, infinite-dimensional spaces are not necessarily complete, since there might be Cauchy sequences which do not converge. An example of a complete infinite-dimensional space is L_2, the space of square-integrable functions, a real- or complex-valued measurable function for which the integral of the square of the absolute value (i.e., the real axis) is finite:

$$f : \mathbb{R} \to \mathbb{C} \text{ square integrable} \quad \Leftrightarrow \quad \int_{-\infty}^{\infty} |f(x)|^2 \, \mathrm{d}x < \infty. \tag{2.3}$$

In this way, a complete space is defined to work in. Thus, when defining the inner product,

$$\langle f|g\rangle = \int_{-\infty}^{+\infty} f(x)\overline{g(x)} \, \mathrm{d}x \tag{2.4}$$

where f and g are both square integrable functions and $\overline{g(x)}$ is the complex conjugate of $g(x)$.

2.2 Bases in a Hilbert Space

2.2.1 Hilbert Space with Orthogonal Bases

Example: Fourier Series

From the perspective of linear algebra, Fourier series is a decomposition of a periodic function into an infinite sum of (simple) harmonic oscillators in terms of a complete orthonormal sequence $\{\phi_n\}_{n=1}^{\infty}$ in a Hilbert space $\mathcal{H}$ (Russel, 2021; Kennedy and Sadeghi, 2013). It should be noted that the underneath definition assumes a robust orthonormal sequence; however, the essential characteristic of the orthonormal sequence is its property of orthogonality (since an orthonormal sequence is a normalized orthogonal sequence).

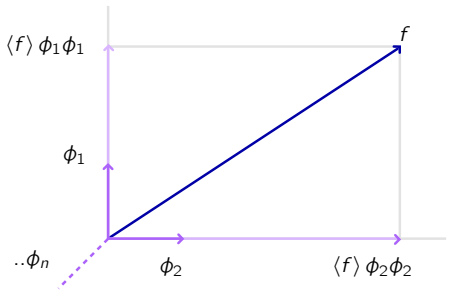

Figure 2.2: Geometrical interpretation of inner products projected along ϕ_n in 2D with orthonormal bases.

Theorem 2.2 (Fourier Series)**.** In a separable Hilbert space, the expansion

$$f = \sum_{n=1}^{\infty} \langle f|\phi_n\rangle \, \phi_n. \tag{2.5}$$

of any $f \in \mathcal{H}$ in terms of a complete orthonormal sequence $\{\phi_n\}_{n=1}^{\infty}$ is called a Fourier series expansion, and the coefficients

$$\langle f|\phi_n\rangle \in \mathbb{C} \tag{2.6}$$

are called the Fourier series coefficients.

Note that we consider here a separable Hilbert space, which means that it only admits a countable orthonormal basis and thus there is a countable dense family of functions.

Furthermore, f can be expressed as a (complex) linear combination of the e_n's, thus the family e_n spans implicitly L^2 in this definition. Geometrically, it simply represents the projection of f along ϕ_n, as illustrated in Figure 2.2.

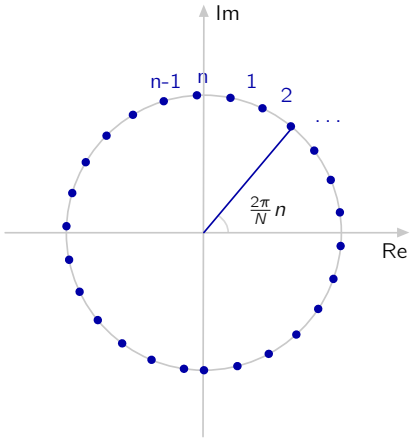

Figure 2.3: A pole-zero diagram representing uniformly-spaced points due to a decomposition into orthogonal bases.

Example: Discrete Fourier Transform

The Discrete Fourier Transform, introduced in Chapter 1, is defined by the discrete orthogonality property of its basis vectors:

$$\sum_{n=0}^{N-1} e^{i\left(\frac{2\pi}{N}\right)nk} e^{-i\left(\frac{2\pi}{N}\right)nl} = \begin{cases} N, k \neq l \\ 0, k = l \end{cases} \tag{2.7}$$

To simplify the expression, we can introduce the variable $\alpha = e^{i\left(\frac{2\pi}{N}\right)n}$, leading to the following formulation:

$$\sum_{n=0}^{N-1} \alpha^{(k-l)} = \begin{cases} N, k \neq l \\ 0, k = l \end{cases} \tag{2.8}$$

Thus, the product of the two exponentials is 0 or N, which corresponds to the same point in the complex plane and so the summation over a period becomes 0.

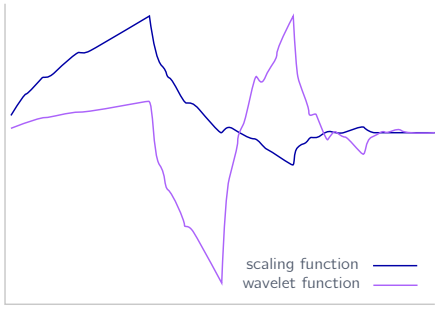

Figure 2.4: An example of a Daubechies 4 tap wavelet (LutzL, 2009).

Example: Daubechies Wavelets

The wavelet revolution in 1986 was started with the creation of the first set of wavelets that were at least as powerful as Fourier components. The technique was published by Pierre Lemarié and Meyer (1986) in *Ondelettes et bases Hilbertiennes*, which literally means "small waves in Hilbert Spaces". Wavelets are a family of differently shaped short-lived oscillations localized in time that can be used to analyze a signal and simultaneously give a solution in time and frequency. They have different characteristics serving for different purposes, such as symmetry, regularity, vanishing moments and orthogonality (Kainulainen and Maercker, 2022). The wavelet transform can be, similarly to the Fourier Transform, categorized as a continuous wavelet transform (CWT)

or discrete wavelet transform (DWT). Within the DWT, a distinction is made between the redundant discrete systems and orthonormal (and others) bases of wavelets (Daubechies, 1992).

Ingrid Daubechies, a famous Belgian physicist and mathematician who obtained a doctoral degree at the VUB in 1980, came up with a family of **orthogonal** wavelets called the *Daubechies wavelets* and characterized by a maximal number of vanishing moments within a specific range. A vanishing moment constrains a wavelet by a polynomial, i.e., a signal with n vanishing moment encodes a polynomial of n coefficients. In practice, improvements in speed for the DWT are also provided with the Fast Wavelet transform. Wavelets can be applied in music for the determination of notes with time and frequency information, by convolving a wavelet with a signal.

2.3 Real-valued signals in a Hilbert space

As previously discussed, a real-valued signal can be expressed as a linear combination of various basis functions, including complex exponentials (Figure 2.5). Notice that digital computers work with sampled functions (Figure 2.6). They are computable finite-dimensional vectors and representable in an n-dimensional Hilbert space (i.e., an inner-product space). We interpret the components of the basis functions e_k and function f as function values:

$$\tilde{f} = \langle e_k | f \rangle = \begin{bmatrix} e_k(x_0) \\ e_k(x_1) \\ . \\ . \\ . \\ e_k(x_L) \end{bmatrix} \times \begin{bmatrix} f(x_0) \\ f(x_1) \\ . \\ . \\ . \\ f(x_L) \end{bmatrix} \tag{2.9}$$

Notice, signal $\tilde{f}$ is a complex-valued function. Since there are theoretically an infinite amount of function values, we write the summation as an integral. To satisfy the mathematical properties of an inner product: $\tilde{f} : \mathbb{R} \to \mathbb{C}$, the first argument in the integral should be the complex conjugate.

$$\tilde{f} = \int_{x_0}^{x_0+L} \overline{e_k(x)} f(x) dx. \tag{2.10}$$

2.4 Summary

We raised our level of sophistication of the analysis of the Fourier transform through the introduction of Hilbert Spaces. We discussed its properties and relation towards Fourier analysis and emphasized the orthogonal basis that are often used in Hilbert Spaces. However, in Chapter 5, the basis will no longer be required to be orthogonal.

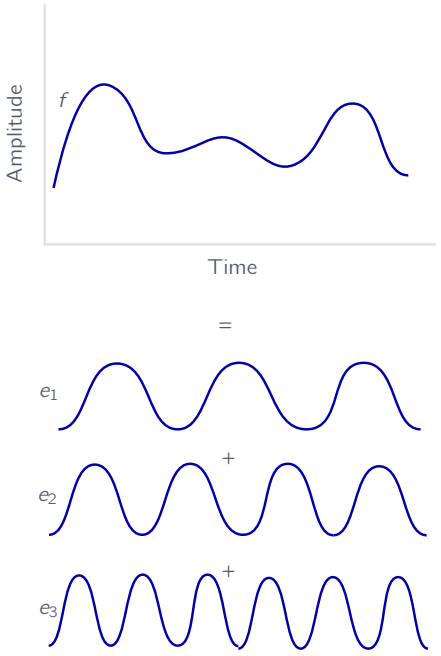

Figure 2.5: A linear combination of the complex exponentials e_1, e_2, e_3 (visualized in Re) approximating function f.

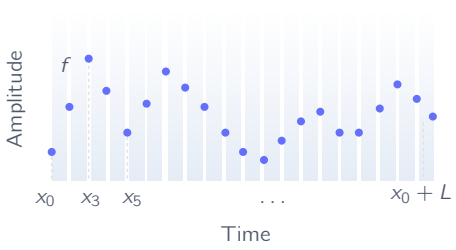

Figure 2.6: A sampled signal f over a period of L in the time domain.

CHAPTER THREE

Psychoacoustics

This chapter will introduce several important concepts in the perception of sound, which will play an important role in the interpretation of resonance spectrograms (Chapter 5) and the creation of hierarchies (Chapter 8) in our knowledge representation. Psychoacoustics is the scientific field that studies the human perception of sound and the psychological responses associated with it. We will discuss the fundamental difference between sensing and perceiving in this chapter and explore how inference influences our perception of sound and can create new (unwanted) tones.

3.1 Sensing and Perceiving

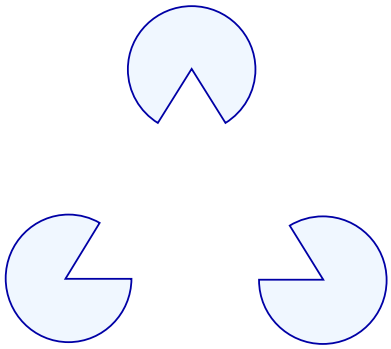

Figure 3.1: A Kanizsa triangle illustrating the difference between sensing and perceiving through the formation of an illusionary triangle from incomplete circles.

Many different descriptions of the distinction between sensing and perceiving were given throughout history. Aristotle described perception as an act of self-consciousness, representing a reflective self-awareness of our perceptual actions (Kosman, 1975). Another commonly accepted explanation is that learning influences perception but not sensation. This means that perception can vary significantly based on what has been learned over different occasions, while sensitivity remains relatively constant over time (unless temporal changes in sensitivity are established) (O'Brien, 2023). Figure 3.1 visually illustrates the distinction between sensing, which involves receiving external information from three circles with triangular gaps, and perception, which involves the inference of additional information from the triangle between the circles with gaps (Kellman and Shipley, 1991).

3.2 Perception of Sound

3.2.1 Missing Fundamental

The phenomenon of the missing fundamental is the perception of the fundamental frequency when it is not physically present in the original sound. This perception occurs because the auditory cortex, the region of the brain responsible for sound processing, interprets repetitive patterns of the overtones that are present (Smith et al., 1978; Zatorre, 2005; Schneider et al., 2005). The model of pitch perception, which will be discussed in Chapter 5, captures this missing fundamental effect (Homer, Harley, and Wiggins, 2023).

3.2.2 Combination Tone

A combination tone refers to the psychoacoustic phenomenon where an additional tone or tones are perceived when two actual musical tones are played simultaneously (Hosch, 2023). There are two types of combination tones: difference tones and summation tones. Difference tones are generated by subtracting the frequency of one tone from the frequency of another tone, as shown in Figure 3.2. On the other hand, summation tones are produced by adding the frequencies of the two tones together.

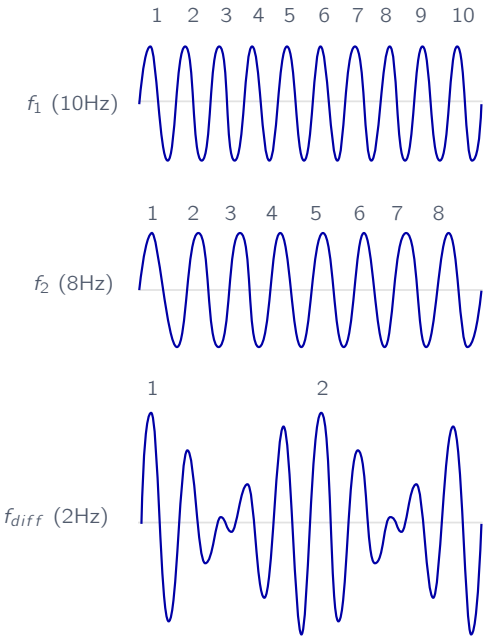

Figure 3.2: Visualization of difference tones, inspired by the book "For the Contemporary Flutist" (Offermans, 2023).

Difference Tone

Difference tones are commonly observed when a harmonic series produces a fifth between notes. This phenomenon can also be heard in a room with sufficient reverberation, where the echo of the initial sound interferes with the live sound. Another situation where difference tones can occur is with ethnic flutes combined with big drums (Offermans, 2023). Figure 3.3 shows a sample excerpt wherein two flutes are playing together, and visualizes the appearing difference tones of this performance.

