

Music Discovery Through Visualisation of Audio Features

Introducing SoundShapes

Thesis, submitted in partial fulfilment of the requirements for the degree of
Master of Science in Applied Data Science.

Abstract

Popular music streaming platforms do not typically display any audio-derived information other than track duration. This limits users' ability to guide their own exploration of music beyond listening to each song recommended by the algorithm or making decisions based on their knowledge of the artist or other metadata. This work proposes visual thumbnails generated from audio features to facilitate music discovery: SoundShapes.

SoundShapes visualise the mood and timbre characteristics of the audio. The valence-arousal scale approximates mood. Timbre is represented by the instruments used and an abstract representation of the genre. Various extraction methods were used to facilitate feature extraction, including low-level signal processing and the use of a convolutional neural network.

A prototype of a user interface was built and used to evaluate the SoundShapes. Although the sample was small and not representative of a wider user population, the results of the evaluation show potential for wider user acceptance.

Author	Beryl van Gelderen
Student ID	4868846
Public defence	Friday June 28, 2024
Version	July 7, 2024
Thesis supervisor	Dr. D. Gauthier
Thesis committee member	Dr. D. Nguyen

An electronic version of this thesis will be made available at
studenttheses.uu.nl.

Graduate School of Natural Sciences
Faculty of Science



**Utrecht
University**

Acknowledgements

The process of creating this thesis has been much more enjoyable and interesting than I ever expected. This is because of David, my supervisor, who provided me with the opportunity to undertake a thesis on music, and the freedom to formulate my own research question. David, thank you. I am very appreciative of your advice and guidance, which have made this project so much more worthwhile.

Thank you to everyone who has helped to evaluate the SoundShapes. This project would not have been possible without you.

Jan, thank you for being the ultimate study buddy. I can not even imagine this masters degree without you.

Vera, thank you for being so supportive, generous and enthusiastic throughout writing this thesis.

Speaking of amazing friends: Fieke, thank you for your encouragement, hospitality at your office and the proofreading. Lisanne, thank you for your advise on how to document a design process and your loving support. Roos, thank you for the encouraging messages and for your feedback on the structure of the design process. Marijn, thank you for hosting the wonderful writing retreat in France with bonus programming and music recommendations at dinner.

Thank you to my colleagues at PRIME, for managing my projects while I finish up my degree, and especially Abel and Rens for taking such an interest in this project.

Thank you to my parents, you have unconditionally supported me through my non-linear educational and career path.

Last but certainly not least, I want to thank Thierry. You have been my rock this whole year, and to be fair, for the better part of a decade as well. I am beyond grateful that you are in my life.

*Beryl van Gelderen
The Hague, July 2024*

Contents

1	Introduction	1
2	Background	3
2.1	Audio Features	3
2.2	Deep Learning.	4
2.3	Music Perception	4
2.4	Feature Selection for Similarity	4
2.5	Music Visualisation.	4
3	Conceptualisation	7
3.1	Use Scenario	7
3.2	Feature Selection.	8
3.3	Visualisation Approach	8
3.3.1	Mood Visualisation	8
3.3.2	Timbre Visualisation	11
3.4	Chosen Visualisation	11
3.4.1	Mood: Valence as Background Colour	11
3.4.2	Mood: Arousal as Line Width of the Shape	12
3.4.3	Timbre: Instruments as a Radar Chart Shape.	12
3.4.4	Timbre: Genre as Colour of the Shape	13
3.4.5	Introducing SoundShapes	14
4	Method	17
4.1	Music Dataset	17
4.1.1	Ethical and legal consideration of the data	19
4.2	Feature Extraction	19
4.2.1	Feature Extraction of Mood: Arousal	19
4.2.2	Feature Extraction of Mood: Valence	20
4.2.3	Feature Extraction of Timbre.	20
4.3	Feature Mapping to Visualisation.	21
4.3.1	From Valence Score to Background Colour	21
4.3.2	From Arousal Score to Line Width	21
4.3.3	From Instruments to Radar Chart Shape	21
4.3.4	From Genres to Shape Colour	22
4.4	Application Design	24
5	Visualisation Results	25
6	User Evaluation	26
6.1	Task to be Evaluated	26
6.2	Participants.	26

6.3	Procedure	26
6.4	Software	27
6.5	Application Design	28
6.6	Survey	28
7	Evaluation Results	29
7.1	System Usability Score	29
7.2	Open-ended Questions	30
8	Discussion & Conclusion	33
8.1	Discussion	33
8.2	Conclusion	34
	References	35
	Appendices	37
A	Code	38
B	User Evaluation Procedure	39
B.1	User Evaluation Protocol for the Researcher	39
B.2	User Evaluation Application Screenshots	41
B.3	User Evaluation Survey Screenshots	45
C	User Evaluation Results	51

1

Introduction

When browsing through popular streaming platforms, it is evident that they do not provide much information about the audio characteristics contained in the music. We see some metadata, such as the album cover, the name of the track and the artist, as shown in the screenshots in figure 1.1. But there is no information about the music itself (audio features or audio-derived features) of any of the tracks, except for the duration of each track, displayed in the album view.

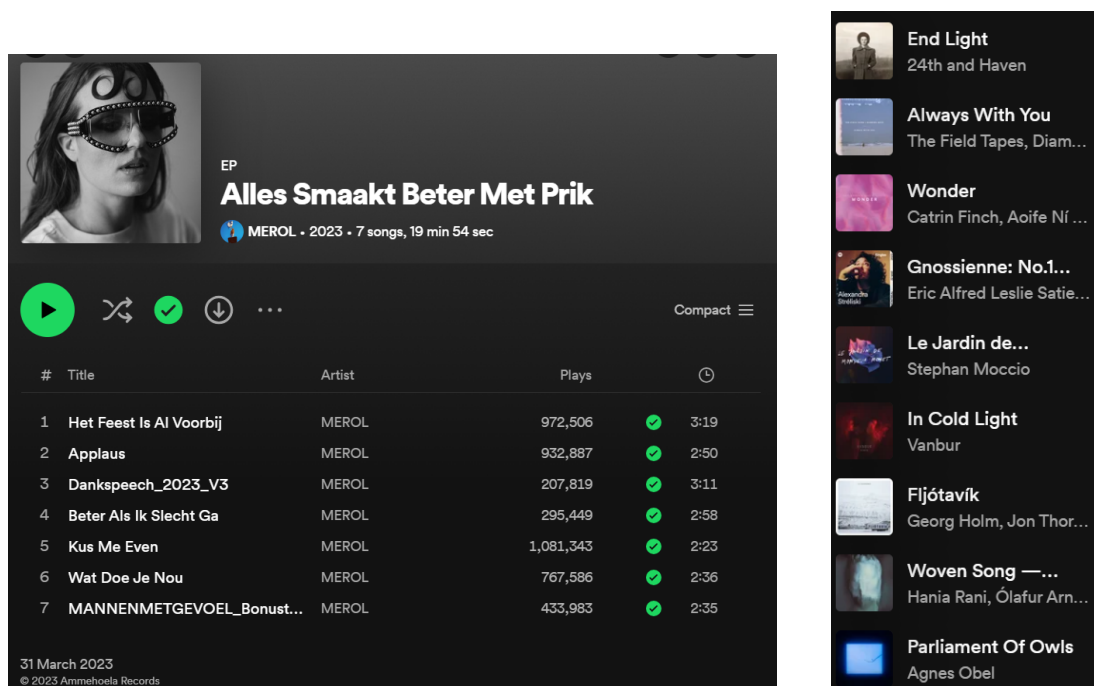


Figure 1.1: Screenshots from a popular streaming platform, Spotify, showing (left) an album view and (right) playlists. No information about the audio features is shown, except for the duration of the track in the album view.

The lack of visible cues related to the audio track is understandable, as most music recommendation systems are strongly dependent on pre-defined

metadata such as artist biographies and genre annotations, and the behaviour of other users in how they listen to tracks and group tracks together in playlists Velankar and Kulkarni, 2023.

Not showing audio features is a missed opportunity, as withholding the knowledge of the audio features of tracks also deprives users of the ability to guide their own exploration of songs, without having to listen to every song the algorithm recommends or basing their decision on pre-acquired knowledge of the artist or album.

In order to remedy this, I explore the automatic creation of visual thumbnails that are generated solely based on the audio track to improve the browsing and searching experience.

Research Question

How do users perceive the effectiveness of visual thumbnail representations of audio-based features in aiding the discovery of similar music tracks?

In order to evaluate the SoundShapes, I explore the user perception of the visualisation in a prototype interface.

The structure of this thesis is as follows. Chapter 2, Background, presents the current state of the art in the relevant literature. The following chapter 3, Conceptualisation, is an account of the visualisation design process. Chapter 4, Method, describes the process of transforming an mp3 to its visual representation. Chapter 5 Visualisation Results, presents the results, SoundShapes. Chapter 6, User Evaluation, describes the process of evaluating the SoundShapes, while 7, Evaluation Results, presents the results from this evaluation. Finally, chapter 8, Discussion & Conclusion, provides an interpretation of the evaluation results and implications for future research.

2

Background

The research area in which this thesis is situated is called Music Information Retrieval (MIR). MIR encompasses research topics related to the modelling and understanding of music (Srinivasa Murthy & Koolagudi, 2018). In addition to MIR, this research intersects with Information Visualisation (InfoVis) and Human-Computer Interaction (HCI). InfoVis involves the leveraging of graphical representations of data to aid in exploring, understanding, and analysing information (Liu et al., 2014). HCI focuses on the design and use of computer technology, emphasising the interfaces between people (users) and computers (Katona, 2021; MacKenzie, 2024).

The state of the art in these areas of research, where relevant to our research question, is described in this chapter.

2.1. Audio Features

Features retrieved from audio are commonly divided into three main categories: low-level, mid-level and high-level features (Afchar et al., 2022; Srinivasa Murthy & Koolagudi, 2018).

Low-level features are extracted using signal processing techniques and are hard for humans to understand. Examples include the MFCC frequency cepstrum, amplitude, zero-crossing rates, beats, and so on. They do not represent the characteristics of the entire track. Mid-level features give characteristics about the whole track or at least about a set or segment. Such features can be about pitch, tempo, rhythm, and harmony.

High-level features are the most human-understandable. High level features are things like danceability, mood and instruments used. Often, they are not solely derived from audio tracks but manually annotated.

The challenge in MIR is in bridging the gap between low-level, machine-understandable data and high-level, human-centric musical concepts, which is sometimes referred to as the “semantic gap”. Approaches to narrowing this gap include the integration of contextual metadata, such as user tags or web-mined information, and listener feedback (Fu et al., 2011).

2.2. Deep Learning

In addition to the direct analysis of audio signals, it is important to recognise the potential of deep learning for content-based music tagging. In tasks relating to music classification and tagging, the best results have been achieved using these methods (Deldjoo et al., 2024).

2.3. Music Perception

Listener perception is difficult to quantify. Listeners perceive songs differently from person to person. For example, for some people the genre is more important than the lyrics, for others the lyrics are more important.

The manner in which music is perceived is not solely dependent on the listener, but also varies according to the context of the listener. The context may encompass a multitude of factors, including the listener's mood, their location, the presence of other listeners, the listener's activities while listening, and so forth. To illustrate, in one instance, a set of tracks may convey the listener's current mood and are thus appropriate. However, in another instance, despite the listener having a similar mood, some of the same tracks are unsuitable. This is due to the listener's partner being present, who dislikes some of the genres.

This implies that the same set of music may be perceived as appropriate and therefore **similar** in one situation, but in another situation some of the same tracks may be inappropriate.

2.4. Feature Selection for Similarity

For an intuitive visualisation design, similar-looking thumbnails should depict similar-sounding music. Identifying which audio features correlate with perceived music similarity is essential in answering our research question, yet there is no consensus in existing literature.

Even with the hypothetical complete understanding and measurement of every acoustic element in music—including pitch, rhythm, dynamics, and timbre, as well as their temporal evolution—a purely signal-centric perspective remains inadequate to model perceived similarity (Knees & Schedl, 2016; Wiggins, 2009).

I acknowledge that a universal set of features to perfectly model music similarity does not exist. Nonetheless, for the purpose of designing the visualisations, and leveraging my musical background, I propose an intuitively selected subset of features, accepting the subjective nature of musical similarity.

2.5. Music Visualisation

The most recent literature review the domain of music visualisations, by Lima et al., 2022, centres on visualising the temporal dynamics within music tracks rather than focusing on a single static icon per track, which is the goal of this project. Therefore, this study is only partially applicable to this project. However, the authors highlight the innovation in integrating music visualisations with Information Visualisation (InfoVis) techniques, including line graphs,

glyphs, and colours. They also anticipate that advancements in feature extraction methods will lead to more sophisticated visualisations in the future.

