POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Visualizing Music and Emotions throughout Metatron: from Audio Analysis to Music Generation

Laurea Magistrale in Music and Acoustic Engineering

Author: Alberto Di Maria

Advisor: Prof. Massimiliano Zanoni

Co-advisor: Marco Accardi

Academic year: 2023-2024

1. Introduction

The evolving landscape of audio analysis and music creation highlights significant limitations in traditional methods, particularly in capturing the nuanced relationship between music and emotions. Conventional approaches often rely on two-dimensional plots of valence and arousal—dimensions representing emotional valence (positive or negative) and arousal (level of activity or excitement). These methods are inadequate for fully representing emotions [2], as they confine visualization to a two-dimensional plane, where features like color define clusters based on instruments, and genre predominates the filtering process.

Text-to-music models face similar limitations. These models often operate with basic interfaces that prioritize functionality over user experience. They present challenging parameters for average users, resulting in a lack of immersive, user-centric design. Additionally, users must write detailed prompts to describe the desired audio sample, which can be daunting for those with limited musical knowledge. This often leads to generic descriptions that inhibit creative exploration, constraining innovation in the generation process by traditional patterns of describing music, such as genre, mood, and instrument.

To address these limitations, this thesis aims to develop an innovative system (Figure 1) that enhances the relationship between music and emotions through advanced visualization techniques, AI-driven music generation, and the integration of sacred geometry. Sacred geometry, an ancient practice using geometric shapes to represent universal patterns, is central to this thesis and provides a novel way to map emotions to audio features.

The project introduces three interconnected clients, each addressing specific limitations. The Latent Space Client (LSC) transitions from traditional 2D plots to a highly customizable 3D space, enabling the representation of space-mapped audio features (Figure 2). The Metatron Client (MC) employs sacred geometry, such as Metatron and Platonic solids, to innovatively map emotions to audio features (Figure 3). Lastly, the Generative Client (GC) uses a text-to-music model to create music based on visual placeholders from the MC, eliminating the need for user-written prompts and enhancing user interaction through visually-driven controls.

This paper is organized as follows: Section 2 ex-

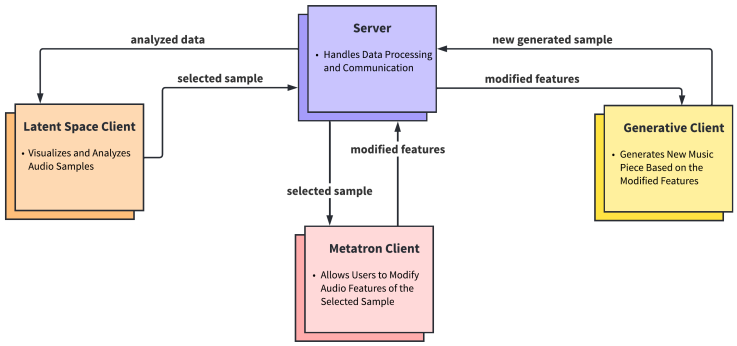


Figure 1: From audio analysis to music generation: system architecture.

plores the detailed functionality and implementation of the Latent Space Client (LSC), explaining how it visualizes audio features in a three-dimensional space. Section 3 focuses on the Metatron Client (MC), describing its use of sacred geometry to create an intuitive, emotion-based mapping of audio features. Section 4 discusses the Generative Client (GC) and its innovative approach to music generation using AI and visual placeholders. Section 5 covers the integration and methodology, detailing how the three clients work together to form a cohesive system. Finally, Section **??** presents the conclusions, summarizing the contributions of the thesis and discussing potential avenues for future research.

2. Latent Space Client

The Latent Space Client (LSC) is designed to overcome the limitations of traditional two-dimensional audio visualizations by transitioning to a customizable three-dimensional space. Utilizing React 3 Fiber [3], a 3D renderer based on Three.js, the LSC dynamically and interactively represents audio features. The cornerstone of this client is its real-time responsiveness; each modification made to the visualization is reflected immediately, providing instant feedback to users.