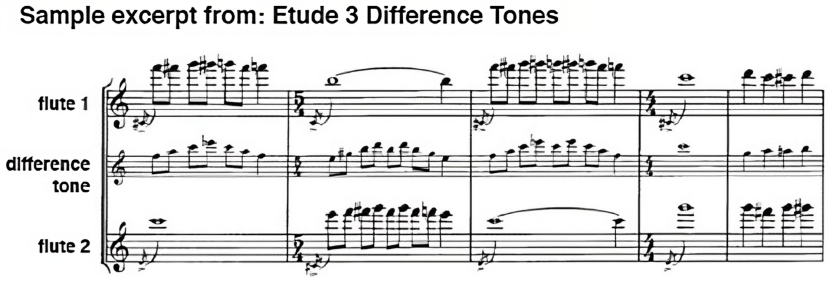

Figure 3.3: Sample of a piece for "For the Contemporary Flutist" illustrating difference tones (Offermans, 2023).

The well-known quote by W. A. Mozart, "What's worse than the sound of a flute? Two flutes.", gains a stronger meaning now in the context of difference tones. While the phenomenon is not exclusive to wind instruments, Mozart clearly expressed his dislike for this additional sound artifact in his compositions. However, setting aside Mozart's disfavor towards flutes, difference tones can be seen as an immense extension to the sound of an orchestra, where the sound of multiple instruments interferes with each other.

3.3 The Fundamental and (Non)harmonic Overtones

A harmonic can be defined as one of the components of the harmonic series, which represents a collection of frequencies that are (nearly) positive integer multiples of a single fundamental frequency. Most real-valued signals, except pure sine waves, consist of a fundamental frequency (the first harmonic) and overtones (higher harmonics), which are sinusoidal components at integer multiples of the fundamental frequency. However, real-valued signals can also contain non-harmonic overtones, which do not follow the harmonic series (Young, 1954). In Figure 3.5, the harmonics are depicted with a

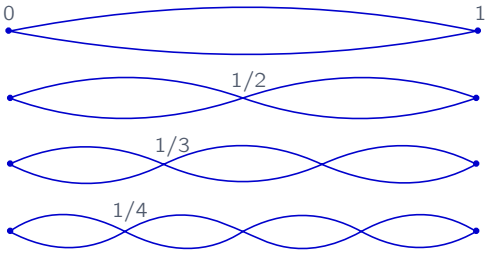

Figure 3.4: A fundamental and its three harmonic overtones.

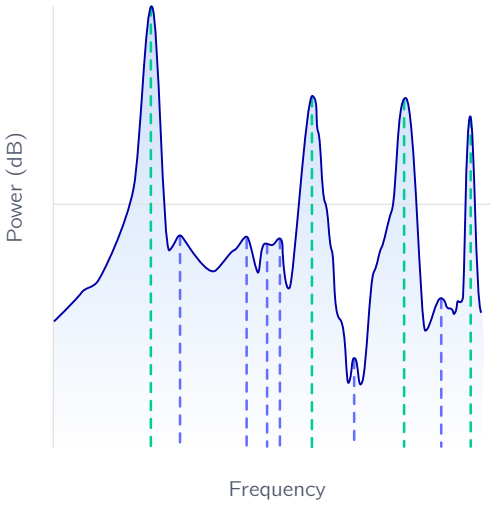

Figure 3.5: This image depicts the distinction between harmonic overtones, represented by the green lines, and non-harmonic overtones, represented by the blue lines.

blue line, while the non-harmonics are shown with an orange line. Organic sounds produced by instruments like guitars and pianos typically include both harmonic and non-harmonic overtones, as the vibrations of metal, wood, and membranes generate non-harmonic overtones, which contribute to the timbre of a sound (McAdams, 2019).

3.3.1 Structualism

According to the theory of structuralism, everyday perception is composed or constructed from basic *sensations*. Psychologists such as Edward Bradford Titchener, who practiced introspection, developed a systematic method to experimentally deconstruct percepts and identify their **constituent elements** in order to understand the underlying structure of perception (Hatfield, 2015). In the context of hearing, various artifacts emerge during the transition from sensing to perceiving. Artifacts such as difference tones and missing fundamentals are just a few examples. Thus, if percepts are syntheses of simpler elements, the question arises whether these elements can be experienced and what they would be in that case. This forms the fundamental philosophy behind the multi-hierarchical structure of constituents developed by Nicholas Harley (2020).

3.4 Summary

This part is a transition to our cognitive approach towards musical analysis and elaborated on the difference between sensing and perceiving. We addressed the psychoacoustic phenomenon where additional tones are perceived due to inference. We introduced the basic concepts of the fundamental tone and non-harmonic overtones. Later in this thesis, we will extract the fundamental and harmonic overtones of musical performances and group the constituent elements of a fundamental together in a knowledge representation.

II

A Cognitive Approach to Musical Audio Analysis

The aim of this part is to present the theoretical foundation for our developed models of perception, cognition, and knowledge representation. The first two chapters are reserved for modelling and explaining the perceptual components of our intelligence. Starting with cochlear mechanics and followed by discrete resonances, which is a simplified model simulating the resonance of the basilar membrane. Additionally, we will examine a density-based clustering algorithm inspired by visual cognition, which holds potential for auditory analysis. Lastly, we will explore the theoretical underpinnings of our knowledge representation, constructed from clustered information.

CHAPTER FOUR

The Mechanisms of Hearing

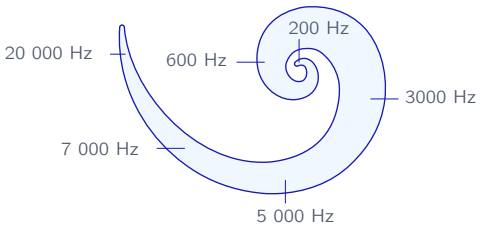

Figure 4.1: Approximated frequency ranges in the cochlea.

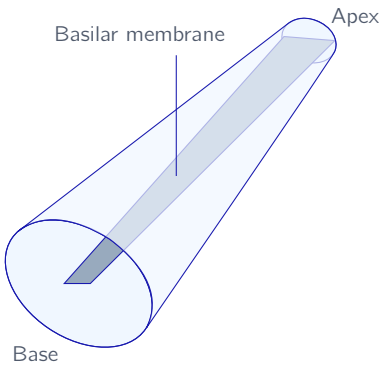

Figure 4.2: A simplified representation of an unrolled cochlea containing the basilar membrane. At the base of the cochlea, the cochlear system decodes high frequency signals and low frequency signals at the apex (Kim and Koo, 2015).

Hearing is both a sensory and perceptual process and involves more than just the transmission of mechanical waves to the auditory cortex. This section will discuss how the Cochlea can be seen as an extremely precise Fourier Analyzer and present the progression of scientific investigations aimed at unraveling the Cochlear mechanics. This will lead to the conceptualization of an idealized model for the receptive fields called resonances, which aim is to simulate (a part of) the inner ear mechanisms.

4.1 Cochlea as a Fourier Analyzer

Sound waves travel through the outer ear canal and initialize vibration in the tympanic membrane by hitting it. This vibration is transmitted to the oval window, creating waves that travel through the fluid of the cochlea, a snail-shaped coiled tube in the human hearing system, which functions as a Fourier analyzer. Inside the cochlea, different spots of the basilar membrane respond to different frequencies of the waves, called tonotopy. This vibration stimulates thousands of about 12 000 tiny cilia (hair cells) inside the organ of Corti, laying ontop of the basilar membrane. It contains outer hair cells (OHC) that convert the auditory signal into electrical signals and inner hair cells that transmit these pulses to the auditory nerves, connected to the auditory cortex in the brain (Elliott and Shera, 2012; Vavakou, Cooper, and Heijden, 2019). Figure 4.1 illustrates the organization of different frequencies on the basilar membrane. The cochlear base, located at the beginning of the tube, is more sensitive to high-frequency waves, while the cochlear apex at the end of the tube is most sensitive to low-frequency waves. The detectable range of sound for the human ear typically falls between 20 Hz and 20,000 Hz.

4.2 Modelling the Mechanisms of Hearing

In the past 200 years, several attempts have been made to understand cochlear mechanics, starting from Helmholtz's resonance theory, which later evolved into traveling wave theory due to the work of Békésy (Manley, Narins, and Fay, 2012). Early experiments conducted by Wegel and Lane (1924) strongly indicated that the resonators in the ear are heavily damped. Later,

Gabor (1947) discussed that the duration of a sound has an influence on the resonance pattern of the inner ear due to the two mechanisms at work: the first mechanism involves the usual resonance pattern of the inner ear, while the second mechanism involves the search for maximum excitation or amplitude, which allows for accurate sound perception when the duration of a pure tone is sufficiently long (as illustrated in Figure 4.3). He explains it as the effect of gradually decrease of stimulation fibers in the auditory system. Additionally, Gabor noted that sound is perceived as "musical" only when the second mechanism comes into play, making it particularly interesting for musical data. For speech perception, on the contrary, it is enough to rely on only the first mechanism.

In more recent work, Vavakou, Cooper, and Heijden (2019) observed that the OHC are operating like envelope detectors, which means that they can detect variations in volume and modify them, by altering their length in response to electrical stimulation before the transmission to the inner hair cell. This property is called OHC electromotility (Brownell, 2017). An interesting detail to mention, is that both Brownell (2017) and Bacon et al. (1999) noticed that the cochlea exhibits less linearity at the base, where the membrane is generally stimulated by high frequencies, compared to the apex, suggesting non-linearity in the excitation of inner hair cells.

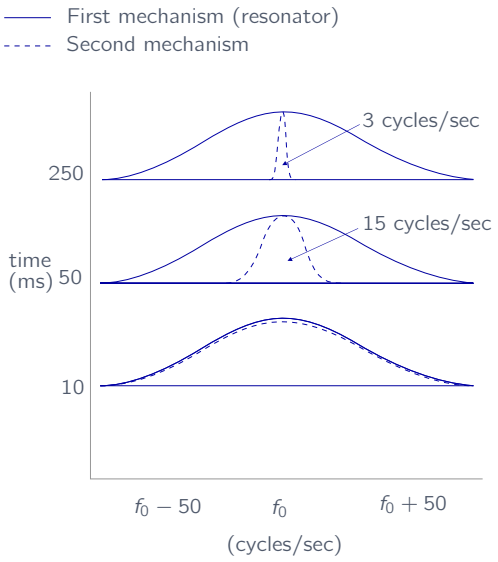

Figure 4.3: The two mechanisms of hearing, wherein the second mechanism tends to approximate the maximum amplitude when the duration of a pure tone increments. The illustration is a reproduction from (Gabor, 1947).

4.2.1 The Fourier Uncertainty Principle and Hearing

It has been proven that linear operators (e.g. windowing, filtering, scaling, ...) cannot exceed the uncertainty bound, and only the trade-off between time and frequency resolution can be improved (Theodor, 1997). The discussed Fourier Uncertainty theorem states that

$$\Delta t \Delta f \geq \frac{1}{4\pi}, \tag{4.1}$$

which implies that it is impossible to localize both a nonzero function and its Fourier transform with great precision in time-frequency analysis, which challenges the STFT to perform with great precision. By precision, we are referring to the capability to accurately track parameters of individual entities (Dubey, 2021; Oppenheim and Magnasco, 2013).

However, Oppenheim and Magnasco (2013) showed that human hearing can discriminate much better than the uncertainty bound, which highlights the relevance of approaching the problem from a perspective that models the hypothetical mechanisms of hearing.

4.3 Summary

We summarized findings about the fundamental processes of auditory perception and illustrated that through the last decades, new details about the inner working of the cochlea have been mentioned in research, agreeing on several behaviors, including the damped or driven behavior of the signal. We grounded our assumptions about the mechanisms of hearing to model the input information for the perception of sound with discrete resonances, which we will discuss in the next chapter.

CHAPTER FIVE

Discrete Resonance Spectrum

The discrete resonances are the input information for an idealized model of auditory receptive fields, and a proxy for the information that is received by auditory cortex from the cochlea. A resonance produces, as described by Homer, Harley, and Wiggins (2023), a neural oscillation that damps in a short period of time. This method is an enhancement in the representation, analysis, and processing of intricate non-stationary auditory signals, outperforming the standard Fourier Transform with orthogonal basis in terms of precision due to nonlinearity. We will start with a mathematical description of discrete resonances and weave afterward some seemingly diverse mathematical topics together for the derivation of the Fast Padé Transform. Finally, the benefits of a non-orthogonal basis in a Hilbert space will be demonstrated.

5.1 Discrete Resonances in Time Domain

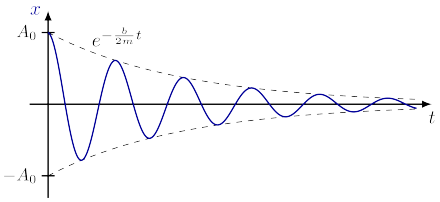

Figure 5.1: Visualization of the real part of a resonance with an exponential decay factor as an envelop (Neutelings, 2021b).

As with Fourier analysis a time-domain signal can be decomposed in different sines and cosines, a signal can also be decomposed in a linear combination of K complex oscillators ("resonances"). These time-domain signals are defined as follows:

$$x(t) = \sum_{k=1}^{K} d_k e^{-i\omega_k t} \qquad d_k, \omega_k \in \mathbb{C}. \tag{5.1}$$

Signal x is formed by a summation of different complex resonances with d_k the initial complex amplitude of the oscillator and a rotating vector $e^{-i\omega_k t}$, dependent on the complex frequency ω_k. It might seem illogical to see the notation of a continuous signal for discrete resonances (i.e., $x(t)$ instead of $x[n]$). We will leave the details behind this formulation beside, the important thing is to know that it is based on the Continuous Fourier Transform, and will only be discretized in the actual implementation.

Just to prevent confusion with the previous chapter: the definition of resonances describes a time-signal that is decomposed in damped or driven oscillators, just like Fourier series did for stationary signals. In resonances, a fourth dimension is added, containing the decay. In a 3D representation,

the following visualization in Figure 5.2 of a resonance can give a more intuitive feeling for its appearance. The decay can be observed as the complex exponential function in Figure 5.1.