When looking at visualisations that just have one visualisation per music track, I find a few examples in the literature. They are described below.

An inspiration for this project is the work by Kim et al., 2009. They implemented music icons according to the emotion plane, as shown in figure 2.1. While the music icons look beautiful, they lack some expressiveness. Other implementations of music thumbnails that serve as an example for this project are those by Hoveling, 2023, which visualises highly abstract data from Neural Network layers, seen in figure 2.2. They were in turn inspired by the file icons by Kolhoff et al., 2008, which are generated by machine learning, changing pre-designed pictures according to features extracted from music files, as shown in figure 2.3. Another inspiration is Thumbnail DJ by Chen and Klüber, 2010, that visualises audio features for DJs, in figure 2.4.



Figure 2.1: Music thumbnails in the emotion plane by Kim et al., 2009.



Figure 2.2: Musicons by Hoveling, 2023.

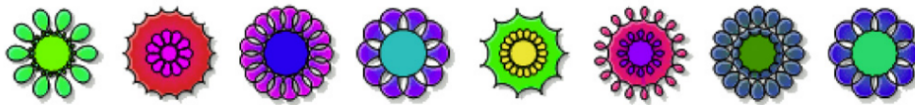


Figure 2.3: Music file thumbnails by Kolhoff et al., 2008.

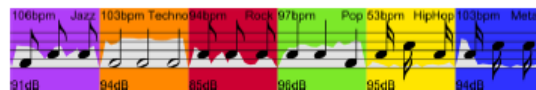


Figure 2.4: Thumbnail DJ examples by Chen and Klüber, 2010.

3

Conceptualisation

This chapter outlines the conceptualisation process for the visualisation design.

To facilitate an intuitive design, similar sounding music should be represented with similar-looking visualisations. In order to clarify the subjective concept of music similarity, a use case is needed for the creation and testing of the music thumbnails. The audio features to be visualised are then chosen with this use case in mind.

The use case, the audio features to visualise and the approach to the visualisation are described in this chapter, as well as the design that was chosen to implement and explore with test users.

3.1. Use Scenario

The thumbnails are used to support an exploratory search by active listeners, aimed at creating background music playlists from a seed song. These playlists are unordered and could be created a variety of contexts, such as:

- social gatherings with drinks
- dinner parties
- focused activities like studying or meditation
- group exercises, like runs with music on speakers

Why active listeners? Active listeners are defined as those who actively seek new music. They do not necessarily have knowledge about music theory.

Why an exploratory search? Thumbnails should facilitate the discovery of music unfamiliar to the listener through visual aids.

Why a seed song? A seed song establishes the initial tone and mood for the playlist, guiding track selection. By displaying the seed song's thumbnail, we can explore if thumbnails aid in identifying suitable songs.

Why an unordered playlist? Seamless song transitions are facilitated by harmonically complimentary tonics and similar tempos. These features fall outside the scope of this project.

Why music in the background? When music is the main focus, like in a concert, context like the artists' biography, the collaborations and the meaning of the lyrics become more significant. As this project aims to solely take into

account the audio features, a background setting is a more appropriate setting to explore this.

3.2. Feature Selection

It is possible to visualise music in a completely abstract way, by visualising concepts that are not high-level features, through the use of a convolutional neural network. This approach was taken by Hoveling, 2023 and Kolhoff et al., 2008. A criticism on their work is that their thumbnails are difficult for users to comprehend, which has the effect of reducing their acceptance. In this project, the decision has been made to create visualisations of high-level features that can be understood by humans.

Considering the use scenario from the previous section, and recognising that choosing features for song similarity can be subjective as described in section 2.4, two high-level features stand out as especially significant for the use scenario: mood and timbre.

For mood, arousal and valence are introduced as proxies, corresponding to the axes of the circumplex model of affect (Russell, 1980). According to this model, all moods can be described with just these two dimensions. This is important for curating playlists that resonate with the intended atmosphere of group settings.

The other feature is timbre. Timbre, sometimes called sound colour or tone quality, describes the characteristic that allows us to distinguish between different sounds when they have the same pitch and loudness. It is what makes a violin sound different from a trumpet, even when they play the same notes. Large changes in timbre are distracting and take away from the cohesion of a playlist, making it less appropriate for a background setting (Bonnin & Jannach, 2014). Therefore, this feature is an important one to visualise as well.

Timbre can be approximated in many ways. The most human-understandable approximations usually take into account either instrumentation or genre (J.-J. Aucouturier et al., 2005)

3.3. Visualisation Approach

This study aims to create a single thumbnail for each track. As an initial guide, Mackinlay's ranking of Encodings was used, succinctly summarised and visualised by Munzner, 2014, shown here in figure 3.1. During the process, other means of visualisation, not in this ranking, were considered as well. Another initial point of inspiration were the music visualisation flowers by Kolhoff et al., 2008.

3.3.1. Mood Visualisation

According to Mackinlay, positioning on a common or unaligned scale, as seen in figure 3.1, is the most or second most effective means of encoding data (Munzner, 2014). However, mapping the arousal-valence scale to unaligned axes, resulted in a visualisation that was difficult to interpret at a glance (see Figure 3.2). Another downside to showing arousal and valence on unaligned

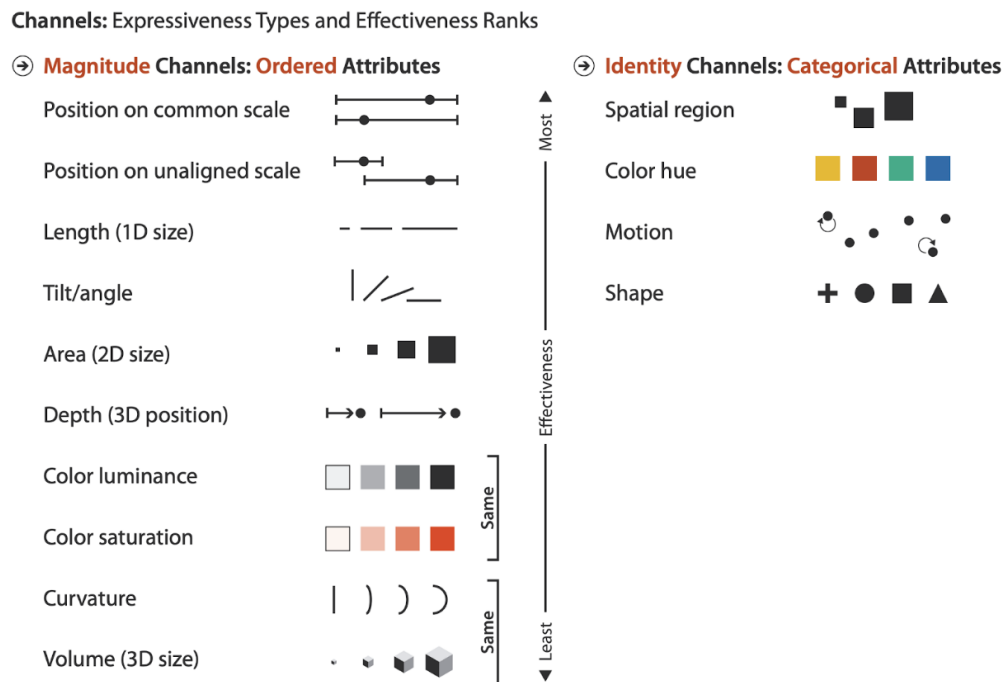


Figure 3.1: Summary diagram of the effectiveness of encoding channels by Munzner, 2014.

axes was that, to improve clarity in positioning, the visualisation requires small points to show the position, which limited the ability to convey information about other features, such as timbre, in that point.

Conversely, the addition of encoding the arousal-valence position in background colour was very distinctive. However, these combined colour scales mixed using gradients, while distinctive, were hard to interpret on the values they were representing once the positional element was removed.

To enhance the clarity of the arousal-valence scale, it would be more effective to plot the visualisations in a single 'flower field' that plots all the music to be shown on the same axes. Some experiments with showing information in other ways using flowers are shown in 3.3. However, the investment in programming that would be required to implement this flower field design in a working user-interface prototype is beyond the scope of this thesis and therefore left for future research.

Mackinlay's ranking of encodings seems more suitable for showing multiple data points (tracks) *in the same plot*, but less suitable when showing a *single visualisation* for each data point (track).

As both valence and arousal have a strong meaning separately, and not just combined, I chose to therefore separate the two in the visualisation, using colour and line width, respectively.

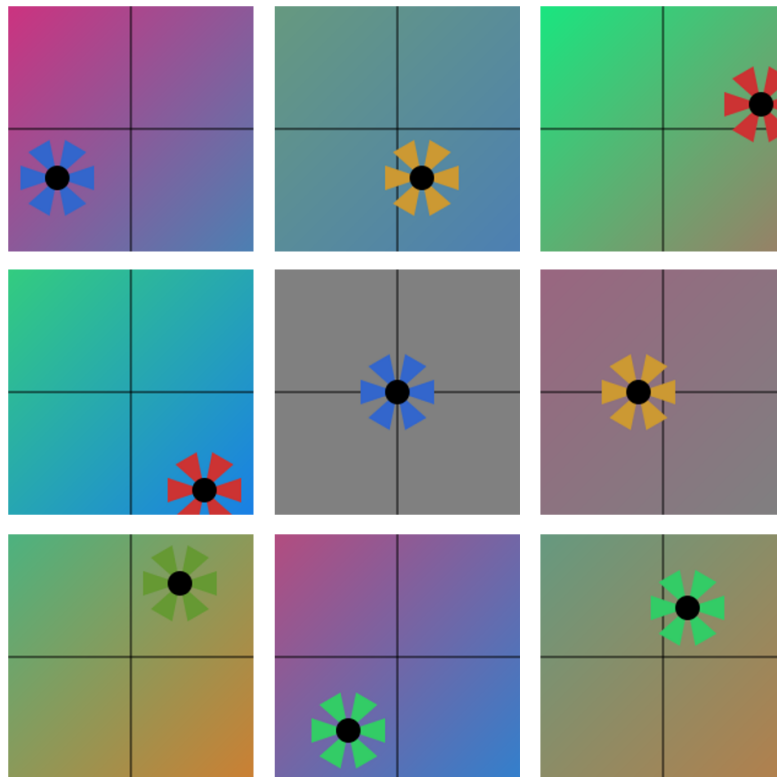


Figure 3.2: A design experiment, visualising music with flowers, using Python Cairo. The location of the flower in the picture shows the arousal and valence of the represented music. Arousal and valence are encoded in their position on the thumbnail, but also encoded in the background colour. This is called double-encoding (Zhang et al., 2023). The colour of the petals signifies the timbre of the music.

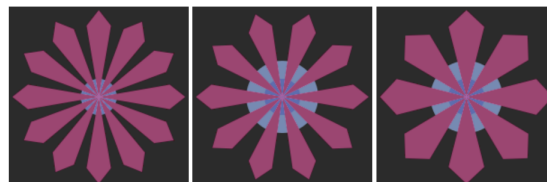


Figure 3.3: Experiments in Cairo drawing flowers with varying amounts of petals, petal widths and centre sizes.

3.3.2. Timbre Visualisation

As mentioned in section 3.2, timbre or tone quality is a concept that covers a wide range of differences in music, making it as elusive as the concept of musical similarity in general. Consequently, it is challenging to identify a singular axis or scale that can be employed to effectively show these differences. As an abstract visualisation of music already exists, my goal here was to create a human-understandable one. Consequently, I decided to utilise the more comprehensible attribute of identified musical instruments.

Nevertheless, the timbre of a song cannot be fully understood without considering other factors in addition to the instruments used. Timbre is influenced by various factors such as the harmonics, dynamic variations, attack and decay, vibrato, and so forth (Downie, 2003). Many of those musical attributes are also reflected in genre (J. J. Aucouturier & Pachet, 2003). While timbre is closely related to genre, it is not the same. Genre is a broader concept, as it encompasses an entire style's musical conventions, while timbre is solely about the sound quality. For the majority of people, genre is a more logical concept, as genres reflect a historical and cultural context. However, there is no single taxonomy of genres that experts, or the vast majority, agrees on (J. J. Aucouturier & Pachet, 2003; Srinivasa Murthy & Koolagudi, 2018).

To illustrate this, a single musical track may correctly be categorised as belonging to each of these genres: classical, romantic, and opera.