Moving away from genre-based classifications and focusing on mood-based analysis, the LSC provides a flexible and comprehensive understanding of audio data. Users can customize the axes with more than ten features, accurately mapped into the 3D space. This allows user to tailor the visualization to their specific analytical needs and enhancing both depth and flexibility in audio analysis.

The innovative filtering system in the LSC further enhances its functionality. Users can apply filters to individual features and perform text searches for desired features or sample names, making audio dataset exploration more intuitive and efficient. This combination of customizable axes and advanced filtering capabilities sets a new standard for interactive and immersive audio analysis, offering a visually appealing and functionally rich interface.

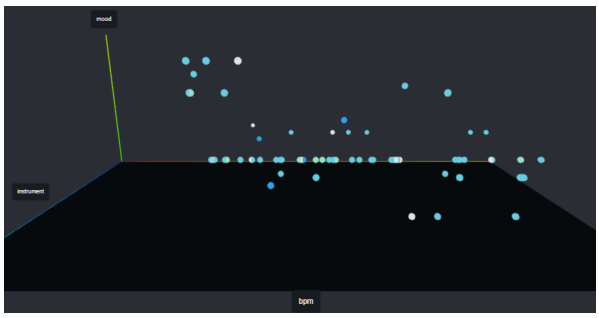


Figure 2: 3D visualization of audio samples in Latent Space Client. Each sphere represents an audio sample positioned according to its characteristics.

3. Metatron Client

The Metatron Client (MC) utilizes sacred geometry to provide an innovative and intuitive approach to audio visualization. Sacred geometry, such as Metatron and Platonic solids, is used to create a unique mapping of emotions to audio features, offering users a novel way to explore the relationship between music and emotions.

Specifically, the MC incorporates Metatron—a complex geometric figure containing all Platonic solids—and the Flower of Life—a pattern of

overlapping circles symbolizing life's interconnection. In the MC, each vertex of Metatron corresponds to two circles representing different audio features. The primary features, such as BPM, mood, and danceability, define the Platonic solid displayed, while the other nine features affect line characteristics. This approach allows users to interactively manipulate these features, gaining real-time visual feedback that enhances their understanding of the emotional dimensions of music.

The MC uses the following Platonic solids, each associated with a specific range of audio features due to its musical attributes:

- Cube: represents the lower spectrum, associated with slower or more stable musical attributes (see Figure 4a).
- Dodecahedron: captures a slightly more complex range, indicative of evolving musical patterns (see Figure 4b).
- Octahedron: sits in the middle range, symbolizing a balance in musical dynamics (see Figure 4c).
- Tetrahedron: represents higher energy and more dynamic attributes of music (see Figure 4d).
- Icosahedron: reflects the highest intensity and complexity in music characteristics (see Figure 4e).

These solids are integrated into the client interface and are linked to interactive knobs, represented by circles at each vertex. Users can adjust these knobs to manipulate the corresponding audio features, with changes reflected in real-time through geometric transformations of the displayed solids.

A key aspect of the MC is its label-free interface, which encourages users to explore and intuitively understand the system without relying on predefined labels. This design choice promotes deeper engagement with the visualization process, as users experiment with different configurations to achieve aesthetically satisfying results. By integrating these ancient geometric principles, the MC provides a holistic and immersive visualization experience.

The MC is seamlessly connected to the Latent Space Client (LSC), allowing users to select audio samples from the LSC and further modify their features using the MC's interactive circles. These modifications are then passed to the Generative Client (GC), enabling a continuous and integrated workflow across the three clients.

