ARTICLE TEMPLATE

# How to sketch timbre: investigating sound-shape associations in free-form graphical representations of sound.

Sebastian Löbbers and György Fazekas

Centre for Digital Music, Queen Mary University of London, United Kingdom

**ARTICLE HISTORY**
Compiled November 30, 2023

**Wordcount:** 10596

**ABSTRACT**

This paper investigates the influence of cross-modal associations on visual representations of sound. Compared to established methods that ask study participants to match existing stimuli, this research explores how people represent sound through free-form graphical sketches when focusing on musical timbre. A total of 2320 sound-sketches were collected in two studies that included 28 and 88 participants. High-level sketch categories were established through qualitative analysis of participant interviews and a card-sorting exercise. Inter-participant agreement on representations and correlations between the auditory and visual domains was computed through statistical analyses of quantitative sound and sketch features. The results show that while sound-shape associations play a significant role in sketched representations, humans incorporate other visual aspects like structural complexity or texture or choose figurative representations of emotions or sound-producing objects. Some level of agreement on how to represent sounds could be found between participants, which appears strongest for sounds dominated by a single perceptual attribute. The analysis further suggests that sound-shape associations found in sound-sketches align with findings from perceptual matching tasks. This research is motivated by designing novel perceptually-informed mappings for digital music production and the results presented in this paper lay the groundwork for the development of a sketch-based sound synthesiser.

**KEYWORDS**

sound-shape associations, sound sketching, cross-modal mapping, timbre perception

## 1. Introduction

Humans make sense of sound in various ways often connecting different sensory domains. Cross-modal associations describe how a stimulus from one modality can induce a response in another modality. While often reported between sounds and colour, they can occur across various modalities, for example between colours and odours, sounds and tastes or sound and shapes (Spence, 2011). Cross-modal associations are sometimes wrongly referred to as synaesthesia. Synaesthesia is a rare condition with estimates of prevalence in the population ranging from 5% (Cuskley et al., 2019) to 0.5% Ramachandran and Hubbard (2001) and 0.05% Baron-Cohen et al. (1996).

CONTACT Sebastian Löbbers. Email: s.lobbers@qmul.ac.uk

Synaesthetes experience cross-modal connections involuntarily and consistently: the same stimulus always induces the same response. Cross-modal associations on the other hand are experienced in some form by most people, but connections tend to be far less consistent and might only occur situationally. Despite this difference, in a study investigating colour-to-sound mappings control participants and synaesthetes employed the same heuristics for linking auditory and visual domains, such as connecting pitch with lightness (Ward et al., 2006). The authors conclude that this type of synesthesia involves utilising mechanisms similar to those in typical cross-modal perception.

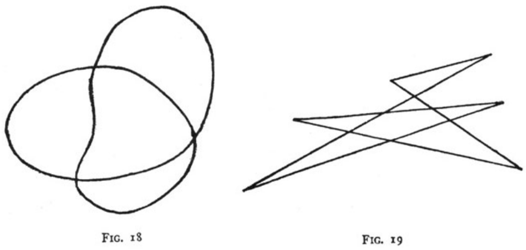



Figure 1.: Visual stimuli used in cross-modality experiments (Köhler, 1929). The left shape is overwhelmingly associated with the made-up word *maluma* or *boubou* and the right one with *takete* or *kiki*.

One of the earliest examples of cross-modal research is provided by Wolfgang Köhler, a member of the *Gestaltpsychology* movement in the 1920s, who found that people associate the made-up words *takete* or *kiki* with sharp, jagged shapes and *maluma* or *boubou* with soft, round shapes (Köhler, 1929). The effect was confirmed in multiple studies (Ramachandran & Hubbard, 2001) and generalised to all phonemes (Nielsen & Rendall, 2013). It was observed across cultures (Davis, 1961; Taylor & Taylor, 1962; Bremner et al., 2013), age groups including toddlers (Maurer et al., 2006), to some extent, with the visually impaired (Bottini et al., 2019) and between movement and phonemes (Shinohara et al., 2016). Similar associations were not only found for phonetic sounds but also for musical instruments (Adeli et al., 2014; Gurman et al., 2021) and abstract sonic textures (Grill & Flexer, 2012). While these studies explicitly ask participants to match stimuli from different modalities, a different approach measures how a stimulus in one domain can influence the perception of another stimulus in a different domain. A famous example is the McGurk effect (MacDonald & McGurk, 1978), a multisensory illusion that occurs when merging video and audio recordings of vocalising different consonants. Other examples include the impact of a mug's colour on the taste of coffee (Van Doorn et al., 2014) and the effect of piano music on sexual attractiveness (Marin et al., 2017).