A song may thus be categorised in multiple genres, and a clear taxonomy is yet to be established. I have chosen to retain this ambiguity in the visualisation.

The findings of the visualisation experiments on the subject of mood have necessitated the setting aside of the ranking of encodings in figure 3.1. Instead, the focus changed to the Radar Chart as a means of information visualisation. Radar charts are especially useful for comparing multiple dimensions for data points (tracks), as they allow for a visual representation of how each variable performs across several dimensions.

Consequently, a suitable visualisation for timbre would be as follows: a radar chart or glyph to indicate the musical instruments, and a colour to show an abstract representation of genre.

3.4. Chosen Visualisation

The previous section left us with a chosen visual representation for each of our features. In this section, on each of these representations will be elaborated. Then, the representations of the audio features will be combined to present the chosen visualisation design: SoundShapes.

3.4.1. Mood: Valence as Background Colour

Valence is represented by colour. According to research by Moharreri et al., 2011 negative emotions are more associated with the colour red, while positive emotions are often associated with the colour blue. These colours were mapped to a scale, with a neutral grey colour in the middle, and used as the background colour for the visualisation, as shown in figure 3.4.

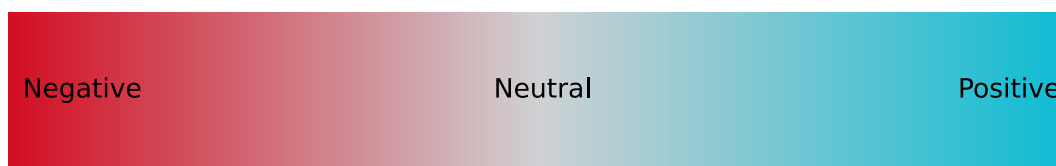


Figure 3.4: Valence encoding in background colour.

3.4.2. Mood: Arousal as Line Width of the Shape

Arousal or energy is depicted by the line width in the visualisation. A wider, or thicker line represents a more energetic track, while a thinner line is used for less energetic music, as illustrated in 3.5.

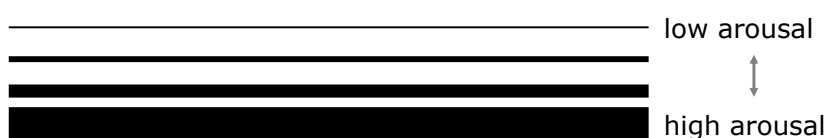


Figure 3.5: Arousal encoding in line width.

3.4.3. Timbre: Instruments as a Radar Chart Shape

Instruments are displayed with a radar chart. Radar charts are suitable for comparing different items on multiple axes with each other (Chambers et al., 1983).

In literature on Information Visualisation, some guidance can be found for effective use of radar charts. The following items should be considered:

- It is recommended that a radar chart should have between three and ten axes. Despite the absence of empirical evidence, to my knowledge, it is generally accepted that the use of more than ten axes in a single radar chart is uncommon (Fuchs, 2015). Accordingly, a limited number of instruments should be displayed.
- Radar charts are easier to interpret if the axes are shown. Thus a guideline is added from the centre, depicting each axis to provide some visual reference.
- A disadvantage of radar charts is that spurious associations may be assumed from adjacent axes. It is therefore advisable to group related axes next to each other (Fuchs, 2015). In this instance, this means that instruments that are similar to each other should be added next to each other.
- The lines that are exactly vertical or horizontal are perceived more strongly (Fuchs, 2015) and should therefore contain the most relevant instruments. As we are working with nine axes, which will be further explained in 4 Method, only one instrument is shown on an axis that is exactly vertical or horizontal. As the presence of vocals seem of particular importance for

our use scenario of background music, the ‘instrument’ *voice* has been selected for this axis.

The resulting axes are shown in 3.6.

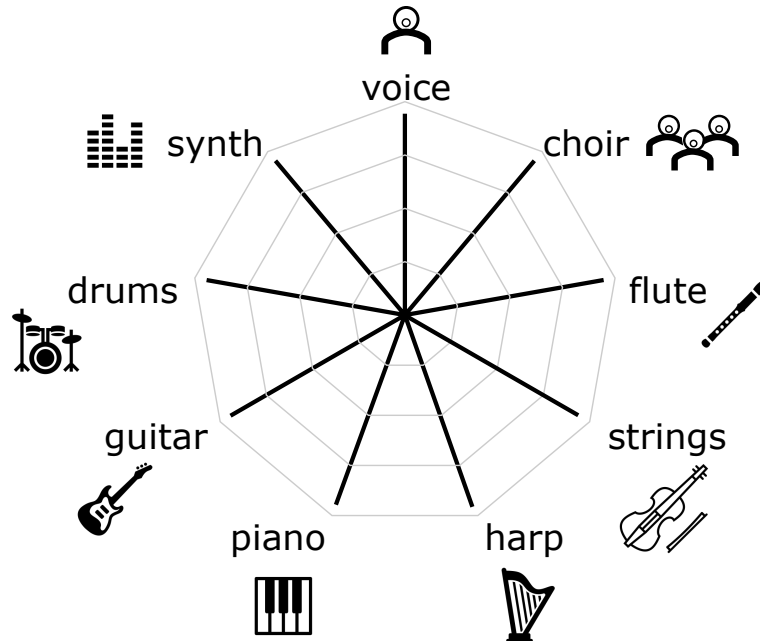


Figure 3.6: Instrument axes for the instrument radar chart.

3.4.4. Timbre: Genre as Colour of the Shape

As discussed in section 3.3.2 Timbre Visualisation, I have chosen to retain some of the ambiguity inherent in genre. Each genre is therefore not represented by a single colour, but instead genre probabilities are mapped onto a 2d colour space. This dimensionality reduction is explained in Chapter 4 Method. The two-dimensional colour space chosen is shown in figure 3.7. Due to the subjective nature of genres, no specific colours were chosen to represent particular genres.

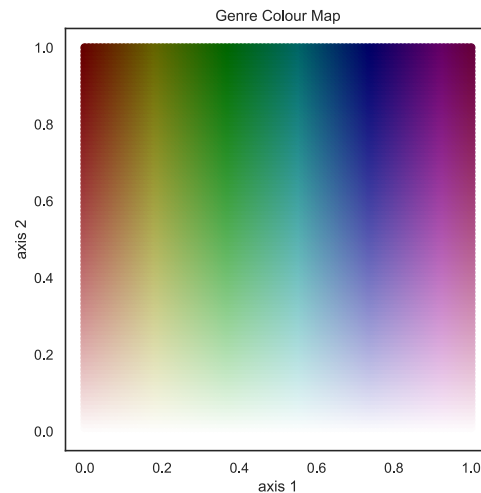


Figure 3.7: The two-dimensional colour map used to represent genre.

3.4.5. Introducing SoundShapes

The selected visual representations for instruments, genre, valence and arousal as described in the previous section, can now be combined into a SoundShape.

To begin with, I will consider the shape, which represents instruments. Figure 3.8 illustrates two examples.

A wider line width of the shape represents a song with higher arousal. This is illustrated in figure 3.9.

The colour of the shape depicts an abstract representation of genre, as illustrated in 3.10.

The background colour shows the valence, which is illustrated in 3.11.

Since the shapes are shown as thumbnails and are therefore quite small, the labels and grid lines are removed. In this way, I get to SoundShapes as illustrated in Fig.3.12.

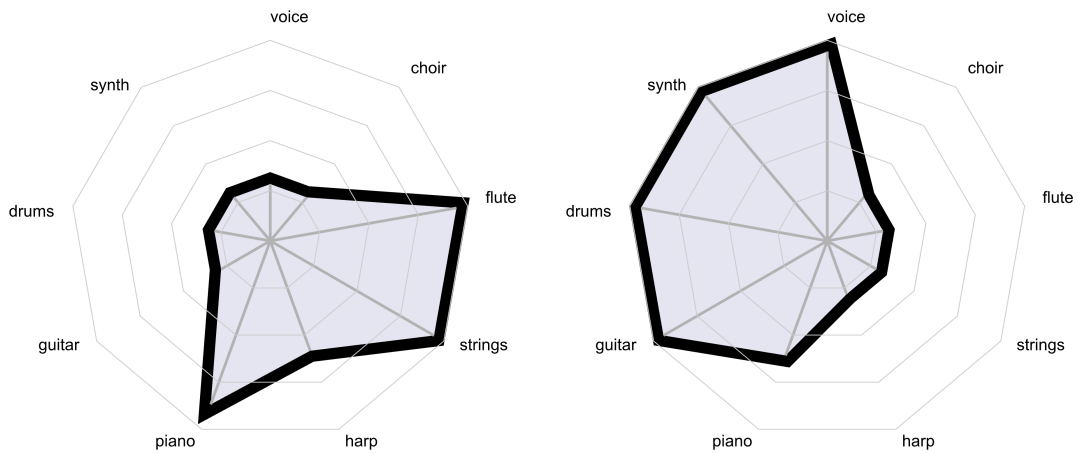


Figure 3.8: The shape on the left represents a song with classical instruments, while the shape on the right represents a sound that could be from a rock band.

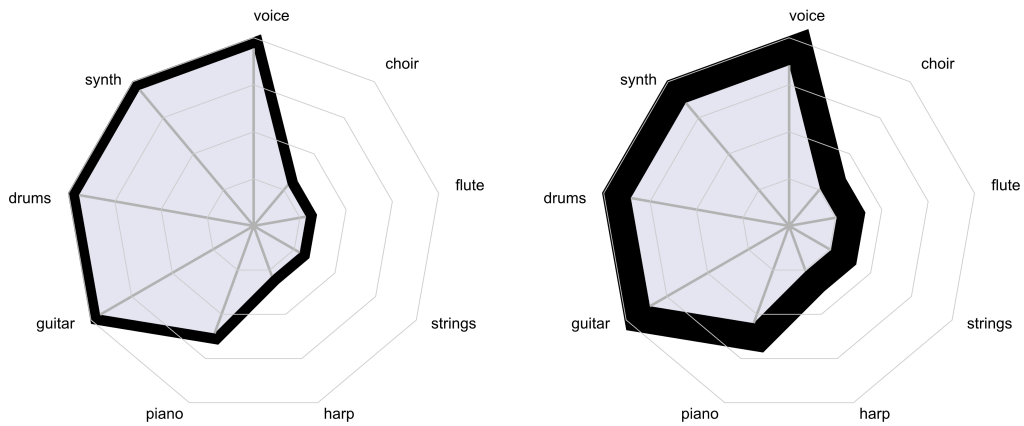


Figure 3.9: The song represented with the shape on the left is much lower in energy, i.e. has a lower arousal, than the song represented on the right.

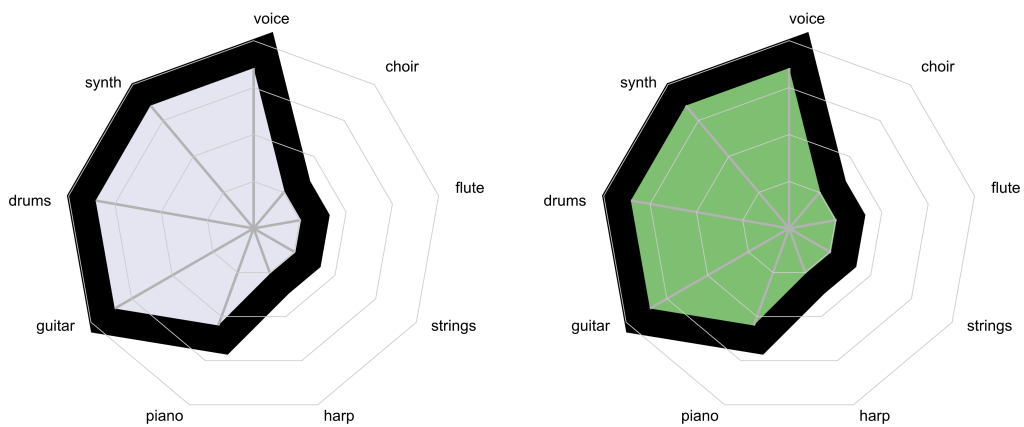


Figure 3.10: These shapes probably represent songs that sound like they are from different genres.