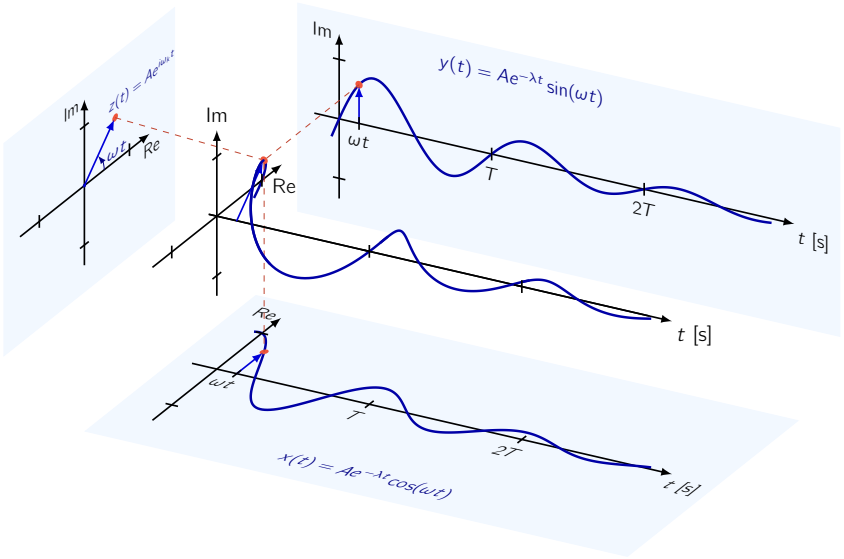

Figure 5.2: A visual representation of a resonance in 3D.

The complex frequency ω_k consists of a real and imaginary part, namely the frequency of the resonance ϕ_k and the rate of decay γ_k as a complex part: $\omega_k = \phi_k + i\gamma_k$. A short mathematical interpretation illustrates the intuition behind this formulation:

$$\begin{aligned} e^{-i\omega_k t} &= e^{-i(\phi_k + i\gamma_k)t} \\ &= e^{-i\phi_k t + \gamma_k t} \\ &= e^{\gamma_k t} e^{-i\phi_k t} \\ &= e^{\gamma_k t}[\cos(\phi_k t) - i\sin(\phi_k t)] \\ &= e^{\gamma_k t}[\cos(\phi_k t) + i\sin(\phi_k t)]^* \end{aligned} \tag{5.2}$$

γ_k describes the decay of the resonance, and the sine and cosine terms are a representation of the oscillatory behavior and denote the frequency. Visually, one can think of it as decayed or augmented sine waves. Furthermore, d_k can be rewritten as $|d_k|e^{i\psi_k t}$, with ψ_k denoting the initial phase of the oscillator. The absolute value of the amplitude is the polar representation of d_k.

$$x[t] = \sum_{k=1}^{K} |d_k| e^{i\psi_k t} e^{-i(\phi_k + i\gamma_k)t} \qquad |d_k|, \psi_k, \phi_k, \gamma_k \in \mathbb{R} \tag{5.3}$$

5.2 Discrete Resonances in Frequency Domain

The Fourier transform is a tool for performing spectral analysis of time-domain signals. It is defined with respect to frequency ϕ. Here, the oscillatory signal

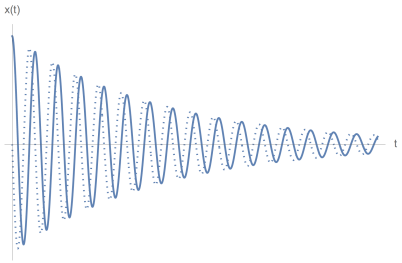

Figure 5.3: A discrete resonance with the complex part visualized with dotted lines and the real part with the continuous line (Homer, Harley, and Wiggins, 2023).

is represented in the frequency domain:

$$X(\phi) = \frac{i}{\sqrt{2\pi}} \sum_{k=1}^{K} \frac{d_k}{\phi - \omega_k} = \frac{i}{\sqrt{2\pi}} \sum_{k=1}^{K} \frac{|d_k| e^{i\psi_k}}{\phi - (\phi_k + i\gamma_k)}. \tag{5.4}$$

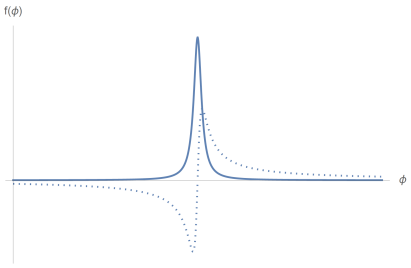

Figure 5.4: A discrete resonance in the frequency domain (Homer, Harley, and Wiggins, 2023).

The absolute value of the amplitude d_k corresponds in the frequency domain to the size of the resonance peak. Frequency ϕ_k defines the location of the resonance peak, and decay γ_k influences the width and polarity of the resonance peak in the frequency domain. Here, again, ψ_k is the initial phase of the oscillator. In the frequency domain, it has an effect on altering the angle of the rotating vector, thus the cotangent of the angle: $\mathrm{Re}[f(\phi)]/\,\mathrm{Im}[f(\phi)]$.

5.3 The Fast Padé Transform (FPT)

In section 1.2.1, we briefly discussed the Fast Fourier Transform. The power for finding the coefficients hides in the specific selection and evaluation of complex numbers sitting evenly spaced on the unit circle, or, in other words, due to the restriction of having orthogonal bases in a Hilbert Space.

However, when the complex numbers are not evenly spaced, and not even necessarily on the unit circle (which is the case for discrete resonances), it becomes a quite expensive operation to find those coefficients. Therefore, Steven Homer (personal communication, 2023) applied The Fast Padé Transform (Belkic, 2019) on resonances to estimate the spectral density function of a time series signal, which is a distribution of the power of a signal across different frequencies. The Fast Padé Transform itself is a numerical algorithm for approximating the power series expansion of a function in a given interval. It is a fast and efficient and can compute the coefficients of a Padé approximant, which is a rational function that interpolates the power series expansion of a function

5.3.1 Preliminary Knowledge

We will begin by refreshing key concepts from linear algebra. This primer will help establish a solid foundation for understanding the mathematical derivation of the Fast Padé Transform.

Definition 5.1 (Generating function)**.** A generating function $G(x)$ encodes an infinite sequence of numbers by treating them as coefficients of a formal power series.

A simple example of a generating function is the encoding of even numbers with

$$G(x) = \frac{2x}{(1-x)^2}, \tag{5.5}$$

which will generate the sequence of even numbers $\{0, 2, 4, 6, 8, \ldots\}$.

Definition 5.2 (Constant-recursive sequence)**.** A constant-recursive sequence is an infinite sequence of numbers, where each number in the sequence is equal to a fixed linear combination of one or more of its immediate predecessors.

This feature will be important in the further derivation, since the sequence is constant recursive, this will allow us to define analytic functions in a Hilbert space.

5.3.2 Mathematical Derivation

We will now discuss the rough structure of a mathematical derivation for the Fast Padé Transform originating from Belkic (2019) and applied to the resonance spectrum by Steven Homer (personal communication, 2023).

Assume the ordinary generation function G of the infinite constant-recursive sequence of numbers (c_n):

$$G(c_n, z) = \sum_{n=0}^{\infty} c_n z^n. \tag{5.6}$$

z denotes a coordinate in polar representation (i.e., $e^{-iw_k t}$).

By definition, a *formal* power series does not have to converge, but since this generating function is defined to be an analytic function in a Hilbert Space, it means that it has a convergent power series expansion. The goal of the derivation is to come define a closed-form expression that can be evaluated and makes the definition valuable. Assuming that c_n is constant-recursive sequence, the generative function can be rewritten as

$$G(c_n, z^{-1}) = \sum_{n=0}^{\infty} c_n z^{-n} = \frac{\sum_{k=0}^{K-1} p_k^- z^{-K}}{1 + \sum_{k=1}^{K} q_k^- z^{-K}} \left(\frac{z^K}{z^K} \right) \tag{5.7}$$

Which is a Padé approximate by definition. Note that, for the sake of math, a multiplicative inverse of the complex number z was used. The modification to the negative angle does not change anything about the real value of the signal, since a real number is equal to a complex number (no matter positive or negative) with its imaginary part equal to zero. The negative annotation at the coefficients was introduced to denote that they are a coordinate in the negative domain. After the derivation, it returns to the positive domain. The derived generic function can be rewritten as

$$\sum_{n=0}^{\infty} c_n z^{-n} = \frac{\sum_{k=1}^{K} p_k z^k}{1 + \sum_{k=1}^{K} q_k z^k} \tag{5.8}$$

This equation is solvable by deriving this equation with respect to the variable z. By introducing the initial complex amplitude d_k, the equality can be rewritten as

$$\sum_{k=1}^{K} \frac{d_k}{1 - (\frac{z_k}{z})} \tag{5.9}$$

For a full derivation, please read Belkic (2019), page 65-77). This form looks exactly as a geometric series $\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$, with $|r| < 1$ and thus if the following inequality holds: $|\frac{z_k}{z}| < 1$, the equality can be written as

$$\sum_{k=1}^{K} d_k \sum_{n=0}^{\infty} (\frac{z_k}{z})^n. \tag{5.10}$$

Therefore,

$$c_n = \sum_{n=1}^{K} d_k z_k^n. \tag{5.11}$$

Notice that z_k represents the complex plane. The derived summation represents a function in n. If the signal is a simple cosine, then it moves around the unit circle when n increases. In the case that the frequency is complex, and the decay is non-zero, the coordinates will fall inside or outside the unit circle. In other words, c_n can be expressed as a sum of oscillators. Solving this equation for d_k and z_k is exactly what the FPT does in a certain interval, i.e., rectangular window, defined as following:

$$\sum_{n=0}^{\infty} c_n z^{-n} - \sum_{n=N}^{\infty} c_n z^{-n} \tag{5.12}$$

Solving the equation by substituting it with the derivation we had before, gives us

$$\sum_{n=0}^{N-1} c_n z^{-n} = \frac{\sum_{k=1}^{K} p_k z^k + \sum_{k=1}^{K} r_k z^{k-N}}{1 + \sum_{k=1}^{K} q_k z^k} \tag{5.13}$$

For a window function where p and r represents the coefficients at the left and right side of the window. which is exactly the FPT (can also be seen as a convolution). Essentially, this derivation allows us to find the c_n's and z, and therefore the oscillators. And since K = N/2, there is a unique solution.

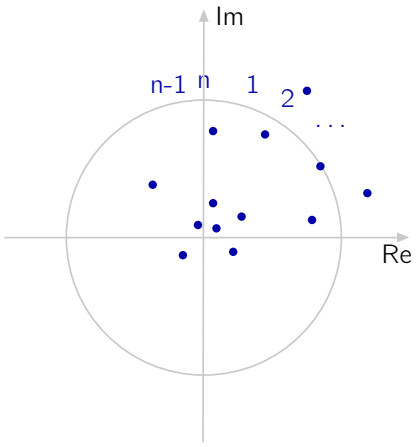

Figure 5.5: A pole-zero diagram consisting of distinct points that are not uniformly distributed due to the non-orthogonal bases and are no longer located on a unit circle due to the complex frequency ω_k.

Non-Orthogonal Bases in a Hilbert Space

Although the usage of orthogonal bases ensures efficient computation and a simple representation, using a collection of non-orthogonal basis functions can be advantageous in some cases, such as in the Hilbert Space of Resonances. The non-orthogonal bases allow us to have non-evenly spaced frequencies γ_k, which is a sequence of numbers. By gaining more freedom with the placement of the frequencies in this space, the frequencies of a signal can be found more precise without limiting itself to a sample rate, as illustrated in Figure 5.5, and a bigger space in the complex plane can be explored.

Consider now the normalized sequence of signals

$$\phi_k(t) = e^{i\omega_k t} \qquad , k \in \mathbf{Z}. \tag{5.14}$$

Here, ω_k is a sequence of numbers and if ω_k = k, we obtain the classical Fourier series basis, with ϕ_k an orthogonal basis. However, if ω_k is not an integer multiple, the signals are not orthogonal (Romberg, 2016). Since ω_k is a complex function in our application of resonances (a complex frequency that makes the sinusoidal decay), we do not have evenly spaced points anymore in the discretization of the formula, as is shown in Figure 5.5. The reason why a signal can equivalently be decomposed into resonances, as Fourier analysis does with sines and cosines, is because

$$e^{\gamma t} cos(\phi_k t) = \frac{1}{2}(e^{i(\phi_k + i\gamma_k)t} + e^{-i(\phi_k - i\gamma_k)t}). \tag{5.15}$$

Summarized, $\langle f|g\rangle = 0$ is not required to be true from an algebraic point of view and the complex frequency with a decaying factor uses non-orthogonal basis.

5.4 Towards higher Precision with Non-Orthogonal Bases in a Hilbert Space

Due to the non-orthogonal property of the Fast Padé Transform, the size of each frequency bin in a time slice is varying with respect to the parameters of a resonance. This has enormous advantages in terms of precision when performing spectral analysis. Figures 5.6 and 5.7 provide a visual demonstration of the fundamental difference between the widely used Short Time Fourier transform and the Fast Padé Transform with a synthetic audio recording of a flute performing the musical note A4. The visualized signals were filtered on power and frequency for the purpose of simple visualization of the main difference between the two methods. The frequency was set in a range between 0 and 2000 since the fundamental frequency and its observable harmonics generally lay between this range (Huang, Sun, and Wang, 2017). Resonances with a relatively small power were also removed from both plots. In Figure 5.6, the resolution of the time and frequency domains are fixed and bound with the STFT, and due to that, the signal is estimated with a division over several nearest frequency bins, bounded by Heisenberg's uncertainty principle (Folland and Sitaram, 1997). However, due to the non-orthogonal requirements of the basis, the size of frequency bins may vary and a more precise estimation of the frequency components can be achieved.