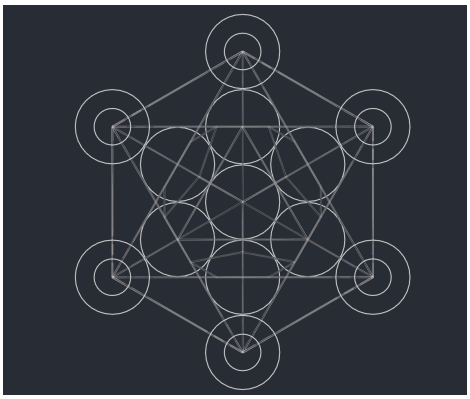
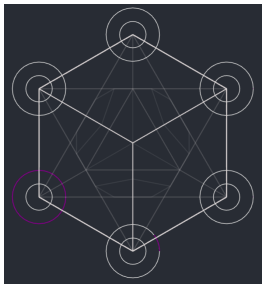
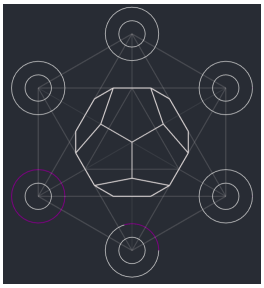


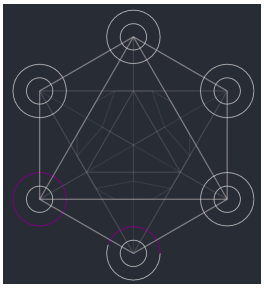
Figure 3: Metatron Client GUI, including Metatron and Flower of Life.



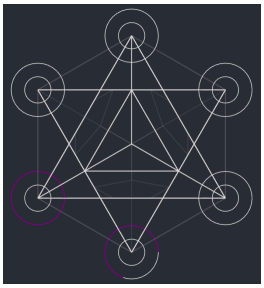
(a) Cube.



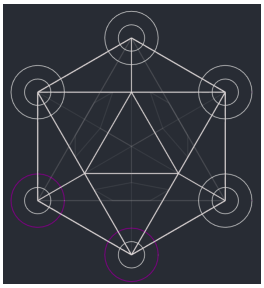
(b) Dodecahedron.



(c) Octahedron.



(d) Tetrahedron.



(e) Icosahedron.

Figure 4: Visual representations of Metatron and the Platonic solids.

4. Generative Client

The Generative Client (GC) introduces a groundbreaking approach to music creation using advanced AI technologies. Unlike traditional use of text-to-music models that require user-written prompts, the GC utilizes visual placeholders derived from the Metatron Client (MC) to generate music. This approach shifts the focus from text-based inputs to visually-driven controls, making the music creation process more accessible and engaging for users.
The core of the GC is the Suno API [4], representing the latest advances in text-to-music generation. By connecting this AI model to the visual elements provided by the MC, the GC allows users to create music by adjusting visual placeholders rather than writing descriptive prompts. This method ensures that the generative process is intuitive and accessible, even for users with limited musical knowledge.
The GC focuses on both analytical audio features, like bpm and danceability, and meta-descriptors, such as clarity and texture. Users can modify these attributes through the GC's interactive interface, which immediately reflects the changes in the generated music. This visually-driven interaction enhances the creative process, making it more engaging and less reliant on technical expertise.
By integrating the Generative Client with the Latent Space Client (LSC) and the Metatron Client (MC), the system creates a seamless workflow that continuously enhances and re-analyzes music samples. This circular process offers users a unique and immersive experience.

4.1. Evaluation and Comparison of Generative Models

A critical part of this project was the detailed analysis of various generative models to identify the most suitable technology for the Generative Client (GC). The motivation behind this comprehensive analysis was to ensure that the chosen model not only provided high-quality music generation but also integrated seamlessly with the visual and interactive components of the system.
Before choosing Suno, the research involved a thorough evaluation of multiple models, each with unique features and challenges. The models analyzed included mostly text-to-music and Variational AutoEncoders (VAE), each scrutinized for their ability to generate high quality music samples, ease of implementation, generation time and computational requirements.
In particular, the evaluation parameters are:

- sample quality: the ability of the model to generate high-fidelity audio that is musically coherent and aesthetically pleasing.
- sample length: the capacity of the model to generate longer audio tracks rather than short samples, which is essential for creating entire music tracks as required by the project.
- integration ease: how straightforward it is to implement and operate the model, considering clear instructions and reasonable computational power requirements. The model should work efficiently on an average good laptop without a high-end GPU.
- generation time: the time taken by the model to generate music, ensuring it is practical for real-time or near-real-time applications.