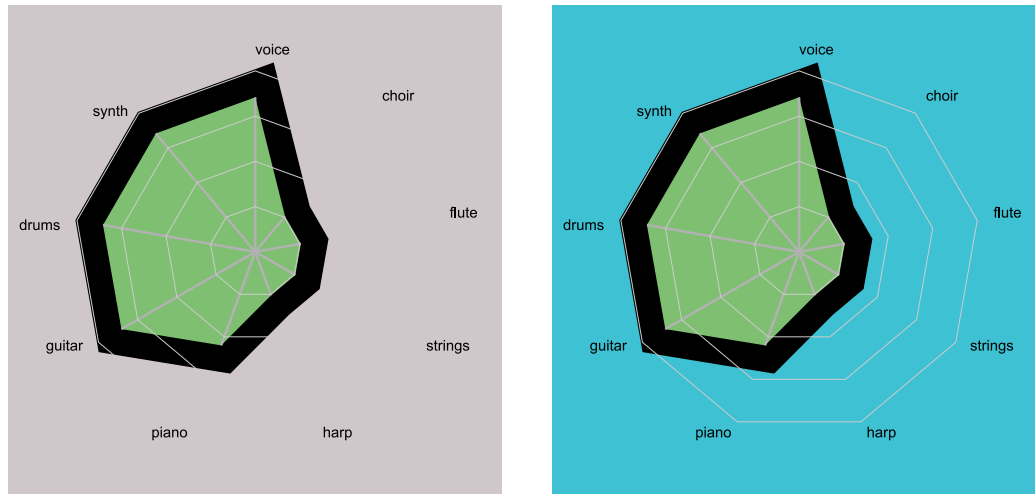


Figure 3.11: The song represented with the shape on the left is neutral with regards to valence, while the song represented on the right is more positive.

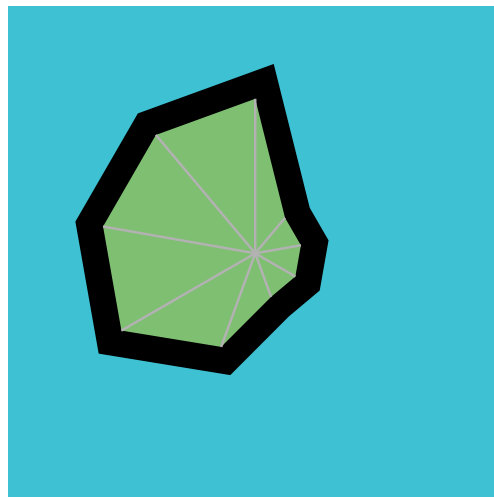


Figure 3.12: The instrument labels and grid lines are removed to improve clarity when scaled down. This is a SoundShape!

4

Method

Given the chosen design for the SoundShapes as described in chapter 3.4, Chosen Visualisation, this section explains the methodology used for data collection, feature extraction, visual mapping and user evaluation.

The workflow of this project is shown in figure 4.1. The steps to turn music into the SoundShapes visualisation are described in this chapter. A visual summary is provided in 4.2.

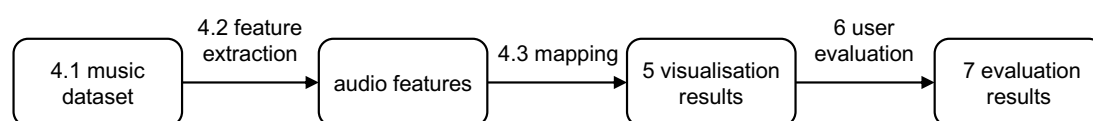


Figure 4.1: The workflow from mp3s to SoundShapes to evaluation results. The numbers refer to sections and chapters in this thesis.

4.1. Music Dataset

The final dataset consisted of 4698 30-second audio excerpts in mp3 format, accompanied by metadata such as artist name, track title, and the Spotify valence score. The dataset was obtained through a multi-step process.

Firstly, a dataset containing Spotify IDs and metadata for approximately 131 000 tracks was downloaded from Kaggle (Gelo, 2019). The dataset was compiled by querying the Spotify API for tracks released in 2018 (Gelo, 2018). The API primarily returned relatively popular tracks as it tends to prioritise such songs in response to queries. The resulting dataset covered a wide range of musical genres.

Subsequently, the Spotify IDs from the initial dataset were used to retrieve 30-second audio excerpts from randomly selected tracks via the Spotify Web API. This process yielded a dataset comprising 4698 30-second MP3 files, each accompanied by metadata such as the artist name, the track title, and the Spotify valence score.

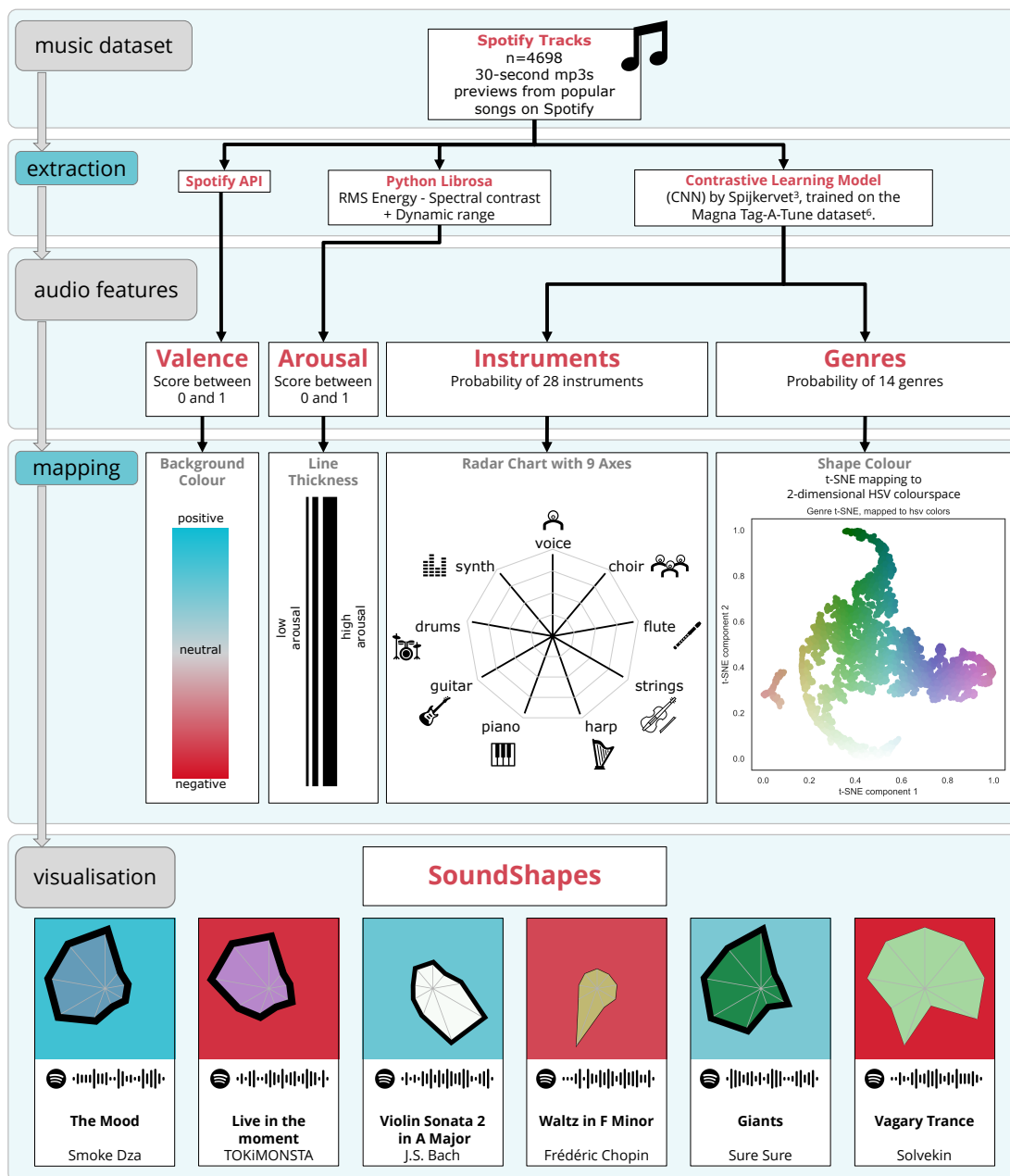


Figure 4.2: Visual summary of the pipeline that creates SoundShapes from mp3's.

4.1.1. Ethical and legal consideration of the data

This project adhered to Spotify’s developer terms of service, as it used the Spotify API to acquire the dataset. Obtaining individual permissions from all artists was unfeasible, and fair use provisions legally permit this usage without explicit consent. The samples used were of a duration of 30 seconds, rather than the full tracks, due to the copyright protections in place at the time, which prohibited the downloading of the full tracks.

Ethically, the project respects intellectual property rights by attributing the creators and, again, limiting the usage to a brief, non-commercial context.

However, copyright limitations pose challenges in accessing large, well-labelled music datasets for training. Notably, Spotify’s updated terms of service no longer permit the training of neural networks on their mp3 previews, a restriction that did not impact this project but highlights a significant limitation for future research in the field of music information retrieval.

4.2. Feature Extraction

The chosen features, as described in section 3.2, are mood, which is defined by valence and arousal utilising the Circumplex Model of Affect, and timbre, which is approximated with instruments and an abstraction of genre. For each of those, we now explain their method of extraction and resulting value range.

4.2.1. Feature Extraction of Mood: Arousal

It is possible to infer arousal directly from low-level audio features. The energy feature of Spotify (Spotify, 2024) was used as a reference point to assess the correlation between various audio feature implementations from the Python Librosa library (Librosa development team, 2023) and Spotify’s energy metric. The results are presented in table 4.1.

L_i	Librosa Audio Feature	c_i	Correlation with Spotify’s Energy
L_1	RMS Energy	c_1	0.78
L_2	Spectral Contrast	c_2	-0.78
L_3	Dynamic Range	c_3	0.60

Table 4.1: Correlation of Librosa Audio Features with Spotify’s Energy Feature

The set of outcomes for each of these Librosa audio features were scaled to values ranging from 0.0 to 1.0, or 1 to 0 for the inversely correlated one, and were then multiplied by their respective absolute correlation factor and summed together. The resulting value was then, again, scaled from 0.0 to 1.0. The process is formalised in equation 4.1.

$$A = \text{scale} \left(\sum_{i=1}^3 \text{scale}(l_i) \cdot |c_i| \right) \quad (4.1)$$

A denotes the arousal score.

$l_i \in L_i$ represents a Librosa audio feature applied to a track, with c_i the corresponding correlation factor of that Librosa audio feature, as depicted in 4.1.

The function $\text{scale}(y)$ scales $y \in Y$ linearly to the range $[0, 1]$, with the exception of the scaling of the Spectral Contrast scale (l_2), which is scaled inversely linearly to the range $[1, 0]$, as it is inversely correlated.

4.2.2. Feature Extraction of Mood: Valence

Valence is challenging to derive from low-level audio features. While Spotify has implemented such a model, they keep implementation details secret. The only clue comes from a [2013 blog post](#), now removed, where a representative mentioned “We have a music expert classify some sample songs by valence, then use machine-learning to extend those rules to all of the rest of the music in the world, fine tuning as we go.” I speculate that the lyrics of a song are also used to extract a sentiment to train the model.

There are some audio features that correspond with valence, such as mode (major/minor) and harmony (consonant/dissonant), and the loudness and pitch in chords. For example loud chords and high-pitched chords usually suggest more valence than soft chords and low-pitched chords, irrespective of mode. However, these qualifications seem to be largely cultural, and remain open for interpretation.

The development of a well-functioning valence scorer represents a significant undertaking that could constitute a full thesis project in and of itself. For those interested in pursuing such an endeavour, I would like to direct them to Ren et al., 2015 and Yang and Chen, 2012 as a starting point.

Consequently, the Spotify API was employed in order to retrieve a valence score for each track in the dataset. The Spotify Valence Score ranges from zero to one, and thus, no scaling was applied.

4.2.3. Feature Extraction of Timbre

Prior to choosing the final design as described in Chapter 3, Conceptualisation, was chosen, multiple attempts were made to approximate timbre with low-level audio features such as Mel Frequency Cepstral Coefficients (MFCCs). The outcomes proved insufficient to use for the purposes of this research. For further details, please refer to Appendix A.

In order to retrieve probability scores for both timbre features, the convolutional neural net presented and trained by Spijkervet and Burgoyne, 2021 was used. At the time of writing in May 2024, this model represents the state of the art in performance on music tagging, which is also demonstrated by its top ranking on the [Papers With Code](#) platform. Moreover, this model is lightweight and well-documented.

This model was employed to extract probability scores for 28 instruments and 14 genres, which are a subset of the 50 most-used labels in the Magna Tag-A-Tune dataset (Law et al., 2009), which the model was trained by (Spijkervet & Burgoyne, 2021).