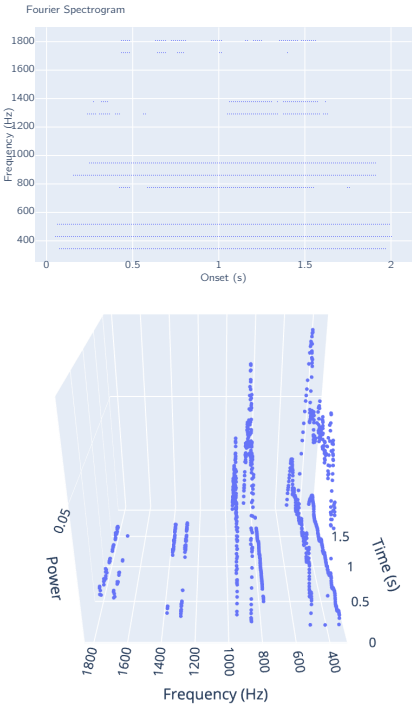

Figure 5.6: A filtered Fourier spectrogram showing the fundamental A4 and its overtones performed by a flute.

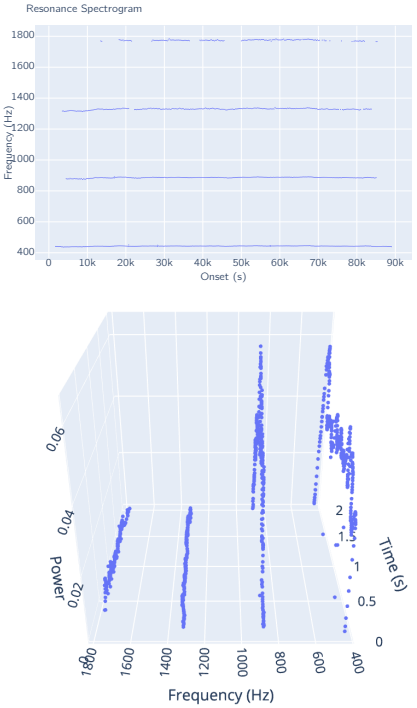

Figure 5.7: A filtered resonance spectrogram representing the same fundamental A4 and its overtones performed by a flute. Note the high precision of this method.

5.5 Extracting Attributes from the Discrete Resonance Spectrum

5.5.1 Dynamic Resonances

The dynamic resonance is a non-explicitly documented technique (Homer, personal communication, 2023), wherein different distance metrics were used to combine two resonances in consecutive slices with each other. The six

distance measures were defined as following: frequency distance, harmonic mean of the d_k and w_k coefficients, residue of the product of the resonances, residue of the product of the resonances weighted by power, residue of the product of the resonances multiplied by the spectra transference function and residue of the product of the resonances multiplied by the spectra transference function weighted by power.

A dynamic resonance is a relation of the distance metric d defined as following:

Definition 5.3 (dynamic resonance)**.** Let S_x and S_{x+1} be non-empty consecutive slices in a spectrogram S with distinct resonances, and let $d(i,j)$ be a distance function defined for all pairs (i,j) representing those resonances, such that $i \in S_x$ and $j \in S_{x+1}$. We define the relation R as follows:

$$\forall i \in S_x, \exists j \in S_{x+1} : d(i,j) \leq d(i,k) \quad \forall k \in S_{x+1} \setminus \{j\}. \tag{5.16}$$

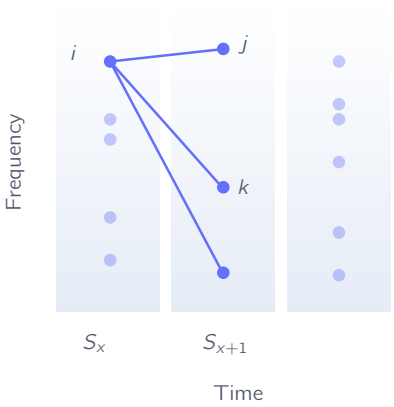

Figure 5.8: Two consecutive slices S_x and s_{x+1} and the relation between two resonances measured with a distance function d.

Due to the slow computational speed of Python and excessive usage of loops, his approach worked significantly slower than our density based approach. However, the insight of using other distance measurements than the Euclidean for the definition of distance inspired us to measure similarity between resonances with the following definition:

$$\cos(d_{jk}) = \frac{\mathrm{Re}\left[\frac{d_j d_k}{w_j - w_k}\right]}{\frac{|d_j|^2}{\gamma_j}\frac{|d_k|^2}{\gamma_k}} \tag{5.17}$$

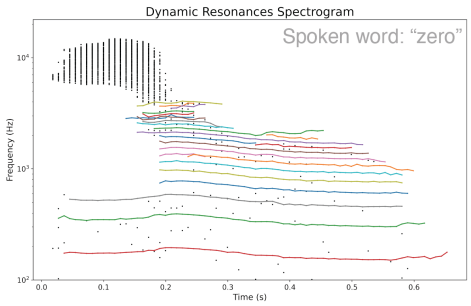

Figure 5.9: Dynamic Resonance spectrum.

After implementing and evaluating this approach, the distance measurement based on similarity of resonances did not contribute to a better clustering. However, it can serve as an inspiration for using other distance measurements in future work to extract more features from the data.

5.6 Summary

We started this chapter with the definition of the discrete resonances in the time and frequency domain and summarized the proof of the Fast Padé Transform, which is a discrete convolution of the coefficients c_n with the coefficients q_k, yielding the coefficients p_k and r_k and zeros. We showed that the main difference with Discrete Fourier Transform is the use of a non-orthogonal basis in a Hilbert Space. We introduced *dynamic resonances* as a novel method for grouping resonances. However, it requires high computational power and therefore, we will introduce a cognitive-based method for cluster analysis for the extraction of musical structures from the discrete resonance spectrum.

CHAPTER SIX

Cluster Analysis of Resonances

Clustering is an unsupervised machine learning technique that involves grouping similar data points based on a specific parameter, such as density or similarity. There are various models known in the literature, with K-means being a conventional one that generates a fixed number of clusters associated with a central point. The Markov Cluster Algorithm, introduced by Stijn Van Dongen (2008), is more appropriate for graphs/networks. Hierarchical clustering on the other hand is often used for the analysis of social network data and biological data analysis (Hexmoor, 2023; Yeturu, 2020). An important drawback of both K-means and hierarchical clustering for our application is that they do not automatically determine the number of clusters. Density-based algorithms, however, such as Mean-Shift, DBSCAN, and HDBSCAN, are more appropriate for this particular problem: resonances require a density-based approach (sudden changes of dense regions imply new musical objects), and they are also capable of automatically determining the amount of clusters based on the input data. The previews provided in Figure 6.1 and 6.2 highlights the benefits of using density-based algorithms for resonance data.

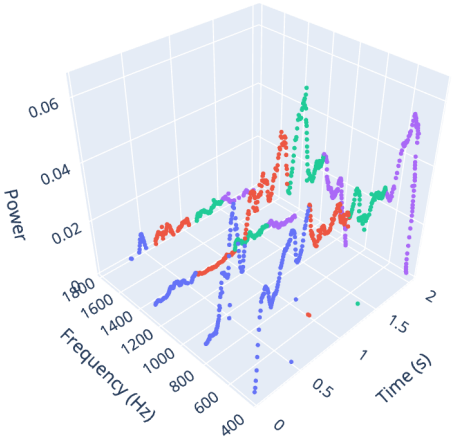

Figure 6.1: The resonances in this figure is a representation of the resonances with the strongest power, extracted from the sound of a flute playing the note A4. They are clustered by the K-means algorithm ($K = 4$) and each cluster is represented with a different color.

6.1 Density-based Cluster Algorithms

First, let us provide a concise overview of the three density-based cluster algorithms mentioned above. The iterative Mean-Shift algorithm moves each data point towards the mean of its respective region to form clusters. This is a centroid-based algorithm and works best for blob-shaped data. DBSCAN is capable of identifying outliers as noise, unlike the Mean-Shift method, and performs effectively on densely populated data with irregular shapes. HDBSCAN is a variation of DBSCAN introduced by Campello, Moulavi, and Sander (2013). In this algorithm, DBSCAN's principle of border points (see further) is abandoned, and only core points are considered as part of the cluster. Even though this method may be beneficial for handling noisy data, DBSCAN is nonetheless deemed to be the most appropriate clustering model for this problem, as it is better to implement custom filtering methods for noise reduction before performing clustering.

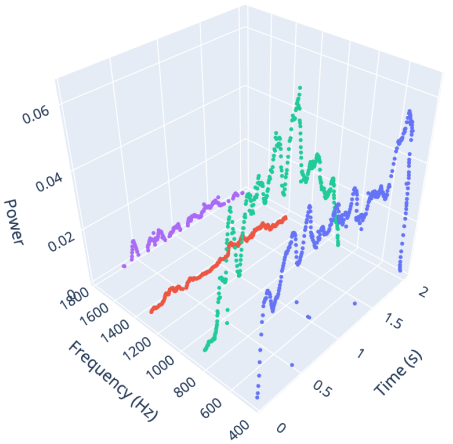

Figure 6.2: Resonances clustered by the DBSCAN algorithm ($\epsilon = 0.4$, $minPts = 4$) are represented with different colors. The labeling mimics how a human would draw circles around resonance groups to extract specific features, which exactly aligns with our desired outcome.

Figure 6.3: Step 1 | Select a point p and assume minPts = 3 and $\epsilon = 0.1$. If at least 3 points are inside radius, mark p as a core point.

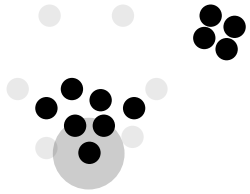

Figure 6.4: Step 2 | Iterate over each point and mark all core points, classify left-overs as non-core points.

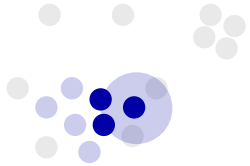

Figure 6.5: Step 3 | Pick a random core point, assign it to the first cluster together with all core points in the radius.

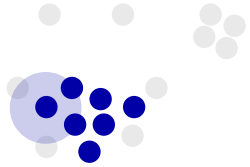

Figure 6.6: Step 4 | Repeat for the neighbouring core points.

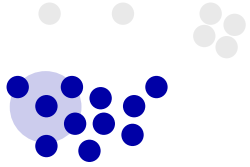

Figure 6.7: Step 5 | Once all the core points have been included in the initial cluster, the border points, which are non-core points within the radius of the core points, are added to the same cluster as well. Note that these border points are not extended iteratively.

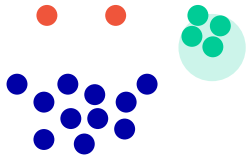

Figure 6.8: Step 6 | Repeat this process to find all clusters. Points that do not belong to any cluster are called noise points and are marked with red.

6.2 DBSCAN clustering algorithm

Density-based Spatial Clustering of Applications (DBSCAN: Ester et al., 1996) is a data based clustering algorithm. The algorithm attempts to imitate the human ability to recognize groups of points with an arbitrary shape that are closely located to each other, and singles out isolated points as noise. Figures 6.3-6.8 visualize the DBSCAN algorithm step by step. The model estimates the minimum density level using a method that relies on a threshold value, minPts, for the number of neighbors within a radius ϵ. The algorithm begins by marking all core points in the dataset. A point p is classified as **core point** if there are at least minPts points (including p) within the radius ϵ. In the next step, a random core point i is selected from which all transitively included (i.e.,density-reachable) core points are identified to form a cluster. Then, all **border points**, which are non-core points in the radius of a core point, are added to the cluster of a core point. Notice that a border point that is in the radius of two core points of different clusters will just be classified in the first cluster that is processed. Any points that are not density reachable from any core points are classified as noise and do not belong to any cluster. To ensure that all points in the same cluster are included, the minimum number of points (minPts) should be set to a relatively low value (Ester et al., 1996; Schubert et al., 2017).

6.2.1 Complexity and Data structure

Multiple implementations of the DBSCAN algorithm exist. A more optimized implementation utilizes a FIFO queue to keep track of the points which are already labelled and a R^*-tree, Kd-tree, or cover tree for performing a continuous search for density points within a tree-like structure (Schubert et al., 2017). The average time complexity is $O(n \log(n))$, since the neighboring queries are executed in logarithmic time, and labeling core and non-core points takes $O(n)$. Worst case, with degenerate data or naive implementations (e.g., not using the index structure), the time complexity becomes $O(n^2)$. We use an adjacency list-based implementation, which is better in terms of running time and memory usage compared to the matrix-based implementation.

6.2.2 DMBSDSCAN

A drawback of DBSCAN is that it performs less well at data with a wide variation in density. DMDBSCAN attempts to solve this problem by introducing a dynamic ϵ estimation, suitable for each density level in the dataset (Elbatta and Ashour, 2013). The algorithm is suggested for future work in case that the static ϵ-estimation would not be sufficient. The required accuracy can currently be pre-defined and adjusted by the user.

6.3 Summary

We delved into a density-based clustering approach, in which labeling mimics how a human would draw circles around resonance groups when plotting it in a certain dimension. The next section will introduce a hierarchy wherein we will be able to represent the obtained clusters and attribute them.

CHAPTER SEVEN

Type-Based Knowledge Representation

Shi (2019) introduced the evolution of the human brain, denoting that the appreciation of music by the brain cortex coincided with the development of human abstract thinking. Knowledge is the possession of the ability to locate information, and therefore, we will introduce Harley's type-based framework for the abstract representation of musical knowledge (Harley, 2020). His framework uses a constituent structure (i.e., a multi-hierarchical information model) that is able to represent the complex hierarchies of musical spaces. The framework has a main, extendable module named CHAKRA, which gives the user the freedom to create structures in terms of constituents, attributes and hierarchies. A submodule of CHAKRA, named CHARM (Wiggins, Harris, and Smaill, 1989), is a particular representation system for the creation of music-specific multidimensional hierarchies. First, we will emphasize the difference between knowledge and data, followed by highlighting the importance of the type-based aspect of the CHAKRA framework, by introducing three perspectives on computation: Type theory, Category Theory and Typed logic. Afterward, the components of the CHARM system will be briefly defined, serving to understand our own software architecture, introduced in part III.