The work presented in this paper focuses on sound-shape associations, a subset of cross-modality research, that describe how people connect timbre with geometrical shapes (Sidhu & Pexman, 2018). Cross-modal associations are typically investigated by asking participants to match visual and sound stimuli. However as pointed out by Soraghan et al. (2018), this approach can show which of the presented stimuli participants are most likely to match and does not account for whether a participant might favour a different representation altogether. This research asked participants in two studies to create monochromatic sketch representations of their associations with sound. The first study uses an exploratory design that imposes minimal restrictions to gain an understanding of the variety in which participants will represent different types of sound graphically. The second study uses a more controlled design that focuses on simple representations of sound produced by a frequency modulation (FM) synthesiser. Alongside sound-sketches, qualitative feedback was collected from participants through interviews and surveys to find out how the tasks were approached. The analysis categorises the various representational approaches and tests for correlations between sounds and shapes through statistical analyses of quantitative audio and visual features. The results lay the groundwork for the broader context of this research that aims to determine if a graphical sketch input can be used to help improve interaction with digital music production software, specifically for controlling digital synthesisers. The design process of the second study included the development of the digital sketching interface from a simple, generic sketchpad to a specialised sketching interface that could be used for controlling a sketch-based synthesiser.

This paper first introduces relevant research into visual representation of sound with a focus on sound-shape associations, musical timbre and sketch recognition in Section 2. Methods and material for both studies are described in Section 3. The results for both studies are presented in parallel in multiple sections: analysis of participant feedback in Section 4, sound-sketch categorisation in Section 5 and statistical feature analysis in Section 6. Discussion and conclusion can be found in Sections 7 and 8. Acknowledgement of the research funding body and a disclosure statement including reference to ethical approval for all studies are found in Sections 9 and 10. A detailed presentation of extracted audio and visual features is provided in Appendix A.

## 2. Background

This section first gives an overview of cross-modal representations of sound with a focus on sound-shape associations that build the basis for this research. It then continues to introduce research into timbre and sketch recognition that is relevant for the analysis of the user studies.

### *2.1. Cross-modality in visual representations of sound*

Cross-modality does not only play a role in human perception of the world but also in executing certain actions. Thoret et al. (2016) found that participants drew circles with a more elliptical skew when listening to sounds that evoked elliptical kinematics. Salgado-Montejo et al. (2016) showed the influence of pitch on the location of free hand movement. They also found movement to be more jagged for higher pitches and rounder for lower pitches. Both examples show that consistencies can be found between people not just in matching or rating tasks but also when given free agency over their response. However, participants' actions were not visualised posing the question of the influence of an action's graphical representation on the cross-modal response. In visual art, connections between the auditory and visual domains are a recurring theme. Notably, Russian artist Wassily Kandinsky developed multiple hypotheses on associations between shapes, colours and music leading to a number of cross-modal artworks inspired by pieces of composer Arnold Schönberg (Rucsanda et al., 2019). Consistent associations were found between complex stimuli of visual artworks and musical compositions (Albertazzi et al., 2015) and piano music excerpts and simple visual structures (Clemente et al., 2020). M. Küssner (2014) investigated how participants represent pure tones varied in pitch, loudness and tempo through drawing and produced similar findings about the connection of space and pitch as Salgado-Montejo et al. (2016). While timbre is acknowledged to play a role in cross-modal associations, the aforementioned research focuses on the musical context of the sound stimuli. Focusing on music timbre and visual texture, Giannakis and Smith (2000) and Giannakis (2006) provide an overview of audio-visual mappings for computer applications. They critically point out the difference between physical and perceptual representations of sound and arbitrary and sensory mappings. Grill and Flexer (2012) showed how sensory mappings, which are based on cross-modal associations, can help participants retrieve sound samples outside of a specific musical context. Knees and Andersen (2016) further developed this idea by proposing sound retrieval from a graphical sketch input and built a non-functioning prototype of such a system. Compared to Küssner's research, timbre plays a more dominant role here which participants tend to visualise through shapes, contours or textures - described by the researchers as *symbolic* representations. Recent research further investigated how participants represent different timbres in a more controlled way and showed how a functioning sketch-based sound retrieval pipeline could be implemented with the help of deep learning (Engeln & Groh, 2020; Engeln et al., 2021). The results showed that participants chose abstract, geometrical shapes as well as more complex images to represent sound. Work by the authors produced similar results (Löbbers et al., 2021) and further suggests that, to some level, free-form sketches consist of universally recognised patterns that allow participants to extract information about a sound's characteristic (Löbbers & Fazekas, 2022). The research presented in this paper takes a closer look at how to classify different representational approaches and collects a comprehensive sound-sketch dataset.