Each track was divided into segments. Then, a probability score was assigned to each segment. The probability scores were then averaged over all segments to obtain the final probability score for each label for the entire track. The manner in which the probability scores for instruments and genres underwent reduction in dimensionality to form visualisable features is outlined in sections 4.3.3 and 4.3.4, respectively.

4.3. Feature Mapping to Visualisation

The visualisation was implemented utilising the Python Cairo library, also called Pycairo (Pycairo, 2024). Cairo enables the ‘drawing’ of visualisations in an SVG vector format.

4.3.1. From Valence Score to Background Colour

Valence was mapped to a Python Matplotlib LinearSegmentedColormap, as described in section 3.4, Chosen Visualisation, and in figure 3.4. The resulting colour was then used to specify the background colour of the SoundShapes.

4.3.2. From Arousal Score to Line Width

Arousal, again as described in section 3.4, Chosen Visualisation, was used to determine the line width of the SoundShape instrument outline. For this use, the score was linearly scaled to the minimum and the maximum width of the outline.

4.3.3. From Instruments to Radar Chart Shape

Feature extraction for timbre, described in section 4.2.3, resulted in probabilities of 26 instruments. As mentioned in 3.4, the maximum amount of axes shown in a radar plot is generally limited to 10. For aesthetic reasons, nine axes were chosen. Most instruments grouped together had a strong correlation with each other, except for harp and harpsichord. This makes sense due to the differing time periods the instruments have been popular in. Two instrument labels were removed: *sitar* and *no vocals*. The *no vocals* labels seemed potentially confusing for the purposes of this visualisation. Most fully instrumental songs scored low on these labels. The *no vocals* labels did seem to indicate a temporary instrumental section in music otherwise including vocals. Thus, the labels do not indicate the absence of vocals over the whole track, and may therefore be confusing, especially compared to the other, more straightforward labels. The *sitar* label was an uncommon instrument in our dataset, which resulted in a limited usage of it. Therefore, both the *no vocals* label and the ‘*sitar*’ label were not used for visualisation.

The chosen groupings are shown in table 4.2. This grouping was done by summing. While averaging might make more sense logically, these groupings made less sense upon auditory inspection. While most of the instruments grouped together have a high correlation, it seems that these values are still too low compared to non-averaged labels. This is illustrated by the fact that no new label, for any song in the dataset, exceeded a probability score of 1.

However, to mitigate some of the heightened sensitivity, especially for the new ‘voice’ label, the new ‘voice’ score was divided by 2.

Magna Tag-A-Tune label	New label
vocal - singing - vocals - male - male vocal - man - male voice - voice - female vocal - female voice - female - woman	voice
choir - choral	choir
flute	flute
strings - violin - cello	strings
harpsichord - harp	harp
piano	piano
guitar	guitar
drums	drums
synth	synth
sitar	<i>excluded</i>
no vocals - no vocal - no voice	<i>excluded</i>

Table 4.2: Original and new instrument labels.

Most instrument scores were very small, between 0.3 and very close to 0. By using a logistic scaling function for each track, differences between instrument scores were emphasised and thus made visible.

To ensure an aesthetic and distinctive visualisation, an instrument score of 0 corresponds to 30% of the maximum axis length (or radius of the shape). It was also ensured that at least one instrument per track reached 100% of the width by an additional, linear, scaling.

4.3.4. From Genres to Shape Colour

At the end of subsection 4.2.3, Feature Extraction of Timbre, probability scores were extracted for fourteen genres. These genres are listed in table 4.3. Two labels, ‘classic’, and ‘classical’ were aggregated by summing similar to the instrument merging, resulting in thirteen genres.

To retain some of the ambiguity inherent in genre, as discussed in section 3.3.2, genre is represented by the colour of the radar chart shape, using the two-dimensional colour mapping presented in subsection 3.4.4.

The thirteen genre scores, between 0 and 1, were each assigned to this two-dimensional colour space with a t-SNE algorithm (Hinton & Roweis, 2002) to a two-dimensional space. This algorithm gave, on auditory inspection, more intuitive results compared to UMAP (McInnes et al., 2018). The two t-SNE components were scaled to a range from 0 to 1, which were then used to extract the colour on the colourmap. The mapping of the genre t-SNE components to the colours is shown in figure 4.3. The colour mapping is defined in the equation below.

Magna Tag-A-Tune label	New label
classical - classic	classical
techno	techno
electronic	electronic
rock	rock
indian	indian
opera	opera
pop	pop
new age	new age
dance	dance
country	country
metal	metal
ambient	ambient
weird	weird

Table 4.3: Genre labels stayed mostly the same. Just the classical label was merged with the classic label.

Let h , s , and v be the hue, saturation, and value respectively.

$$\begin{aligned}
 h &= \text{TSNE_component_1} \times 0.9 \\
 s &= \text{TSNE_component_2} \\
 v &= 1 - (\text{TSNE_component_2} \times 0.6)
 \end{aligned}
 \tag{4.2}$$

By multiplying *hue* by 0.9, the last 10% of the hues are unused. This avoids confusion between genres with the lowest and highest values of the first t-SNE component, as they would otherwise appear similar due to the cyclic nature of hue. The *value* component does not utilise 40% of the lower brightness possibilities. This is for both aesthetic as well as for clarity reasons. By capping to a higher (lighter) brightness the SoundShapes look more cohesive, are better distinguishable on a light background, and maintain a stronger contrast with the black outline of the shape.

The resulting colour was then used as the colour of the inner shape formed by the instrument radar chart.

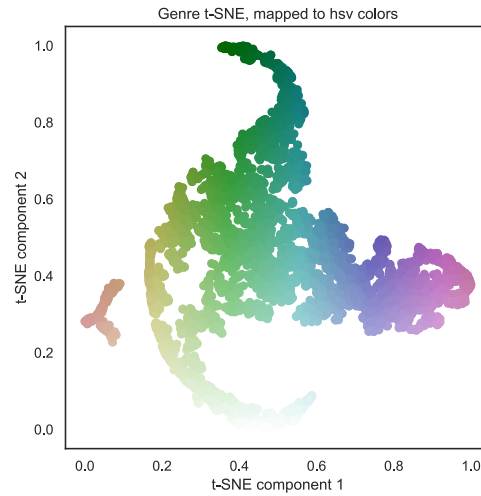


Figure 4.3: Track's genres are mapped with t-SNE to a colour.

4.4. Application Design

The prototype application was developed and deployed on a website using the Streamlit library in Python and the associated Streamlit deployment platform (Snowflake Inc, 2024). Streamlit was chosen for its rapid prototyping capabilities, for being a Python library so that the project remained in one programming language, and for its ease of deployment. The prototype application enables the browsing of music with the SoundShapes. To build this interface, the minimal functionality for browsing music was investigated, as illustrated with the sketch in figure 4.4. This design was then implemented.

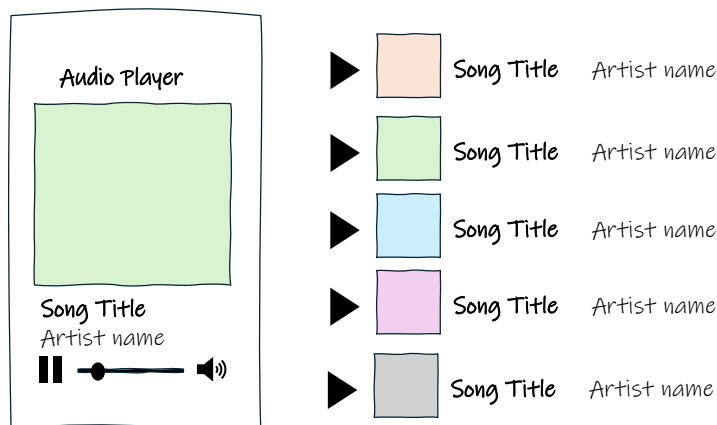


Figure 4.4: Sketch of an audio player application with minimal functionality.

5

Visualisation Results

A visual music representation was designed and implemented, resulting in SoundShapes for 4698 mp3s. In figure 5.1, a few examples of the SoundShapes are shown.

An application was built to explore the SoundShapes as well. To explore the SoundShapes through this application, visit SoundShapes.Streamlit.app.

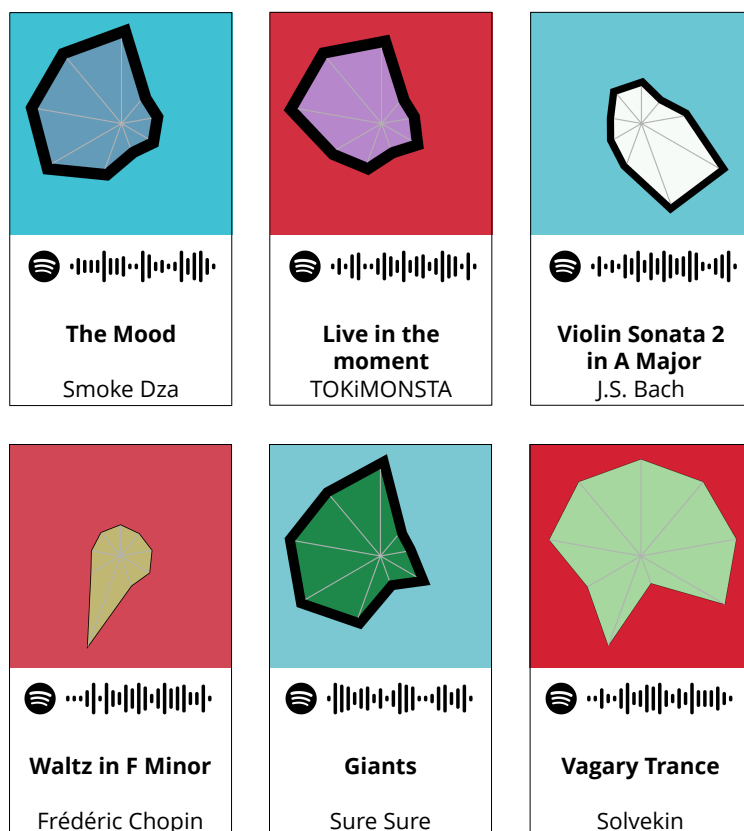


Figure 5.1: Some of the SoundShapes that were created. To listen to each track, scan the code with the Spotify app.

6

User Evaluation

In order to evaluate the ShoundShapes, the perception of the usability of the visualisations was explored.

6.1. Task to be Evaluated

The use scenario as described in 3.1, Use Scenario, is repeated below.

The thumbnails are used to support an exploratory search by active listeners, aimed at creating background music playlists from a seed song. These playlists are unordered and could be created for a variety of contexts.

The task, which is analogous to this use scenario, was presented to participants as follows. Participants were presented with a randomly selected song, designated as the seed song. They were then instructed to consider a hypothetical context in which they would listen to that song. Should they be unable to envisage such a context, participants were able to change to another randomly selected seed song to fulfil the task with.

Once a suitable seed song had been identified, participants were then instructed to describe the context in which they would listen to it. If they identified a context where they would not listen to the song in the background, but instead give it their full attention, they were asked to either propose a new context or request to receive a new, randomly selected, seed song.

Subsequently, participants were requested to identify another song that would be suitable for the listening context they had just described. Once they had identified such a song, the task was considered completed.

6.2. Participants

A total of nine participants were recruited from the author's social circle. The participants' ages ranged from 24 to 68, with five identifying as women and four as men. Approximately half of the participants report engaging in musical activities on a regular basis.