7.1 The Three Perspectives on Computation

The Three Perspectives on Computation, also called *The Computational Trinity*, is the central organizing principle of a theory of computation that unifies Logic, Type Theory, and Category Theory (Eades, 2012; Goguen, 1991). The three fields look different but are nevertheless equivalently treatable. One may think of it more intuitively as the unification of a subfield of logic, programming, and mathematics, wherein every proof can be written as a program, every program corresponds to a mapping, and every mapping to a proof (Harper, 2011).

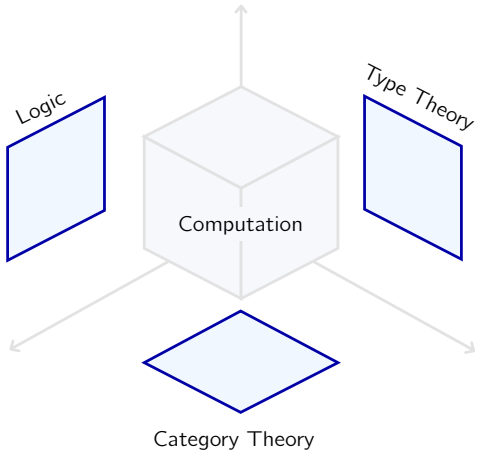

Figure 7.1: Computational Trinity: a tripartite correspondence between Logic, Type Theory, and Category Theory.

7.1.1 Type Theory

At the early 20th century, Bertrand Russell introduced type theory to cope with a paradox in naive set theory, which was expressed as follows:

$$H = \{x | x \notin x\} \Rightarrow H \in H \Leftrightarrow H \notin H. \tag{7.1}$$

The paradox arises when one considers whether H contains itself or not. If H contains itself, then by definition it is a set that does not contain itself, which is a contradiction. On the other hand, if H does not contain itself, then by definition it is a set that should contain itself, which is also a contradiction. This problem arises when impredicative universal quantification is allowed, which means that the definition of the object involves quantifying over all objects, including the object being defined itself. Russel's type theory resolved this problem by defining objects as part of a specific group. Consider a type n, then we can redefine H as

$$H^n = \{x^{n-1} | x^{n-1} \notin x^{n-1}\} \Rightarrow H^n \in H^n \Leftrightarrow H^n \notin H^n. \tag{7.2}$$

In this case, the paradox is false since the sets are defined by the types, and type $n-1$ excludes type n (Eades, 2012).

Type theory became the formal presentation that models objects and relations, such as a variable, function or substitution, with types. For example, variable 10 has the type of natural numbers ($\mathbb{N}$), which is in the built-in notation written as $10 : \mathrm{nat}$. From this term, other typed terms can be constructed. As illustrated in (Hoang, 2014), terms of the type $\mathbb{N}$ can be constructed just by defining the variable as we defined before, or as a successor function $succ(n) : \mathbb{N}$:

$$0 \xrightarrow{succ} 1 :\equiv \mathrm{succ}(0) \xrightarrow{succ} 2 :\equiv \mathrm{succ}(1)$$

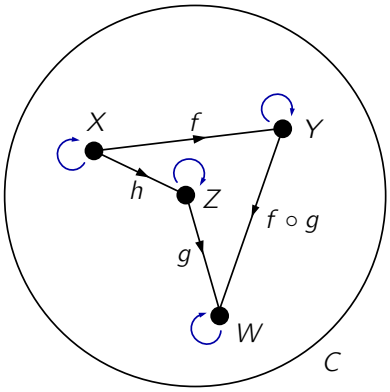

Figure 7.2: Category C with a set of objects $\{X, Y, Z, W\}$, morphisms $\{f, g, h\}$ and composition of morphisms $\{f \circ g\}$. Each object also has an identity morphism: an arrow points to the object itself.

7.1.2 Category Theory

Tom Leinster described category theory as a bird's eye view of mathematics. From high in the sky, details become invisible, but we can spot patterns that were impossible to detect from ground level (Leinster, 2016). Formally, category theory is the study of mathematical structures using abstractions of functions called morphisms, as well as a mathematical workspace and theory (Barr and Wells, 2012; Eades, 2012). It provides the concepts to meaningfully compare and combine unrelated systems by understanding their patterns (Harley, 2020). A category C is an abstract object consisting of a set of objects, morphisms and compositions of morphisms. The objects and relations can visually be represented as directed graphs, as illustrated in figure 7.2.

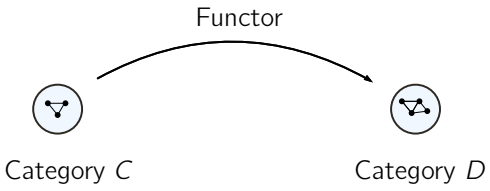

Figure 7.3: A functor represented as a directed edge from category C to category D.

Categories are connected through structure-preserving maps named functors. They can be considered as morphisms in a category of subcategories. This theoretical framework of formalization is especially useful for combining different levels of abstraction (e.g. the unification of CHARM with other modules as will be explained further), and for formal descriptions of systems.

7.1.3 Typed logic

The perspective on type theory, from a logical aspect, is the definition of certain rules that must hold. An example of a formal language using these inference rules for programming purposes, is Calculus of Inductive

Constructions (Cic). An example of a typing rule in Cic is defined as following:

$$\frac{E[\Gamma] \vdash T : s \quad s \in \mathcal{S} \quad x \notin \Gamma}{\mathcal{WF}(E)[\Gamma :: (x : T)]} \tag{7.3}$$

Where a term t is correctly typed in a global environment E if and only if there exists a local context Γ and a term T such that the judgement $E[\Gamma] \vdash t : T$ can be derived from the following rules, as literally mentioned by Inria (2018). However, the interpretation of this rule is not important in context of the thesis and only serves as an illustration.

7.2 The CHAKRA System

Type theory is a strong and important foundation for the implementation of the *Common Hierarchical Abstract Knowledge Representation for Anything* (CHAKRA) framework, since it allows the integration of heterogeneous data and avoids the paradox of naive set theory. CHAKRA was originally defined in Coq[1], a library with Calculus of Inductive Constructions as underlying formal language (a mapping from logic to programming). Afterward, an implementation of the CHAKRA-framework was written in Julia [2].

7.2.1 The CHARM System

The *Common Hierarchical Abstract Representation of Music* (CHARM) intends to provide a logical specification of an abstract representation of music (Pearce, 2005; Wiggins, Harris, and Smaill, 1989), regardless of the particular style, data source or application (Smaill, Wiggins, and Harris, 1993). It inherits the abstract structural components Constituent, Id and Hierarchy from CHAKRA and applies it in the musical domain (Harley, 2022b).

Musical Objects with Attributes in a Space

Musical objects are seen as statements about the world in the musical space. These atomic entities automatically add semantics to data wherefore they can be considered as **knowledge**, rather than data. Note that data can be seen as the simplest form of raw values. Knowledge, in contrast, is a statement about something in the world space that could be true or false.

We call every existent musical object a *Constituent* and use them in a hierarchical structure. Musical objects contain attributes such as $frequency$, $amplitude$ and $onset$ which can be defined in an (abstract) musical space. Often the attribute name and the name of the space is the same (for example, the attribute pitch and musical space pitch). However, sometimes they do not overlap (for example, the attributes *onset* and *off-set* both have the musical space of time, but it is important that their attributes names are different) (Harley, 2020). In Figure 7.4, an example of a musical object (i.e., constituent) is given. The constituent is formed from a joint pair of

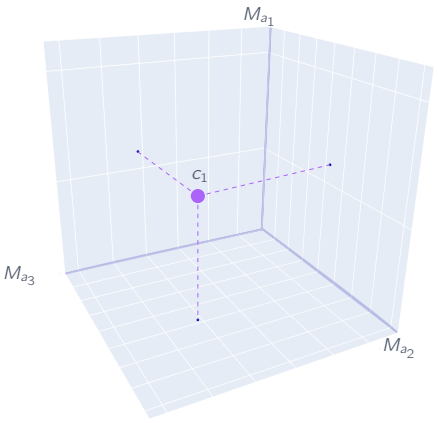

Figure 7.4: A simplification of a constituent $c_1 \in C$ defined in the musical space M.

[1] https://nick-harley.github.io/chakra-coq/chakra.html
[2] https://github.com/nick-harley/Chakra

resonance r_1 and $-r_1$, and has 10 attributes. The attributes of musical objects are equivalent to the features of a dataset.

Definitions

The first concept defined in CHARM, is the Constituent C, a high-level grouping of musical objects defined in (multiple) musical spaces M (Wiggins, Harris, and Smaill, 1989). Musical objects are atomic entities of the human conceptualizations of music, such as resonances, slices in time of an audio signal or onsets of notes. The set of locations in musical space $M = (M_a)_{a \in A}$ is a family of sets indexed by attributes A and M_a is the subspace or dimension indicated by the perspective a (Harley, 2020, p. 111). For example, a constituent representing a resonance, contains the attributes frequency, onset, decay and amplitude.

Theorem 7.1 (Attribute)**.** The set of Attribute names A is a collection of keys for key-value pairs, where key $a \in A$.

Constituents are used to construct hierarchical structures. They are connected in a directed acyclic graph and its relation is called a Hierarchy H. Note that a single constituent can be defined in multiple hierarchical structures.

Definition 7.2 (Constituent)**.** A musical object (i.e., a constituent) is composed of a tuple <i, P(i), R(i)> where identifier = i, particles = $P(i)$ and attributes =$R(i)$ (Harley, 2020, p. 112).

We explicitly use the notation R(i) here instead of A because A is the set of names and R a set of relations between constituents and values. Thus, $(a, v) \in R_i$ means that the value of an attribute should be taken.

Definition 7.3 (Hierarchy)**.** $H \subset C \times C$ is a binary relation C, such that (C, H) forms a (simple) directed acyclic graph (DAG). The graph (C, H) captures the hierarchical structure of the domain, where $(c, c') \in H$ indicates that c' is part of c (Harley, 2020, p. 111).

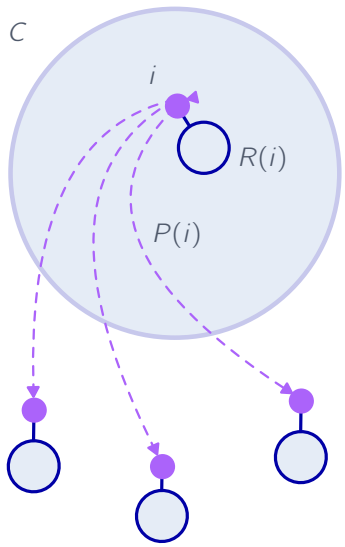

Figure 7.5: A constituent C consisting of the identifier, particles and attributes.

.

7.3 Summary

We introduced the CHAKRA system as a model for the human-like ability to locate and use information. The theoretical underpinnings of this framework were explored from three different perspectives, providing a better understanding of its principles. Finally, we defined the constituent elements of CHAKRA in context of music, setting the foundation for the actual implementation in the following part.

III

Software Architecture

In this last part, we will demonstrate our own contribution, holding the connection between the *already* developed model of perception with our own model of cognition and knowledge representation. We will start with the section that introduces our knowledge representation, since it gives an overall view of obtained clusters of constituents. The hierarchies can generally be categorized in four musical levels: frame-level, note-level, stream-level and notation-level (Benetos et al., 2019). We demonstrate the possibilities of the developed software for a note-level description of music and touch the frame-level description with a first proposition.

CHAPTER EIGHT

Knowledge Representation Applied to Audio Files

8.1 Specification of the CHAKRA abstraction

We created a constituent structure (i.e., a multi-hierarchical information model) to model a theoretical estimation of the underlying structure of perception discussed in 3.3.1 Structuralism. This constituent structure was an expansion of Harley's type-based framework for the abstract representation of anything. Learning and expanding knowledge about musical data was achieved through density-based clustering techniques. The acquired knowledge creates new dimensions in the hierarchical structure and gives more structural insights about musical structures in audio fragments. We will start with a discussion of the components of the music-specific CHAKRA implementation on resonances. Afterward, we briefly explain several engineering considerations during the development of this software.

The central concept of the CHAKRA system is that it establishes the fundamental abstract types known as a *Constituent, Identifier* and *Hierarchy*. The user of this framework is then responsible for creating a customized implementation for these abstract types, inheriting their underlying structures.

8.1.1 Constituent

Every musical object in the hierarchical structure is called a *Constituent*. New constituents are formed from finite sets of other constituents and are composed of an identifier i, particles $P(i)$ and attributes $R(i)$, as has been defined in 7.2. The value of an identifier is unique in a single set of identifiers. We call particles the children of a constituent (i.e., node) and the attributes are specific features of a constituent (e.g., onset, pitch, duration).

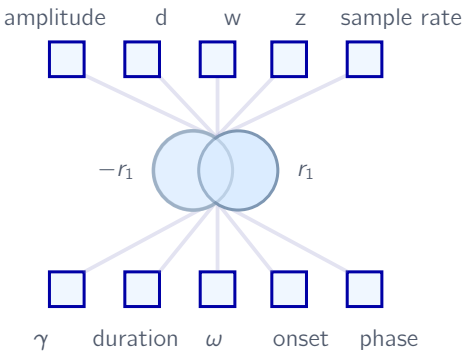

Figure 8.1: Visualization of a constituent containing a joint pair of resonances and its musical spaces.

Constituents of a DRS

The smallest musical objects in our particular hierarchy are resonances, they lay at the lowest layer of the hierarchy. We define each resonance as a constituent, containing 10 attributes, as illustrated in Figure 8.1.

Each resonance has a conjugate pair with similar attributes; although the conjugate has the opposite sign of the frequency. For sake of completeness, the imaginary part of the signal is kept and the resonance and its conjugate

are combined in a new *pair*-constituent. From those pairs, slices in time are taken. A slice is defined as a small period of time containing a subset of the resonances (Figure 8.2). These slices are, in their turn, grouped by a DRS-constituent. This constituent represents all the resonances in the time-domain of a Discrete Resonance spectrum.