The motivation for this analysis was to equip the project with the most advanced and suitable technology available, ensuring that the generative process could be both high-quality and efficient. Additionally, the detailed comparison and evaluation of these models provide valuable insights about the current state of art in generative music, useful for future research and development in the field.
The evaluation identified Suno as the most suitable model for high-quality music generation, offering a good balance between quality and performance and the flexibility required by the project. Tables 1 and 2 show the most interesting models analyzed and their respective performances.

5. Methodology

The integration of the Latent Space Client (LSC), the Metatron Client (MC) and the Generative Client (GC) forms a cohesive system that significantly enhances the overall user experience in audio analysis and music creation. These clients operate as interconnected modules managed by a central server, ensuring seamless interaction and data flow.
The backend of this system consists of a Python server utilizing Essentia [1], a comprehensive li-

Model	Sample Quality	Sample Length	Integration Ease	Generation Time
MusicVAE	Low	Short	Easy	Low
RAVE	High	Long	Difficult	Very Low

Table 1: VAEs models comparison.

Model	Sample Quality	Sample Length	Integration Ease	Generation Time
Suno	High	Long	Easy	Low
Stable Audio	High	Long	Difficult	Low
Jukebox	High	Long	Difficult	Long
MusicGen	High	Short	Medium	Long

Table 2: Text-to-Music models comparison.

brary for audio analysis and feature extraction. The server architecture ensures efficient data management and real-time processing, enabling users to interact with the system without noticeable delays.
The data flow between the clients is facilitated by Flask-SocketIO, a technology that supports real-time communication. This setup allows for sequential data processing from the LSC to the MC and finally to the GC. Each client performs its specific tasks and passes the modified data to the next client, creating a continuous loop of interaction and enhancement. For instance, new music samples generated by the GC are reanalyzed by the LSC, ensuring that the visualization and generative processes are dynamically linked.

6. Conclusions

This project introduces a novel system for audio analysis and music creation, enhancing the user experience through advanced visualization techniques and AI-driven generation.
The LSC introduces a customizable 3D space for mood-based analysis, moving beyond traditional genre classification. The MC employs Metatron and Platonic solids, demonstrating that it is possible to map emotions to audio features using sacred geometry. The GC utilizes the latest AI technologies to generate music based on visual placeholders, eliminating the need for written prompts and simplifying the user experience.
Additionally, a comprehensive analysis of various generative models was conducted to ensure that the chosen model, Suno, provided the best balance of quality and performance while integrating with the system. This analysis also aims to serve as a valuable resource for future research, offering a detailed understanding of the state of the art in generative music models.
This work serves as a starting point for future research, particularly focusing on further exploring how Metatron can be utilized in audio analysis and creation. The groundwork laid here provides a robust foundation for future developments, advancing the state of the art in user-centric audio analysis and music creation.
The key contributions of this thesis include the integration of sacred geometry in music visualization, the development of an interactive and user-friendly music generation system, and the detailed evaluation of generative models for high-quality music production. These innovations significantly enhance the relationship between music and emotions, providing a new direction for future research and applications in this field.

References

[1] Dmitry Bogdanov, Nils Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, Juan R Zapata, and Xavier Serra. Essentia: An audio analysis library for music information retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 493–498, 2013.

[2] Johnny R. J. Fontaine, Klaus R. Scherer,

Etienne B. Roesch, and Phoebe C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057, 2007.

[3] Paul Henschel and contributors. React three fiber. `https://github.com/pmndrs/react-three-fiber`.

[4] Suno. Suno model. `https://suno.ai/`, 2024.