6.3. Procedure

Participants were asked to go to the website with the SoundShapes, and, if meeting online, to share their screen and sound. Participants were then asked

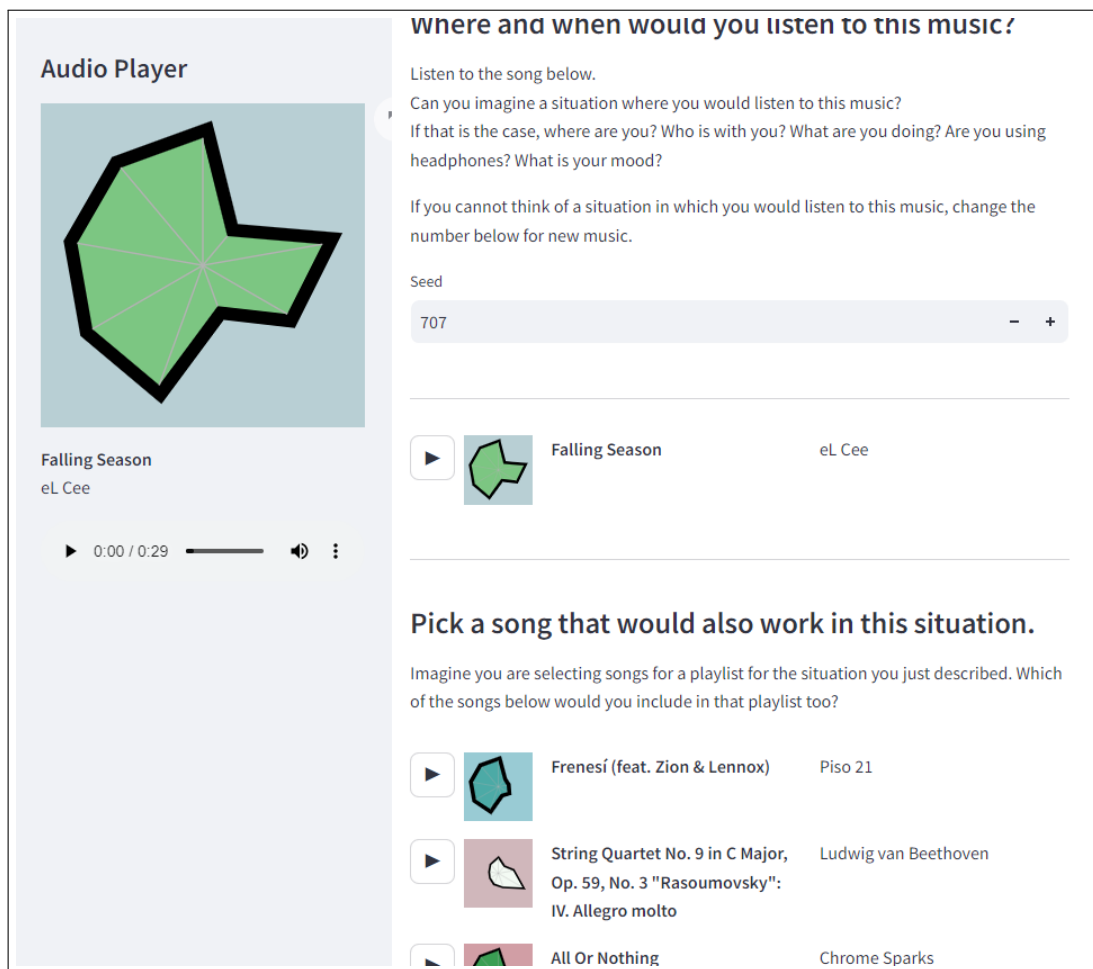


Figure 6.1: Screenshot of the application for the user evaluation with instructions for the participants.

to perform the task for evaluation using traditional album covers. This way, participants could get used to the interface. After completion, a short explanation about the SoundShapes and the meaning of their design was shown. Then, the participants performed the task again, now with the SoundShapes shown, instead of album covers. After this task was completed, participants were provided with the link to the survey, which they then responded to in private. A screenshot with the instructions is shown in 6.1. The full procedure description is included in Appendix B, User Evaluation Procedure.

6.4. Software

For the same reasons as provided in section 4.4, Application Design, Streamlit was used to prototype and deploy the user interface.

The survey was conducted with Qualtrics, as this software has been designed to address the most common concerns regarding data privacy and security in survey research (Utrecht University, 2024).

6.5. Application Design

The minimalist design from the previous chapters was built upon to enable user evaluation. Two identical pages were built, one with the original album covers and one with the SoundShapes. Both versions contain the same information about the task that the participant should perform. A new page was added containing some introductory information about the SoundShapes design. For user convenience, the following features were implemented: an input field allowing the seed song to be changed by modifying the random seed, and a button to increase the number of songs that are shown. A screenshot is provided in figure 6.1

The application, including the user study pages, can be found at [Sound Shapes.Streamlit.app](https://SoundShapes.Streamlit.app).

6.6. Survey

The sole form of data collection for this user evaluation was done with a survey. In the survey, participants were asked questions about their perception of the SoundShapes, using the System Usability Score (Brooke et al., 1996) to gauge usability, and four open-ended questions, presented as prompts, to elicit a more qualitative understanding of their perceptions. The System Usability Score was chosen as it is 'the most widely used standardised questionnaire for the assessment of perceived usability' (Lewis, 2018). The full survey can be found in Appendix B, User Evaluation Procedure.

7

Evaluation Results

The perception of the usability of SoundShapes was explored, as described in the previous chapter. This chapter presents the results of this evaluation in two parts: the responses to the System Usability Score (SUS) (Brooke et al., 1996), and the responses to the four open-ended questions. All responses can be found in Appendix C, User Evaluation Results.

7.1. System Usability Score

System Usability Scores are calculated such that agreement with a positive statement has the same effect on the score as disagreement with a negative statement to facilitate comparison. This (standard) version of scoring is used in all reporting below. For more information on how the score is calculated, please refer to Brooke et al., 1996.

All participants responded with System Usability Scores all equal to or exceeding 75/100. This indicates that all participants showed a willingness to use the SoundShapes. The average score was 84/100.

Compared to the questions about the other topics, question 5 and 6 had had relatively low scores. These questions are about the integration or consistency of the SoundShapes. These responses indicate that the SoundShapes were not always consistent with participants expectations from their previous experiences with the SoundShapes.

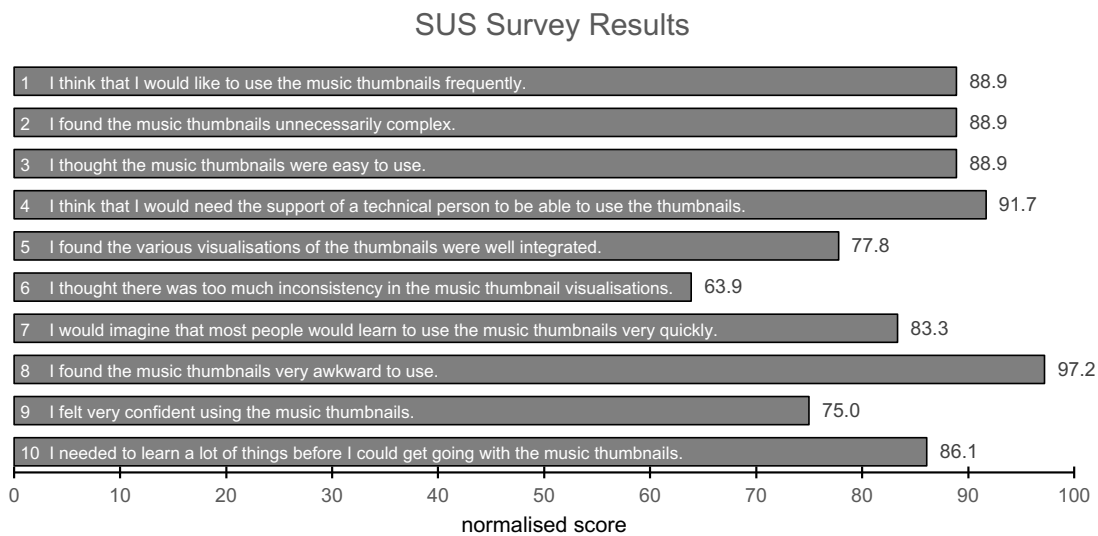


Figure 7.1: SUS scores per question.

7.2. Open-ended Questions

The evaluation reveals a strong positive reception towards the intuitive and playful nature of the SoundShapes. Users appreciate the ability to explore music based on visual patterns, though they find genre identification challenging, and find some inconsistencies in the SoundShapes.

Below, a few themes that were often mentioned are highlighted, with a selection of quotes from the participants.

Theme: intuitive, playful or fun

Four out of nine participants commented that they found the Soundshapes intuitive or playful in use.

- “Intuitive & playfull, bij elke plaatje krijg je meteen een gevoel en naarmate je het meer gebruikt wordt dit gevoel steeds meer bijgeschaaft en bevestigd.”
Translation: “Intuitive & playfull, you immediately get a feeling with every picture and the more you use it, the more this feeling is honed and confirmed.”
- “[What I liked most was] the game of whether I could quickly find similar sounding music just by looking at the thumbnail.”
- “origineel, grappig”
Translation: “original, funny”

Theme: shape colour as genre is difficult to interpret

Four out of nine participants commented that they found it difficult to interpret genre.

- “the current colour of the graph to genre mapping was the most difficult for me to connect with.”

- “De verschillende variabelen hebben verschillende ‘learning curves’, bijv. de lijndikte herken je snel in de intensiteit van de muziek, maar genre herkennen bij kleuren heeft bij mij meer tijd nodig.”
Translation: “The different variables have different ‘learning curves’, e.g. line thickness can be recognised quickly in music intensity, but genre recognition in colours takes more time with me. ”

Theme: a new way of exploring music

Three out of nine participants commented on how the SoundShapes how influenced their exploration of music.

- “I honestly love this for exploring new music, it stops traditional biases I may have about the album art/artist or genre when those are the features displayed.”
- “[What I liked most was] Exploring the different ways a song can be similar or dissimilar, finding new songs based on instruments, or on positivity, or on energy :).”
- “ Er komt onverwachte en onbekende muziek naar boven.”
Translation: “ Unexpected and unfamiliar music emerges. ”

Theme: inconsistencies with expectations

Three out of nine participants commented on how the SoundShapes did not always match their expectations.

- “[What did you like least about using the new thumbnails?] That it is not 100% correct yet.”
- “er komt ook wel eens iets naar boven wat binnen de verwachting van de thumbnail niet helemaal klopt.”
Translation: “something sometimes comes up that is not quite right within the expectation of the thumbnail.”

Theme: useful

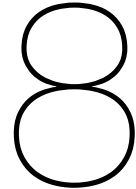
Three out of nine participants commented on how they found SoundShapes useful.

- “I found that the music thumbnails were quicker in filtering than reading the text”
- “ With these thumbnails I can quickly scan across hundreds of songs and identify which patterns would suit my taste or mood for a playlist.”

Theme: independence from a recommendation algorithm

Two out of nine participants commented that ShoundShapes helped them find music independently of a recommendation algorithm.

- “They music thumbnails really helps to find new, fitting music autonomously instead of relying on the algorithm.”
- “fijn om te weten dat je niet gestuurd wordt door een algoritme.”
Translation: “nice to know you are not being directed by an algorithm.”



Discussion & Conclusion

8.1. Discussion

A visual music representation was designed, implemented, and evaluated with the objective of investigating the perception of users on the effectiveness of visual thumbnail representations of audio-based features in aiding the discovery of similar music tracks.

The resulting SoundShapes visualise valence and arousal as a proxy for mood, and instruments and genre as a proxy for timbre.

SoundShapes show potential

The SoundShapes were subjected to evaluation by nine participants. All participants exhibited a willingness to accept the SoundShapes, as indicated by their SUS scores.

This acceptance was also visible from the responses given to the open-ended questions. Six out of nine of participants found the SoundShapes either intuitive, fun or useful. While the SoundShapes were not entirely consistent with expectations, these comments seem to indicate that they are consistent enough for user acceptance.

However, the small sample size may not accurately reflect the characteristics of the user population. Furthermore, the evaluation was conducted using a scenario provided by participants, which may not fully reflect real-world conditions. This makes it challenging to draw generalised conclusions. Nevertheless, the early-stage acceptance observed suggests a promising potential for music visualisations aiding in music discovery.

SoundShapes are an addition, not a replacement, to album art

Listeners feel strongly about album art, and album covers have a valuable historical context. Therefore, I advise using SoundShapes or other music audio features visualisations as an addition to facilitate in music discovery, and not as a replacement to album art.

Computationally lightweight

The visualisations can be generated with minimal computational resources, though I should note that the valence feature was not extracted from the mp3 by me. Nonetheless, the entire process, from mp3 to visualisation, for the dataset of 4698 30-second tracks, required approximately 15 minutes on a laptop without the use of a GPU. When my own implementation of the arousal feature was excluded and the Spotify energy feature was utilised instead, the same process required approximately one minute. This indicates that, at least in the case of Spotify, which already possesses arousal (energy) and valence scores for the majority of its catalogue, a visualisation such as SoundShapes could be implemented with a minimal computational investment.

Recommendations for future research

The performance of the ShoundShapes varies by genre and is now optimised for Western music. Further investigation into the use of different models, or re-training models with different labels, or personalisation may be beneficial.

The use of colour as genre proxy, at least in the way it was implemented now, was confusing to participants. The genre labels that were available from the Magna Tag-A-Tune dataset may not have been as relevant to the genres of music in our dataset. Colour mappings such as this one are not often used, but I think there is still a place for them, although they require a bit of a learning curve.