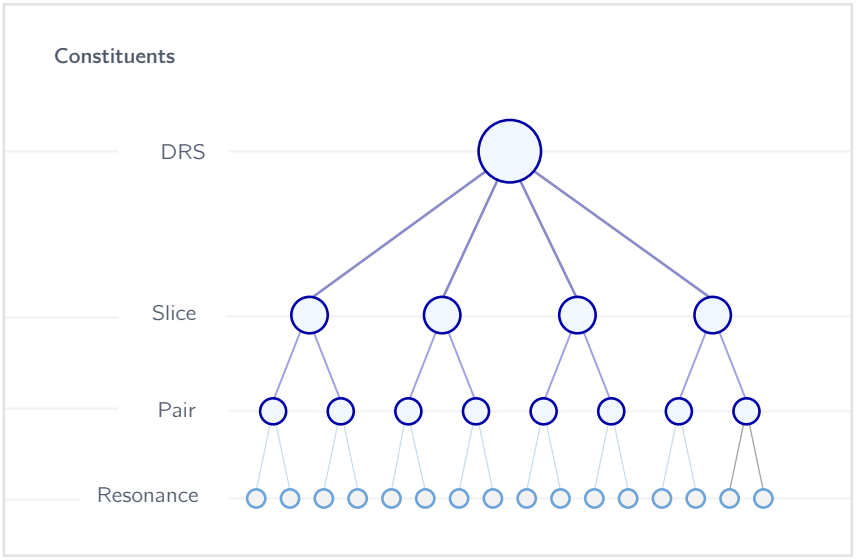

Figure 8.3: Two-dimensional hierarchical structure of the DRS hierarchy, consisting of four levels of constituents, denoted with circles.

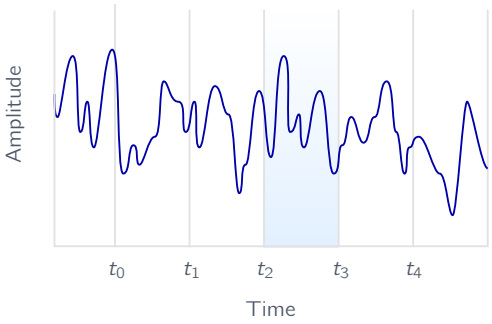

Figure 8.2: A rough illustration of a slice in the time-domain representation of a signal. All musical objects falling inside a time-slice t_n are grouped together by a slice constituent.

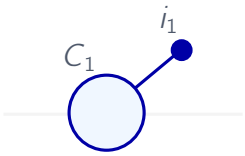

Figure 8.4: A constituent c_1 and its corresponding ID i_1.

8.1.2 Identifier

The relation between Constituents and Identifiers is bijective in the multi-dimensional hierarchical structure. Their value can be repetitive through different sets, but will always contain a unique value in a single set of identifiers (e.g. ResIDs). The identifiers are important assets for finding components of Constituents. By returning the identifiers of the underneath level instead of the knowledge, improvements over speed and memory are achieved. Figure 8.5 emphasizes the usage of sets for the definition of different IDs.

8.1.3 Hierarchy

A Hierarchy *H* is a direct acyclic graph of constituents and is defined as the abstract type *Hierarchy*. Subtypes of *Hierarchy* directly correspond to specific implementations of this abstract type. We will now explore four implementations of the hierarchies currently being developed, which can be further expanded by incorporating additional machine learning techniques.

Definition of Several hierarchies

The **DRS hierarchy** originates from the idea of a Discrete Resonance Spectrum. It groups resonances (both negative and positive) based on information from the time domain. Note the difference between the DRS constituent and the DRS Hierarchy. Although they have the same name, they do not

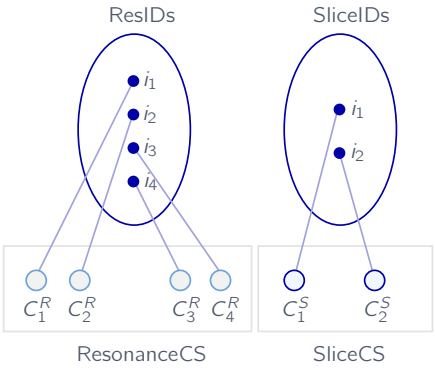

Figure 8.5: The different sets of constituents and their corresponding IDs allows them to have similar names as long as they are defined within a different set.

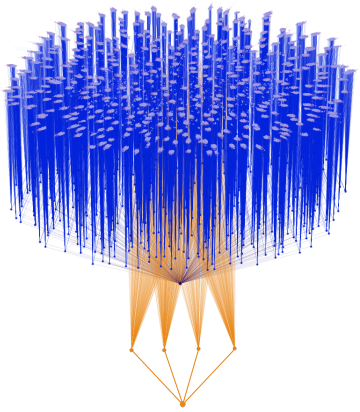

Figure 8.6: Real-valued representation of the DRS and HARM hierarchy for an audio fragment of a flute playing A4. The blue-like points are part of the DRS structure, including the resonances, pair, slice and DRS constituents. The orange points represent the constituents of the HARM hierarchy (the graph was generated with Gephi).

hold the same semantic value. Since half of the resonances are a subset of the negative part, they are often not needed for the analysis of real audio signals. Therefore, we defined the **NEG hierarchy** to group the resonances with a negative frequency together. This connection can make filtering easier if the imaginary part of the signal is not required. The **NOTE Hierarchy** was created for the extraction of musical notes from spectral data. The constituent representing a musical note is defined attributed with *onset, duration* and *pitch* and can formally be defined as following:

$$\text{NOTE\_Dataset} : \text{CSPEC} := \forall_{\text{Parts}}(\text{Note}) \tag{8.1}$$

Where CSPEC is the *type* of constituent specifications defined in Calculus of Inductive Constructions (see Chapter 7.1.3). Finally, the **HARM Hierarchy** contains constituents of overtones corresponding to a specific fundamental tone. A conceptual representation of the four hierarchies is represented in Figure 8.7 and a real-valued representation of the DRS and HARM hierarchies has been plotted in Figure 8.6.

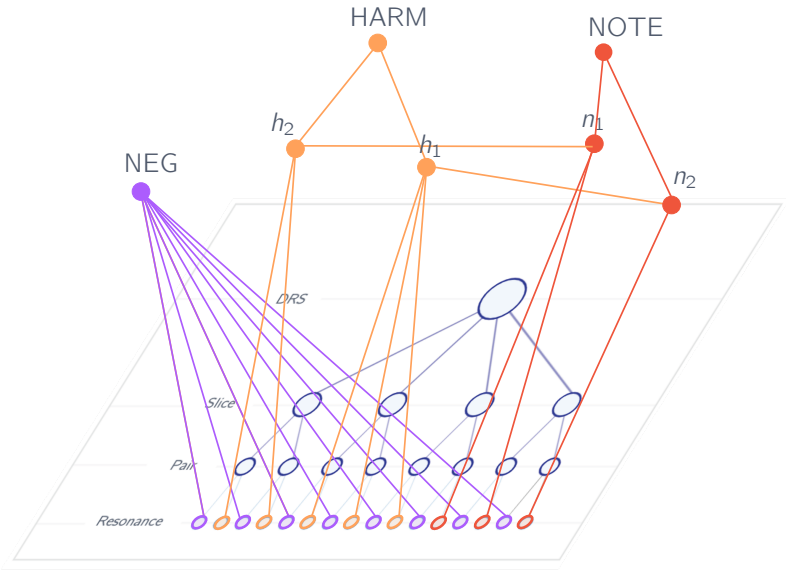

Figure 8.7: A practical visualization of the state-of-the-art hierarchies in our software. Each hierarchy is visualized in its own color, containing Constituents represented as nodes.

8.1.4 Operations

The two most important operations in the hierarchical structures are *parts* and *find*. Parts gives all the particles of a selected constituent. The find operation returns all the attributes of a constituent or collection of constituents.

8.2 Engineering Considerations

We chose to use Julia for knowledge representation modeling and generating new constituents in the hierarchies. This decision was primarily influenced by

two factors: firstly, the existing implementation of CHAKRA in Julia, and secondly, the superior speed performance of this uniquely typed language compared to other high-level languages like Python. Julia has many features, wherefrom multiple dispatch and duck typing are far away the most interesting to mention for our application. Multiple dispatch allows multiple functions with the same name, which is fast and was consequently exploited in the implementation of our software. Languages supporting Duck typing allow changing objects by adding new methods or attributes to those objects. Duck typing is also present in languages like Python and C++, and is a major type system category. Our main concern about Julia is, however, the fact that the increase of its speed is mainly caused by the cost of the initial compilation time.

8.3 Summary

This section discussed the actual implementation of the knowledge representation applied to the resonance information, which was, in its turn, extracted from Audio files. We proposed several hierarchies, including the HARM and NOTE hierarchy, for the extraction of notes and harmonies from an audio file. The final chapter will discuss how the constituents of both hierarchies were created and attributed.

CHAPTER NINE

Note-level description

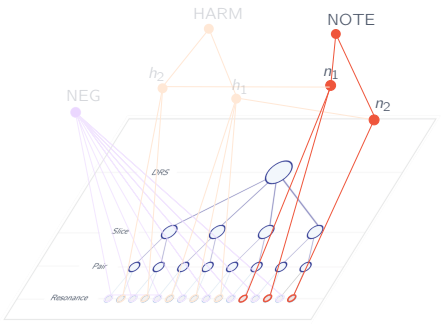

Figure 9.1: The hierarchy of notes is part of the Note-level description of music, highlighted in red.

The goal of this section is to illustrate how musical structures can be extracted for a constituent abstraction in our knowledge representation. Musical structures can in general be divided in a frame-level, note-level, stream-level or notation-level description. The note-level description of music often estimates the pitch, onset time and offset time in literature and corresponds to the descriptive constituents in the NOTE Hierarchy. We will first show why the estimation of those attributes requires the extraction of the fundamental frequencies, and will then extract them through a relative-harmonics based approach. This subset of fundamental frequencies will then be used in the hyperparameter-tuned DBSCAN clustering algorithm from Chapter 6 to recognize individual notes. Finally, we apply a power-based filter to enhance the extraction of notes from a resonance spectrum. This provides a refinement for the pitch extraction of notes. In the last section of this chapter, we also pitch our idea towards the extraction of harmonic overtones.

9.1 Fundamental Frequency Detection

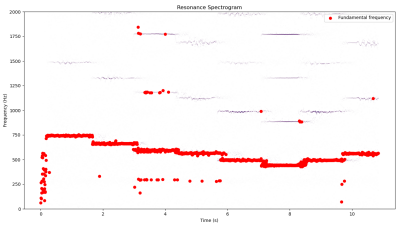

Figure 9.2: A slice (the first two measures) from a real audio recording of a violin performing *Canon in D*, by Johann Pachelbel. Resonances that are estimated to belong to tonic root are highlighted in red.

As discussed earlier, musical instruments do not generate perfect sinusoidal waves. The reverberation of the sound in the environment, unique construction of the instrument, the inimitable single performance of a musician as well as the quality of recording equipment all contribute to the unique sound we capture in an audio recording. Noise, as well as harmonic and non-harmonic tones, interfere with each other, which makes the analysis of music relatively complex. Since humans perceive harmonics as a combination of overtones in harmonic series, we modelled the pitch perception by estimating the fundamental by the frequencies with the greatest relative harmonicity (Figure 9.2). Harmonicity refers to the distribution of power in a resonance, and since harmonic overtones are integer multiples of a fundamental tone, it is possible to estimate the fundamental, even in cases that the fundamental is missing, as was described in Chapter 3 Psychoacoustics. For example, if a flute is playing an A4, the relative harmonicity with respect to A4 will be larger than the harmonicity of B4. The measurement of the relative harmonicity is examined through a resonance-harmonic inner product, described by Homer, Harley, and Wiggins (2023) and defined as cosine similarity:

$$S_C\left(f, Hg_\eta - g_\eta\right) = \cos\theta = \frac{\mathrm{Re}\left[\langle f \mid Hg_\eta - g_\eta\rangle\right]}{\|f\| \cdot \|Hg_\eta - g_\eta\|} \tag{9.1}$$

The fundamental frequency $\eta_0 = \arg\max_\eta \left[S_C\left(f, Hg_\eta - g_\eta\right)\right]$, which exactly estimates the frequencies with the greatest relative harmonicity. Homer, Harley, and Wiggins (2023) are also engaged in refining a method called the Rameau fundamental [3], which can currently be used for the Tonic root estimation, but should be further elaborated for recordings containing different instruments.

[3] Jean-Philippe Rameau (1722) founded the modern musical theory with the publication *Traité de l'harmonie réduite à ses principes naturels*, by mathematically proving that every pitch consists of a harmony. Rameau believed that the rules of harmony were derived from nature, called *The vibrating world*, and these rules governed all music.

9.2 Clustering

We performed the DBSCAN algorithm (discussed in Chapter 6) on the detected fundamental frequencies in terms of frequency and onset to group resonances belonging to a certain note. In a slice (the first two measures) of a real audio recording of a violin performing *Canon in D*, by Johann Pachelbel (Bridget, 2019), the DBSCAN algorithm has been performed with variable values of ϵ and $minPts$. Frequencies labeled with 0 are removed from the Figure 9.6 and refer to all non-clustered resonances extracted from the audio file, frequencies labeled as -1 are labeled as noise from the Fundamental Frequency detection algorithm.

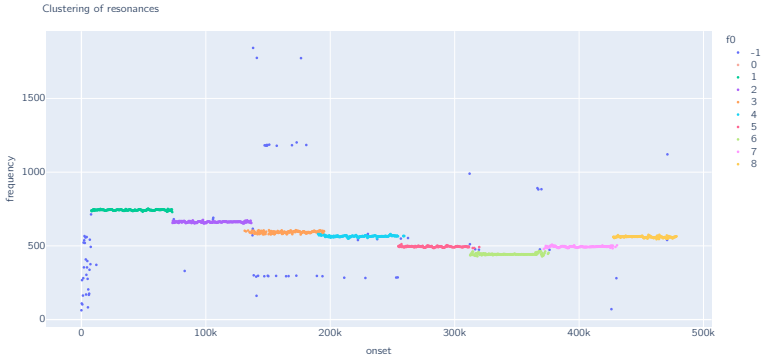

Figure 9.6: A correct labeling with the DBSCAN through correct parameter estimation. Performed on the first two measures of Canon in D.

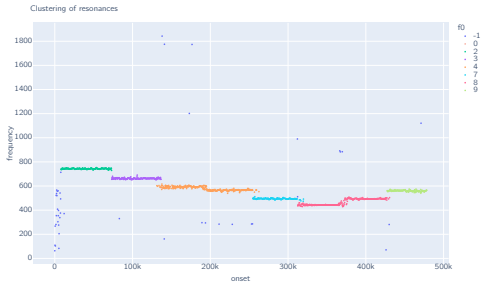

Figure 9.3: ϵ=0.1, minPts=4.

Figure 9.4: ϵ=0.1, minPts=10.

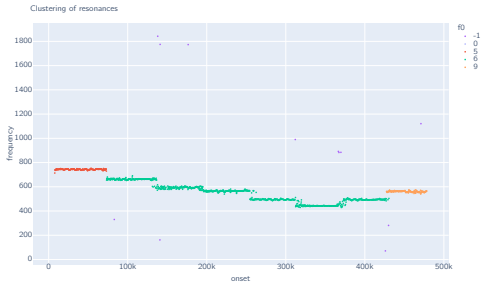

Figure 9.5: ϵ=0.3, minPts=10.

9.2.1 Parameter estimation

As mentioned earlier, ϵ and $minPts$ are the two parameters to be estimated. Variations in clustering when varying the two parameters are represented in Figures 9.3-9.5. Therefore, we will introduce two automatic parameter estimators: the silhouette score and *kneedle* method.

Silhouette Score

The silhouette score is a normalized metric for the evaluation of the quality of a clustering technique:

$$\text{silhouette score} = \frac{\beta - \alpha}{\max(\alpha, \beta)}. \tag{9.2}$$

α denotes the average distance between the points inside a cluster, and β the average distance between all clusters. A high positive value implies that the clusters are will distinct from one another, indicating a good clustering performance. Values close to 0 imply cluster overlapping, and negative scores tending towards -1 imply wrongly assigned points (Rousseeuw, 1987).

Kneedle Method

Several methods exist to estimate an optimal value for ϵ. Satopaa proposed the knee method in *"Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior"* (Satopaa et al., 2011). From a k-distance graph, with values sorted from small to large (or vice versa), the optimal parameters can graphically be estimated by finding a knee or elbow in the graph as illustrated in Figure 9.7. The mathematical definition of the curvature is the basis definition for the knee estimate. Satopaa et al. (2011) defined the Kneedle-algorithm as following:

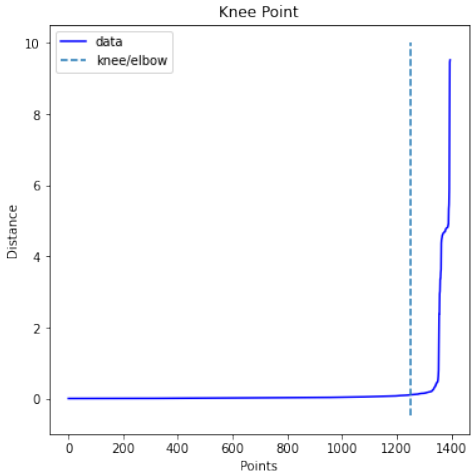

Figure 9.7: Kneedle of a harmonic series from C5 to F played by a flute.

Definition 9.1 (Kneedle)**.** For a continuous function f, there exists a $K_f(x)$ that defines the curvature of f at any point as a function of its first and second derivative:

$$K_f(x) = \frac{f''(x)}{(1 + f'(x)^2)^{1.5}}. \tag{9.3}$$

The point of maximum curvature is used in the Kneedle-algorithm to select the optimal value for ϵ, which is (1 - normalized value) of the knee locator (or just the normalized value if inverse density ordering was implemented). It is worth mentioning that the clustering results are significantly impacted by the choice of ϵ. A small value of ϵ will lead to inadequate clustering, whereas a high value will result in most objects being merged into a single cluster.

9.2.2 Clustering Performance Evaluation

The rule of thumb for a threshold value of the minimum amount of neighbors within a radius ϵ, is *minPts* = dim$*$2 (Ester et al., 1996; Sander et al., 1998). However, if the dataset is large or contains noisy data, a larger value for *minPts* can be required. The threshold value was therefore estimated with the silhouette score and a method comparison study for the ϵ-estimation was performed to evaluate the difference between the Kneedle-algorithm and silhouette score on the quality of clustering. We examined the detection of 159 synthetically generated notes. The evaluated notes differ in duration, pitch and distance. The hit or miss criterion was based on whether a group of resonances representing a note was detected or not. We noticed that the silhouette score performed a significantly better clustering than the *kneedle* method. However, in both methods, due to the not-yet-perfect f_0 estimation, noise was clustered together and influences the results.

9.2.3 Clusters of noise

In this specific section, when we mention *noise*, we are referring to resonances that are detected along with the fundamental tone, even though they are not actually part of it. Most of them are classified as -1 by the DBSCAN algorithm, but closely laying noisy points form sometimes unwanted little clusters of overtones. However, they mostly have a relatively low power compared to the resonances presented in the real fundamental. Therefore, we removed the clusters containing a significantly lower power on average. The difference between a plot with and without power is shown in Figure 9.8. It is crucial to get rid of these clusters of noise for attributing constituents to a particular note in our system, as well as for the musical transcription of the sound.

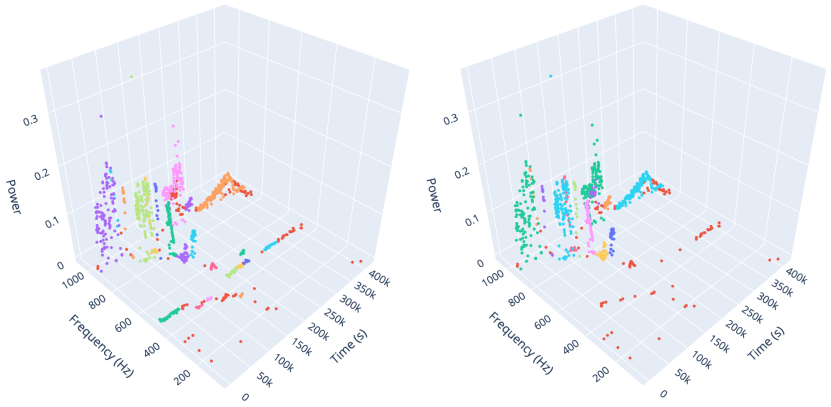

Figure 9.8: A comparison between two three-dimensional plots with a third axis denoting the power. Clusters with a small average power, relative to the other ones, are removed.

9.3 Attributing a Note Constituent

Resonances belonging to a cluster are each assigned to a constituent representing a musical *note*, attributed with the average frequency, onset, off-set and relative duration. Extracting the real rhythm of a live performance is more complicated than it seems, since people never play perfectly on the beat as written on paper. Therefore, we focused on the accurate extraction of pitch for this thesis, and suggested a relative estimation of the duration in our software. It correctly estimates the duration of notes for computer-generated audio files, but needs a refinement for audio files performed by humans in future work.

Example: Syrinx (Flute Solo)

We analyzed the pitch estimation of a slice of *Syrinx*, by Claude Debussy, performed by a real and artificially generated flute. The ground truth music score is presented in Figure 9.9. The first two measures have strong variations in rhythm and the notes are laying nearby, which makes is harder for the cluster algorithm to recognize two different objects. With the application of our method on this sample, the pitch of each note was correctly reconstructed based on the frequency data within each cluster. Please note that there is a slight difference in notation due to the key signature in the original piece.

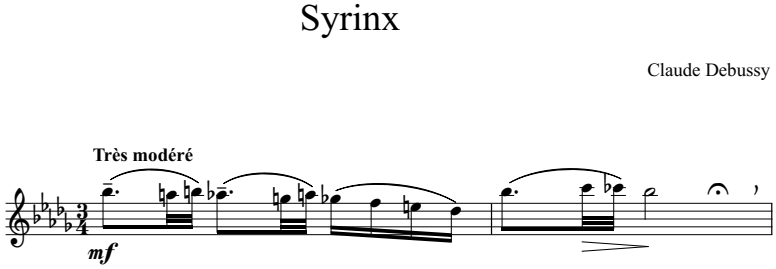

Figure 9.9: Western score notation of the first two measures of Syrinx, by Claude Debussy.

[4] About musical notation: The flat (♭), sharp (♯), and natural (♮) signs preceding a note indicate that the note should be played a semitone lower, higher, or disregarding the sign in the key signature or the modification of a tone within the same measure.

Thus, the pitch of each note in this slice has been assigned correctly[4], as illustrated in Figure 9.10.

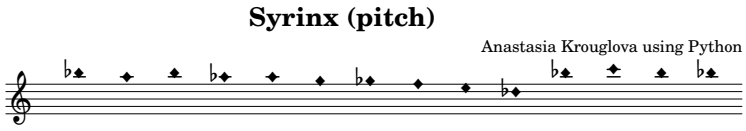

Figure 9.10: The pitches in the first measure of Syrinx, by Claude Debussy were correctly assigned and serve as one of the attributes for the Note constituent.

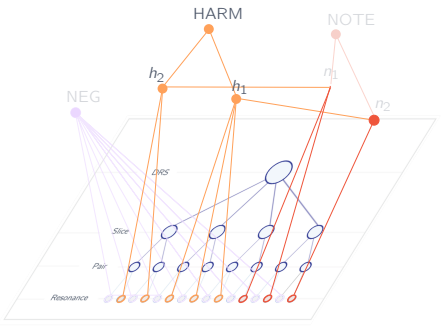

Figure 9.11: The hierarchy of harmonics is part of the Frame-level description of music, highlighted in purple.

9.4 Towards a Frame-level description

We introduce our approach towards a subtask of the frame-level description of music, namely the extractions of harmonic overtones. The frame-level description (i.e., multi-pitch estimation) estimates the number of notes that are simultaneously played in a slice of time (Benetos et al., 2019). It is currently based on the assumption that the fundamental tone is known, but should be expanded to a relative estimation and extraction of overtones for polyphonic music.

9.4.1 Attributing the Harmonic Constituent

As discussed in the chapter about psychoacoustics, a real musical tone often consists of a fundamental, but also from harmonic and non-harmonic overtones. Since harmonic overtones are integer multiples of the fundamental tones, they can be found by defining them in a space of the so-called f_0-likeliness (H):

$$H = E(\frac{f}{f_0} \mod 1) \tag{9.4}$$

$E \sim \mathcal{N}(0.5, 0.5)$ simulates the idea of entropy for the definition of how likely a resonance is part of an overtone. We used a real audio recording of a violin performing Canon in D, by Johann Pachelbel (Bridget, 2019) to analyze the overtones of a violin. We limited the space with a bound defined in the H space, which give us a first approximation of the overtones if the fundamental is known.

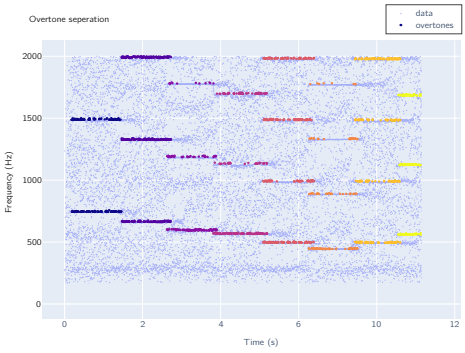

Figure 9.12: The extraction of the harmonic overtone from a real audio recording performed by Bridget (2019) (produced and approved for use by Stringspace).

Conclusion

The aim of this thesis was to create a multipurpose cognitive framework for musical analysis. We modelled human-like perception with the discrete resonance spectrum, grouped this information with cognitive models and structured the clusters in a type-based knowledge representation. We extracted musical structures from audio files and stored them in a hierarchical structure. This way, a bidirectional link between knowledge and data was obtained. Our methodology allows us to infer knowledge from different methods and build a system for a long-term perspective. Moreover, by using a cognitive clustering-approach, one of the fundamental music transcription challenges of overlapping tones has been resolved. To conclude, we created a tool that gives a range of possibilities in different subtasks of musical analysis and can push this research domain further by allowing inference between various machine learning models.

Further Work

For future work, we suggest to search for a cognitive approach towards a multi-pitch estimation of music, since the inference of musical attributes, such as pitch, in a polyphonic musical signal is a highly challenging problem (Benetos et al., 2019). Further, we also propose to improve the relative estimation of note durations which is currently present in our software, for a more accurate musical transcription of musical signals. It would also be interesting to extract less mainstream musical structures, such as phrases. For instance, they can be perceived as a musician gently inhaling before playing wind instruments, the subtle shift in bow direction for string instruments, or the relative power of notes in key instruments.

Nomenclature

Symbols

d_k	Initial complex amplitude of a resonance
ψ_k	Initial phase of a resonance
ϕ_k	Real frequency of the resonance
ω_k	Complex frequency of a resonance $(\phi_k + i\gamma_k)$
γ_k	Rate of decay
$\mathcal{F}[f(t)]$	Fourier transform of a time-domain signal
$\mathcal{F}^{-1}[\hat{f}(\omega)]$	Inverse Fourier transform of a frequency-domain signal
$\|\|f\|\|$	Norm of f
$\langle\alpha\|\beta\rangle$	Inner product
$\overline{f(x)}$	Complex conjugate
$P(i)$	Particles of a Constituent
$R(i)$	Attributes of a Constituent

Acronyms and Abbreviations

Cic	Calculus of Inductive Constructions
Coq	library with Calculus of Inductive Constructions as underlying formal language
CSPEC	Type of constituent specifications
CHAKRA	Common Hierarchical Abstract Knowledge Representation for Anything
CHARM	Common Hierarchical Abstract Representation of Music
DAG	Directed Acyclic Graph
DBSCAN	Density-based spatial clustering of applications with noise
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
L^2	Space of absolutely square-summable functions
STFT	Short Time Fourier Transform
VUB	Vrije Universiteit Brussel

Bibliography

Arvin, Farshad and Shyamala Doraisamy (2009). "Real-Time Pitch Extraction of Acoustical Signals Using Windowing Approach". In: *Australian Journal of Basic and Applied Sciences* 3, pp. 3557–3563.