The SoundShapes are not robust against colour blindness. Adding some double encoding for the colours used, for example by adding curvatures, textures or depth, would help make the SoundShapes more widely accessible, as well as more distinctive to each other.

8.2. Conclusion

This study explores the potential of SoundShapes in enhancing music discovery through visual representations of audio-based features. The positive reception from participants, despite the limitations of a small sample size and controlled scenarios, suggests that such visualisations could be potentially useful.

However, further research is necessary to address the identified limitations, such as varying performance over musical genres and accessibility issues, to improve the robustness and general applicability of the SoundShapes.

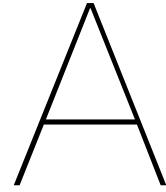
Future research should explore the use of different models, re-training with varied labels, and personalisation to enhance the effectiveness of SoundShapes. Additionally, accessibility for colour-blind users through double encoding methods should be ensured. While SoundShapes show promise as a supplementary tool for music discovery, they should be used in conjunction with, rather than as a replacement for, traditional album art.

References

- Afchar, D., Melchiorre, A. B., Schedl, M., Hennequin, R., Epure, E. V., & Mousallam, M. (2022). Explainability in music recommender systems. *AI Magazine*, 43(2), 190–208. <https://doi.org/10.1002/aaai.12056>
- Aucouturier, J. J., & Pachet, F. (2003). Representing musical genre: A state of the art. *International Journal of Phytoremediation*, 21(1), 83–93. <https://doi.org/10.1076/jnmr.32.1.83.16801>
- Aucouturier, J.-J., Pachet, F., & Sandler, M. (2005). "the way it sounds": Timbre models for analysis and retrieval of music signals. *IEEE Transactions on Multimedia*, 7(6), 1028–1035.
- Bonnin, G., & Jannach, D. (2014). Automated generation of music playlists: Survey and experiments. *ACM Computing Surveys (CSUR)*, 47(2), 1–35.
- Brooke, J., et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4–7.
- Chambers, J., Cleveland, W., Kleiner, B., & Tukey, P. (1983). *Graphical methods for data analysis*. Wadsworth.
- Chen, Y.-X., & Klüber, R. (2010). ThumbnailDJ: Visual Thumbnails of Music Content. <http://www.bordercommunity.com>
- Deldjoo, Y., Schedl, M., & Knees, P. (2024, February). Content-driven music recommendation: Evolution, state of the art, and challenges. <https://doi.org/10.1016/j.cosrev.2024.100618>
- Downie, J. S. (2003). Music information retrieval. *Annual review of information science and technology*, 37(1), 295–340.
- Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2), 303–319. <https://doi.org/10.1109/TMM.2010.2098858>
- Fuchs, J. (2015). Glyph design for temporal and multi-dimensional data: Design considerations and evaluation.
- Gelo, T. (2018). Collecting data from the spotify web api using spotipy. <https://nbviewer.org/github/tgel0/spotify-data/blob/master/notebooks/SpotifyDataRetrieval.ipynb>
- Gelo, T. (2019). Spotify audio features dataset. <https://www.kaggle.com/datasets/tomigelo/spotify-audio-features>
- Hinton, G. E., & Roweis, S. (2002). Stochastic neighbor embedding. *Advances in neural information processing systems*, 15.
- Hoveling, V. (2023). *Musicons* [Master's thesis, Delft University of Technology]. <http://musicons.io/>
- Katona, J. (2021). A review of human–computer interaction and virtual reality research fields in cognitive infocommunications. *Applied Sciences*, 11(6), 2646.

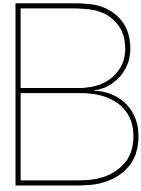
- Kim, H.-J., Yoo, M.-J., Kwon, J.-Y., & Lee, I.-K. (2009). Generating affective music icons in the emotion plane. *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, 3389–3394. <https://doi.org/10.1145/1520340.1520491>
- Knees, P., & Schedl, M. (2016). *Music Similarity and Retrieval*. <http://www.springer.com/series/6128>
- Kolhoff, P., Preuß, J., & Loviscach, J. (2008). Content-based icons for music files. *Computers and Graphics (Pergamon)*, 32(5), 550–560. <https://doi.org/10.1016/j.cag.2008.01.006>
- Law, E., West, K., Mandel, M. I., Bay, M., & Downie, J. S. (2009). Evaluation of algorithms using games: The case of music tagging. *ISMIR*, 387–392.
- Lewis, J. R. (2018). The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, 34(7), 577–590. <https://doi.org/10.1080/10447318.2018.1455307>
- Librosa development team. (2023). Librosa. <https://librosa.org/doc/>
- Lima, H. B., Santos, C. G., & Meiguins, B. S. (2022). A Survey of Music Visualization Techniques. *ACM Computing Surveys*, 54(7). <https://doi.org/10.1145/3461835>
- Liu, S., Cui, W., Wu, Y., & Liu, M. (2014). A survey on information visualization: Recent advances and challenges. *The Visual Computer*, 30, 1373–1393.
- Mackenzie, I. S. (2024). Human-computer interaction: An empirical research perspective.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*.
- Moharrerri, S., Dabanloo, N. J., Parvaneh, S., & Nasrabadi, A. M. (2011). The relation between colors, emotions and heart response using triangle phase space mapping (tpsm). *2011 Computing in Cardiology*, 573–576.
- Munzner, T. (2014). *Visualization analysis and design*. CRC press.
- Pycairo. (2024). Pycairo. <https://pycairo.readthedocs.io>
- Ren, J.-M., Wu, M.-J., & Jang, J.-S. R. (2015). Automatic music mood classification based on timbre and modulation features. *IEEE Transactions on Affective Computing*, 6(3), 236–246.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6), 1161.
- Snowflake Inc. (2024). Streamlit. <https://streamlit.io/>
- Spijkervet, J., & Burgoyne, J. A. (2021). Contrastive Learning of Musical Representations. <http://arxiv.org/abs/2103.09410>
- Spotify. (2024). Web api reference. <https://developer.spotify.com/documentation/web-api/reference/get-audio-features>
- Srinivasa Murthy, Y. V., & Koolagudi, S. G. (2018, April). Content-based music information retrieval (CB-MIR) and its applications toward the music industry: A review. <https://doi.org/10.1145/3177849>
- Utrecht University. (2024). Qualtrics. <https://students.uu.nl/en/node/6/qualtrics-a-survey-tool>

- Velankar, M., & Kulkarni, P. (2023). Music recommendation systems: Overview and challenges. In A. Biswas, E. Wennekes, A. Wieczorkowska, & R. H. Laskar (Eds.), *Advances in speech and music technology: Computational aspects and applications* (pp. 51–69). Springer International Publishing. https://doi.org/10.1007/978-3-031-18444-4_3
- Wiggins, G. A. (2009). Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music. *ISM 2009 - 11th IEEE International Symposium on Multimedia*, 477–482. <https://doi.org/10.1109/ISM.2009.36>
- Yang, Y. H., & Chen, H. H. (2012, May). Machine recognition of music emotion: A review. <https://doi.org/10.1145/2168752.2168754>
- Zhang, Y., Wang, Z., Zhang, J., Shan, G., & Tian, D. (2023). A survey of immersive visualization: Focus on perception and interaction. *Visual Informatics*, 7(4), 22–35.



Code

The code for implementing SoundShapes, the user evaluation application, and some experiments that were ultimately not incorporated into SoundShapes can be accessed at github.com/BerylVanGelderren/SoundShapes



User Evaluation Procedure

This section contains a description of the user evaluation protocol for the observer, screenshots of the application used for the evaluation, and a screenshot of the questionnaire.

B.1. User Evaluation Protocol for the Researcher

Preparation

1. Scheduling

- Schedule meeting, if online; discuss preferred platform (Zoom or Teams).
- Ask participants to be prepared with the following: a laptop with Chrome installed, their preferred videoconferencing tool installed (if meeting online), an internet connection, speakers/headphones, a maximum of about 30 minutes of their time.

2. Enable Screen Sharing

- Ensure screen sharing is enabled for all participants.

Participant Introduction

1. Explain the Study

- **Music Visualisation on Streaming Platforms:** Explain the concept and purpose.
- **Study Structure:** Mention that the study consists of a practical part and a questionnaire.
- **Practical Part:** Inform them that they will perform the same task twice – once with traditional album covers and once with the new music thumbnails.
- **Encourage Thinking Out Loud:** Ask the participant to verbalise their thoughts during the tasks.

- **No recording:** Tell participants that they are not being filmed for this test.
- **Voluntary Participation:** Assure them they can stop at any time without providing a reason. Explain how the data will be used and stored.

2. Open the Application

- Instruct the participant to open the application in Chrome: <https://musicthumbnails.streamlit.app>
- If online, assist participant in sharing their screen and sound.

Conducting the Evaluation

1. Practical Part

- Give short explanation about task, then let participants read and figure it out for themselves.
- Answer any questions participants might have, don't volunteer information and let participants explore on their own pace.
- If the context is not for a group, let participant pick another context/seed.
- If no context is found, let participant pick another seed.
- After participant has chosen a similar song, ask them to scroll down to click to continue.
- Give short explanation about SoundShapes with the explanation page, reassure participants that they do not need to remember all of it.
- Repeat task, same as before.

2. Questionnaire

- After completing the practical part, ask the participant to fill out the provided questionnaire (linked from the application) after I have left.

Conclusion

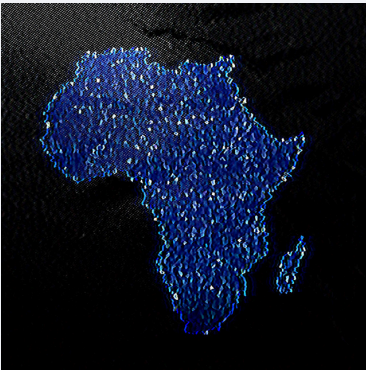
1. Thank the Participant

- Thank the participant for their time and feedback. Remind them not to forget to respond to the survey.
- Invite the participant to graduation talk.

B.2. User Evaluation Application Screenshots

Screenshots of the user application are on the next pages. The buttons to navigate through pages are at the bottom of these pages and thus not visible in the screenshots. To explore the application as if you were participating in the user study, please visit SoundShapes.Streamlit.app and click *Participate in the study*.

Audio Player



6mdm
Saint Mark

▶ 0:00 / 0:29

SoundShapes- test part 1

Where and when would you listen to this music?

Listen to the song below.

Can you imagine a situation where you would listen to this music?

If that is the case, where are you? Who is with you? What are you doing? Are you using headphones? What is your mood?

If you cannot think of a situation in which you would listen to this music, change the number below for new music.

Seed









706

- +

▶  6mdm Saint Mark

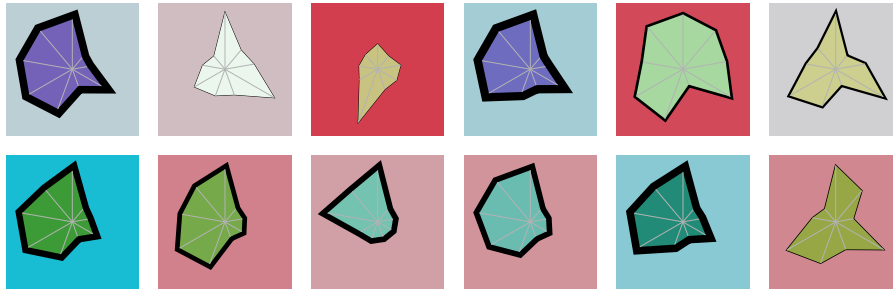
Pick a song that would also work in this situation.

Imagine you are selecting songs for a playlist for the situation you just described. Which of the songs below would you include in that playlist too?