Bacon, Sid P. et al. (1999). "Growth of simultaneous masking for fm<fs: Effects of overall frequency and level". In: *The Journal of the Acoustical Society of America* 106.1, pp. 341–350. doi: 10.1121/1.427060.

Baraniuk, Richard (2020). *8.2: Continuous Time Fourier Transform (CTFT)*.

Barr, Michael and Charles Wells (2012). *Category Theory for Computing Science*.

Belkic Karen, Dzevad Belkic (2019). *Signal Processing in Magnetic Resonance Spectroscopy with Biomedical Applications*. Boca Raton: CRC Press. isbn: 978-0-429-13019-9. doi: 10.1201/9781439806456.

Benetos, Emmanouil et al. (2019). "Automatic Music Transcription: An Overview". In: *IEEE Signal Processing Magazine* 36, pp. 20–30. doi: 10.1109/MSP.2018.2869928.

Bridget (2019). *Stringspace String Quartet & Jazz Band | Pachelbel Canon - Stringspace Solo Violin recording*.

Brownell, William E. (2017). "What Is Electromotility? -The History of Its Discovery and Its Relevance to Acoustics". In: *Acoustics today* 13.1, pp. 20–27.

Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander (2013). "Density-based clustering based on hierarchical density estimates". In: ed. by Jian Pei et al. Vol. 7819. ISSN: 1611-3349 Num Pages: 13. Berlin, Germany: Springer, pp. 160–172. doi: 10.1007/978-3-642-37456-2_14.

Chung, S. H., A. G. Pettigrew, and M. Anson (1981). "Hearing in the Frog: Dynamics of the Middle Ear". In: *Proceedings of the Royal Society of London. Series B, Biological Sciences* 212.1189. Publisher: The Royal Society, pp. 459–485.

Collins, Tom (2018). "Expressive Completeness Versus Structural Generality: Can a Single Music Representation Support Both?" In: Proceeding.

Cooley, James W. and John W. Tukey (1965). "An Algorithm for the Machine Calculation of Complex Fourier Series". In: *Mathematics of Computation*

19.90. Publisher: American Mathematical Society, pp. 297–301. doi: 10.2307/2003354.

Daubechies, Ingrid (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics. isbn: 978-0-89871-274-2. doi: 10.1137/1.9781611970104.

Donnelly, Patrick Joseph (2012). "Bayesian Approaches to Musical Instrument Classification Using Timbre Segmentation". PhD thesis.

Draguns, Andis et al. (2021). *Residual Shuffle-Exchange Networks for Fast Processing of Long Sequences*. arXiv:2004.04662 [cs, eess] version: 4. doi: 10.48550/arXiv.2004.04662.

Dubey, Kabir (2021). *The Fourier Uncertainty principles*. Tech. rep.

Eades, Harley (2012). *Type Theory and Applications*. Tech. rep.

Elbatta, Mohammed and Wesam Ashour (2013). "A dynamic Method for Discovering Density Varied Clusters". In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 6, pp. 123–134.

Elliott, Stephen J and Christopher A Shera (2012). "The cochlea as a smart structure". In: *Smart materials & structures* 21.6, p. 064001. doi: 10.1088/0964-1726/21/6/064001.

Ester, Martin et al. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, pp. 226–231.

Folland, Gerald B. and Alladi Sitaram (1997). "The uncertainty principle: A mathematical survey". In: *Journal of Fourier Analysis and Applications* 3.3, pp. 207–238. doi: 10.1007/BF02649110.

Gabor, D. (1947). "Acoustical Quanta and the Theory of Hearing". In: *Nature* 159.4044. Number: 4044 Publisher: Nature Publishing Group, pp. 591–594. doi: 10.1038/159591a0.

Goguen, Joseph A. (1991). "A categorical manifesto". In: *Mathematical Structures in Computer Science* 1.1, pp. 49–67. doi: 10.1017/S0960129500000050.

Harley, Nicholas (2020). "Abstract Representation of Music: A Type-Based Knowledge Representation Framework". Accepted. Thesis. Queen Mary University of London.

Harley, Nicholas (2022a). *Chakra*. https://github.com/nick-harley/Chakra.

Harley, Nicholas (2022b). *Charm*. https://github.com/nick-harley/Charm.

Harper, Robert (2011). "The Holy Trinity". In: *Published in the Existential Type weblog (cit. on pp.,)*

Hatfield, Gary (2015). "Objectifying the Phenomenal in Experimental Psychology: Titchener and Beyond". In: *Philosophia Scientiæ. Travaux d'histoire et de philosophie des sciences* 19-3. ISBN: 9782841747276 Number: 19-3 Publisher: Université Nancy 2, pp. 73–94. doi: 10.4000/philosophiascientiae.1133.

Hexmoor, Henry (2023). *Diffusion and Contagion | Elsevier Enhanced Reader*. doi: 10.1016/B978-0-12-800891-1.00006-8.

Hoang, Lê Nguyên (2014). *Type Theory: A Modern Computable Paradigm for Math*.

Holzapfel, AndrÉ and Yannis Stylianou (2008). "Musical Genre Classification Using Nonnegative Matrix Factorization-Based Features". In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.2, pp. 424–434. doi: 10.1109/TASL.2007.909434.

Homer, Steven T, Nicholas Harley, and Geraint A Wiggins (2023). "Modelling of Musical Perception using Spectral Knowledge Representation". In: *Journal of Cognition*. In press.

Hosch, L. William (2023). *Perception - Synthesis of constituent elements | Britannica*.

Huang, Wentao, Hongjian Sun, and Weijie Wang (2017). "Resonance-Based Sparse Signal Decomposition and Its Application in Mechanical Fault Diagnosis: A Review". In: *Sensors (Basel, Switzerland)* 17.6, p. 1279. doi: 10.3390/s17061279.

Inria (2018). *Calculus of Inductive Constructions — Coq 8.9.1 documentation*.

Kainulainen, Jouni and Matthias Maercker (2022). *Image Processing - Wavelet Transform II*.

Kellman, Philip J and Thomas F Shipley (1991). "A theory of visual interpolation in object perception". In: *Cognitive Psychology* 23.2, pp. 141–221. doi: 10.1016/0010-0285(91)90009-D.

Kennedy, Rodney A. and Parastoo Sadeghi (2013). "Hilbert Space Methods in Signal Processing". In: Edition: 1. Cambridge University Press. isbn: 978-0-511-84451-5. doi: 10.1017/CBO9780511844515.

Kim, Jinsook and Miseung Koo (2015). "Mass and Stiffness Impact on the Middle Ear and the Cochlear Partition". In: *Journal of Audiology & Otology* 19. doi: 10.7874/jao.2015.19.1.1.

Kosman, L. A. (1975). "Perceiving that We Perceive: On the Soul III, 2". In: *The Philosophical Review* 84.4. Publisher: [Duke University Press, Philosophical Review], pp. 499–519. doi: 10.2307/2183851.

Leinster, Tom (2016). *Basic Category Theory*. arXiv:1612.09375 [math]. doi: 10.48550/arXiv.1612.09375.

Lemarié, Pierre Gilles and Yves Meyer (1986). "Ondelettes et bases hilbertiennes." In: *Revista Matemática Iberoamericana* 2.1-2, pp. 1–18.

López-Serrano, Patricio et al. (2019). "NMF Toolbox: Music Processing Applications of Nonnegative Matrix Factorization". In: International Conference on Digital Audio Effects (DAFx-19).

LutzL (2009). *D4 Wavelet*.

Manley, Geoffrey A., Peter M. Narins, and Richard R. Fay (2012). "Experiments in comparative hearing: Georg von Békésy and beyond". In: *Hearing Research* 293.1-2, pp. 44–50. doi: 10.1016/j.heares.2012.04.013.

McAdams, Stephen (2019). "The Perceptual Representation of Timbre". In: *Timbre: Acoustics, Perception, and Cognition*. Ed. by Kai Siedenburg et al. Springer Handbook of Auditory Research. Cham: Springer International Publishing, pp. 23–57. isbn: 978-3-030-14832-4. doi: 10.1007/978-3-030-14832-4_2.

Neutelings, Izaak (2021a). *Fourier series & synthesis*.

Neutelings, Izaak (2021b). *Harmonic oscillator plots*.

O'Brien, Daniel (2023). *Epistemology of Perception, The | Internet Encyclopedia of Philosophy*. The University of Birmingham.

Offermans, Wil (2023). *For the Contemporary Flutist | Difference Tones*. Awarded by the Newly Published Music Committee of the National Flute Association, USA in 1993. Zimmermann Verlag.

Oppenheim, Jacob N. and Marcelo O. Magnasco (2013). "Human Time-Frequency Acuity Beats the Fourier Uncertainty Principle". In: *Physical Review Letters* 110.4. Publisher: American Physical Society, p. 044301. doi: 10.1103/PhysRevLett.110.044301.

Pearce, Marcus (2005). "The construction and evaluation of statistical models of melodic structure in music perception and composition". In: *City University, London*.

Rameau, Jean-Philippe) (1722). *Traité de l'harmonie réduite à ses principes naturels*. First Edition. Paris. isbn: 978-84-86230-06-7.

Romberg, J (2016). *ECE 6250, Fall 2016, Notes – Justin Romberg*.

Rousseeuw, Peter J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. doi: 10.1016/0377-0427(87)90125-7.

Russel, Herman (2021). *9.2: Complex Exponential Fourier Series*.

Sander, Jörg et al. (1998). "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications". In: *Data Mining and Knowledge Discovery* 2.2, pp. 169–194. doi: 10.1023/A:1009745219419.

Satopaa, Ville et al. (2011). "Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior". In: *2011 31st International Conference on Distributed Computing Systems Workshops*. Minneapolis, MN, USA: IEEE, pp. 166–171. isbn: 978-1-4577-0384-3. doi: 10.1109/ICDCSW.2011.20.

Schneider, Peter et al. (2005). "Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference". In: *Nature Neuroscience* 8.9. Number: 9 Publisher: Nature Publishing Group, pp. 1241–1247. doi: 10.1038/nn1530.

Schubert, Erich et al. (2017). "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN". In: *ACM Transactions on Database Systems* 42.3, pp. 1–21. doi: 10.1145/3068335.

Shi, Zhongzhi (2019). "Cognitive Machine Learning". In: *International Journal of Intelligence Science* 09, pp. 111–121. doi: 10.4236/ijis.2019.94007.

Sleep, Jonathan (2017). "Automatic Music Transcription with Convolutional Neural Networks using Intuitive Filter Shapes". In: *Master's Theses*. doi: 10.15368/theses.2017.95.

Smaill, Alan, Geraint Wiggins, and Mitch Harris (1993). "Hierarchical Music Representation for Composition and Analysis". In: *Computers and the Humanities* 27.1. Publisher: Springer, pp. 7–17.

Smith, James C. et al. (1978). "Human Auditory Frequency-Following Responses to a Missing Fundamental". In: *Science* 201.4356. Publisher: American Association for the Advancement of Science, pp. 639–641.

Smith, Steven (1997). *The Scientist and Engineer's Guide to Digital Signal Processing by Steven W. Smith*. isbn: ISBN 0-9660176-3-3.

Temperley, David (2004). "Bayesian Models of Musical Structure and Cognition". In: *Musicae Scientiae* 8, pp. 175–205. doi: 10.1177/102986490400800204.

Theodor, Popescu (1997). "Time-frequency analysis, by L. Cohen, Prentice Hall Signal Processing Series, Prentice Hall, Englewood Cliffs, New Jersey, 1995 - Book review". In: *Control Engineering Practice* 5, pp. 292–294. doi: 10.1016/S0967-0661(97)90028-9.

Van Dongen, Stijn (2008). "Graph Clustering Via a Discrete Uncoupling Process". In: *SIAM Journal on Matrix Analysis and Applications* 30.1, pp. 121–141. doi: 10.1137/040608635.

Vavakou, Anna, Nigel P Cooper, and Marcel van der Heijden (2019). "The frequency limit of outer hair cell motility measured in vivo". In: *eLife* 8. Ed. by Andrew J King et al. Publisher: eLife Sciences Publications, Ltd, e47667. doi: 10.7554/eLife.47667.

Wegel, R. L. and C. E. Lane (1924). "The Auditory Masking of One Pure Tone by Another and its Probable Relation to the Dynamics of the Inner Ear". In: *Physical Review* 23.2. Publisher: American Physical Society, pp. 266–285. doi: 10.1103/PhysRev.23.266.

Wiggins, G. A., M. Harris, and A. Smaill (1989). *Representing Music for Analysis and Composition*.

Wiggins, Geraint A. (2020). "Creativity, information, and consciousness: The information dynamics of thinking". In: *Physics of Life Reviews* 34-35, pp. 1–39. doi: 10.1016/j.plrev.2018.05.001.

Yeturu, Kalidas (2020). "Chapter 3 - Machine learning algorithms, applications, and practices in data science". In: *Handbook of Statistics*. Ed. by Arni S. R. Srinivasa Rao and C. R. Rao. Vol. 43. Principles and Methods for Data Science. Elsevier, pp. 81–206. doi: 10.1016/bs.host.2020.01.002.

Young, R. W. (1954). "Inharmonicity of piano strings". In: *Acta Acustica united with Acustica* 4.1, pp. 259–262.

Zatorre, Robert J. (2005). "Finding the missing fundamental". In: *Nature* 436.7054. Number: 7054 Publisher: Nature Publishing Group, pp. 1093–1094. doi: 10.1038/4361093a.