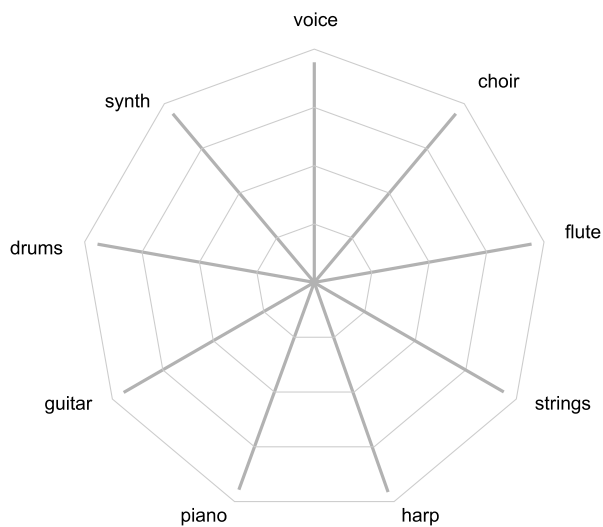
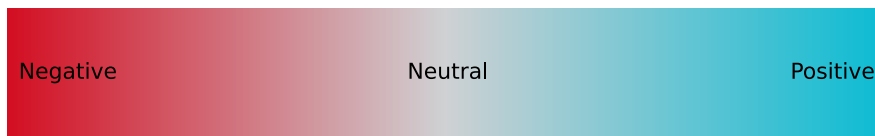
- ▶  Be Like That (Remix) Bobby Raps
- ▶  Valses oubliées, S. 215: No. 4, — Franz Liszt
- ▶  Die Kunst der Fuge, BWV 1080 (Arr. for Chamber Ensemble): Fugue No. 7 Johann Sebastian Bach
- ▶  Where You Are Tenille Townes
- ▶  Das Orgelbüchlein: No. 45, Alle Menschen müssen sterben, BWV 643 Johann Sebastian Bach
- ▶  Silent Treatment Eric Bellinger
- ▶  String Quartet No. 7 in F Major, Op. 59, No. 1 "Rasoumovsky": IV. Allegro - Thème russe Ludwig van Beethoven
- ▶  Grind Somburd

SoundShapes explained

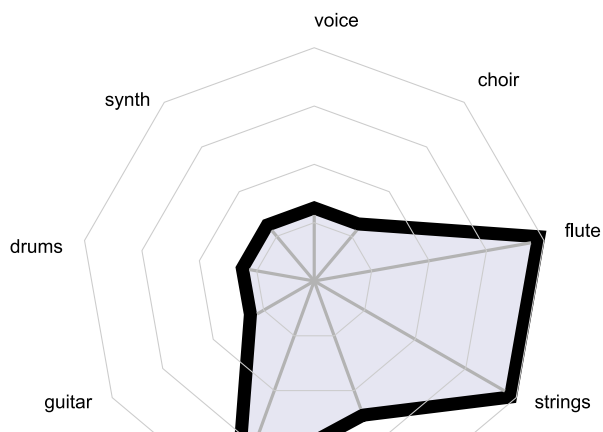
Here are some examples of SoundShapes, previously known as Music Thumbnails. Each SoundShape represents an audio file.



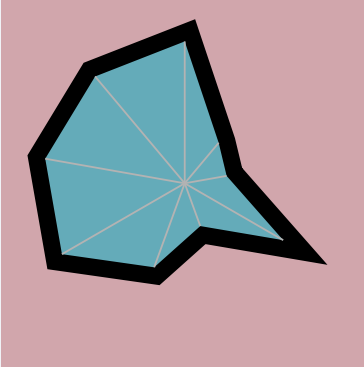
The background color of each thumbnail corresponds to the 'valence' or positivity of the music.



The shape of the thumbnail corresponds to the instruments that are detected.



Audio Player



With You
Mýa

▶ 0:00 / 0:29 — 🔊 ⋮

SoundShapes- study part 2

For part 2, we are going to do exactly the same thing as in part 1. Using the button below, you can go back to see the explanation again, but don't worry: sometimes it's best to just dive in!

[See explanation](#)

Where and when would you listen to this music?

Listen to the song below.

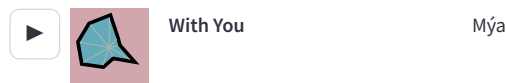
Can you imagine a situation where you would listen to this music?

If that is the case, where are you? Who is with you? What are you doing? Are you using headphones? What is your mood?

If you cannot think of a situation in which you would listen to this music, change the number below for new music.







Seed

513



Pick a song that would also work in this situation.

Imagine you are selecting songs for a playlist for the situation you just described. Which of the songs below would you include in that playlist too?

- ▶  N.Y. State of Mind feat. National Symphony Orchestra - (Live) Nas
- ▶  Homecoming Party Old Crow Medicine Show
- ▶  After All All Saints
- ▶  Concerto for Viola d'amore in D Minor, RV 395: II. Andante Antonio Vivaldi
- ▶  I Believe (Record Of The Week) [ABGT279] Ilan Bluestone
- ▶  Gibberish Neffex

B.3. User Evaluation Survey Screenshots

On the following pages, screenshots can be found covering the full survey as seen by participants, including the System Usability Score and the open-ended questions.

Thank you for helping me graduate!

This survey is meant to be done **after performing the user study tasks on musicthumbnails.streamlit.app**.

This survey consists of two parts.

- Part 1 will ask you to rate ten statements about the music thumbnails usability. Use the scale to indicate agreement with each statement.
- Part 2 contains five optional prompts. Feel free to answer any or all of them.

Your responses are anonymous and will only be used for research purposes. If you have any questions while completing the survey, please let me know.

Thank you! - Beryl

How would you rate the following statements?

	Strongly Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I think that I would like to use the music thumbnails frequently.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found the music thumbnails unnecessarily complex.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I thought the music thumbnails were easy to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I think that I would need the support of a technical person to be able to use the music thumbnails.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found the various visualisations of the thumbnails were well integrated.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I thought there was too much inconsistency in the music thumbnail visualisations.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I would imagine that most people would learn to use the music thumbnails very quickly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found the music thumbnails very awkward to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt very confident using the music thumbnails.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I needed to learn a lot of things before I could get going with the music thumbnails.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

The open questions below are all optional. Treat them as prompts, just to provide me with some deeper insights on what you think of the music thumbnails.

How do the new music thumbnails compare to using the traditional album art to find music for your playlist?

What did you like most about using the new thumbnails?

What did you like least about using the new thumbnails?

What suggestions do you have for improving the visualisations usability or functionality?

What suggestions do you have for improving the visualisations usability or functionality?

I have another comment:

I have an idea for a name for the music thumbnails:

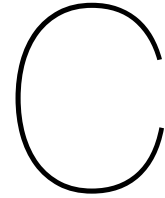
Anything else:



Your response has been recorded.

Thank you for participating, dank je wel!

Cheers, Beryl



User Evaluation Results

The responses to the survey from the user evaluation are provided in this appendix.

On the next page, the responses to the SUS survey are shown. The SUS scores have not been calculated here, just the raw responses are shown.

On the page that follows, all responses to open questions are shown, as well as the way they were counted for the reporting in Chapter 7, Evaluation Results.

Raw Responses to the System Usability Score Survey

0 = Strongly disagree, 1 = Somewhat disagree, 2 = Neither agree nor disagree, 3 = Somewhat agree, 4 = Strongly agree

	Participant number:								
	1	2	3	4	5	6	7	8	9
Q1	How would you rate the following statements? - I think that I would like to use the music thumbnails frequently.								
	4	4	4	4	4	3	3	3	3
Q2	How would you rate the following statements? - I found the music thumbnails unnecessarily complex.								
	1	0	0	1	0	1	0	1	0
Q3	How would you rate the following statements? - I thought the music thumbnails were easy to use.								
	3	3	4	3	4	3	4	4	4
Q4	How would you rate the following statements? - I think that I would need the support of a technical person to be able to use the music thumbnails.								
	0	2	0	1	0	0	0	0	0
Q5	How would you rate the following statements? - I found the various visualisations of the thumbnails were well integrated.								
	3	3	4	3	3	3	2	3	4
Q6	How would you rate the following statements? - I thought there was too much inconsistency in the music thumbnail visualisations.								
	2	1	0	1	2	1	3	2	1
Q7	How would you rate the following statements? - I would imagine that most people would learn to use the music thumbnails very quickly.								
	2	3	4	4	3	4	4	3	3
Q8	How would you rate the following statements? - I found the music thumbnails very awkward to use.								
	0	0	0	0	0	0	0	1	0
Q9	How would you rate the following statements? - I felt very confident using the music thumbnails.								
	4	2	4	3	2	2	3	4	3
Q10	How would you rate the following statements? - I needed to learn a lot of things before I could get going with the music thumbnails.								
	2	2	0	0	0	0	0	0	1

Participant number	Q1	Q2	Q3	Q4	Q5	Theme mentioned						
						Q1	Q2	Q3	Q4	Q5	Q6	
1	The music thumbnails say much about the contents of the music. I found I did not use album art at all when looking the first time, just looked at the song name/description and artists. When I know what I'm looking for (colours and shapes), I found that the music thumbnails helped me to skip over songs that could be potential candidates.	The game of whether I could quickly find similar sounding music just by looking at the thumbnail.	It takes some getting used to, especially the different axes of the radar plot. I wasn't expecting which axis is which.	What suggestions do you have for improving the visualisations usability or functionality?	I have another comment: I have an idea for a name for the music thumbnails: Anything else:	1						
2	the thumbnails could give me a better insight of how the music will sound, but a well designed album cover could do the same. my friends years very visual, that's why I like it. I think it could be a disadvantage with the thumbnails.	they give you an insight of what to expect before you play the song. And it is a easy to understand visual.	the background color didn't quite match my expectations of the song.	perhaps a sorting button where the songs are listed according to their similarity in shape and colours. Also, perhaps an explanation on the side with what the axes mean?	would be fun if you could search songs by creating thumbnails by drawing a shape and changing the colours. I like that idea. I can explain by creating the colour and then getting a list of recommended songs that are close.							
3	original, graphic, fun, copy, when able, niet gestuurd word, dans en muziek, van, over, moment, want houd ik eigenlijk van, op welk moment	likken of het beeld, het is ook wettelijk. Om het er ook een beetje in te krijgen, of er misschien wel langer mee bezig dan is bedoeling. M.a.w. nieuwsgierigheid wordt aangewakkerd.	Verassigendement is, het is een of komt ook wel eens voor, maar het is niet het thema van de thumbnail, het thema is de muziek.	Bepaalde muziek, bijvoorbeeld jazz of klassieke muziek, kan het heel goed gebruiken voor de muziekvisualisatie. Het is een beetje anders, maar het kan heel leuk zijn. Het is een beetje anders, maar het kan heel leuk zijn.	muziek, figuratief, muziek zien, ook jij, kan muziek (zemen)	1						
4	Album art is usually marketing-driven, in the best case aesthetically pleasing. Musical interpretation is not always enhanced by album art.	finding out whether or not the music was what I expected after trying to interpret the thumbnail	like the spiderweb graph, for different sort of instruments. Outline thickness for energy? is a bit more difficult to interpret, maybe that could be done with a colour? The score could be of the graph to get a better idea of the score.	making sorting options to select make quick selections	something like "music mood model"?							1
5	I really like the way of visualising songs, album art does not do it really for me, when you look at the cover of a CD, you see the album art, but when you look at the song, you see the features of the song. Putting these features of music into a graphical visualisation makes it much easier for my pattern recognition optimised monkey brain.	I honestly love this for exploring new music. It helps me find new music, but I don't know how to use it. I like the features of the song, but I don't know how to use it. I like the features of the song, but I don't know how to use it.	like the spiderweb graph, for different sort of instruments. Outline thickness for energy? is a bit more difficult to interpret, maybe that could be done with a colour? The score could be of the graph to get a better idea of the score.	I think I want to see some 'stronger' graphical features to represent 'genre' or 'mood', the hue of the graph was a bit difficult to interpret.	Songs or (identical) songs (based on icons)							
6	They enable people to select more what they want once they have figured out the system.	It is a nice tool also for the fun of it.	That it is not 100% correct yet.	More accurate	Shipping sound tracks	1						
7	You can choose music depending on your mood, or your activities at that moment.	Understanding what I like in terms of energy, aggressivity, positivity, etc.			My music							
8	They music thumbnails really helps to find new, fitting music autonomously instead of relying on the algorithm.	Exploring the different ways a song can be similar or dissimilar, finding new songs based on instruments, or on positivity, or on energy.	That the genre is still the heart of the thumbnail, and that that is the most difficult to learn or interpret.	More distinct genres, and perhaps an explanation of the colours to which genre they match? Instrument, energy, and positivity were all very easy to understand and work with.	Great work boy!! I think this could really be a separate product on its own. D.							1
9	De thumbnail communiceren een sfeer	instructie & playfull, bij elke plaatje krijg je meteen een gevoel en meteen je het meer gestructureerd wordt. Het is een beetje anders, maar het is een beetje anders. Het is een beetje anders, maar het is een beetje anders. Het is een beetje anders, maar het is een beetje anders.	Ik haalde nog niet meteen de muziek bij, zoals hip-hop of jazz, dus de muziek is soms een verrassing - wat ook juist leuk kan zijn, want zo ontdek je verschillende wijzen, anders niet had aangrijpelijk.	Awsome work!	1							
					Total number of respondents mentioning theme:	4	4	3	3	3	3	2