### *2.2. Description of musical timbre*

Musical timbre is generally described as the qualitative aspects of sound that cannot easily be quantified like loudness or pitch; however, it remains a concept that lacks a unified definition. An often-cited definition from the American National Standards Institute (ANSI) states that timbre is the auditory attribute that enables listeners to distinguish two sounds with the same loudness and pitch (ANSI, 1994). This is often criticised for only describing what timbre is not, but not defining what it actually is (Siedenburg & McAdams, 2017). While humans typically have an intuitive understanding of what constitutes timbre, the language they use to describe it varies individually (Saitis et al., 2020) often borrowing concepts of other sensory domains (Saitis & Weinzierl, 2019). Despite this lack of a common vocabulary, Wallmark and Kendall (2021) argue that some concepts like luminance (bright), texture (rough), and mass (heavy) widely appear across language groups. In the realm of contemporary music production, where digital technologies play a pivotal role, musical timbre occupies a prominent role for many modern music styles that distinguish themselves through their 'specific' sound rather than harmony, melody or musical structure (Provenzano, 2018; Blake, 2012). Finding or crafting a desired sound becomes an integral part of the production process which often involves searching large sample libraries or tweaking software parameters. These parameters often relate to the underlying digital signal processing (DSP) rather than perception of sound which can make it difficult to realise sound ideas or explore sonic spaces in an intuitive way (Seago, 2013). The perceptual attributes used to process a sound can be influenced by the source itself or individual factors of the listeners like preference, domain knowledge, cultural upbringing or listening context. This becomes apparent in the descriptors of sound libraries and synthesiser presets in digital audio production that can range from describing an acoustic instrument (keys, strings, drums), a compositional function (lead, pad, bass), a playing style (staccato, legato, arpeggio), references to genre (rock, hip-hop), hardware (808, clavinet), jargon (wobble bass, acid), mood (happy, sad) or describing a sound's timbre directly often with the help of cross-modal associations (bright, rough, shrill). Without tagging guidelines or a unified approach to describing sound, retrieving specific sounds from sample or preset libraries can be a cumbersome task that impairs the creative process. An increasingly popular approach for the organisation of these libraries is to display samples in a 2-dimensional space where similar samples are displayed close to each other (Fried et al., 2014; Bruford et al., 2019; Garber et al., 2021). The groundwork for this approach was laid by Grey (1977) and further popularised by Iverson and Krumhansl (1993) and McAdams et al. (1995), who obtained dissimilarity ratings through participant studies and using multi-dimensional scaling (MDS) to low-dimensional timbre space. For applications that are designed to work with a user's own sample library like the *Timbral Explorer*[1] it is not feasible to use samples annotated by humans. Instead, they use computationally extracted audio features. A popular and successful feature is the mel-frequency cepstral coefficient (MFCC) that can be used as input for deep-learning classification of diverse environmental sounds (Cramer et al., 2019). However, MFCCs are not very good at communicating sound characteristics to humans or helping explain the perception of sound (Siedenburg et al., 2016). Other popular computational features that, while still based in acoustic feature analysis, show better alignment with human perception are *spectral flatness* and *zero-crossing rate* that describe a sound's noisiness, *spectral centroid* that serves as a measure for brightness and the *root-mean-square* (RMS) describing a sound's loudness (Peeters, 2004; McFee et al., 2015). Pearce et al. (2019) aim to bridge the disconnect between computational features and features relevant to human perception by developing a model that, through a mixture of human-annotated sound samples and computed features, predicts a sound's *hardness, depth, brightness, roughness, warmth, sharpness, booming* and *reverberation.* Further research explores how, prompted by adjective, participants synthesise novel timbres through frequency modulation (FM) (Wallmark et al., 2019). Similarly, Hayes et al. (2022a) implemented a simple FM synthesiser to run in a web browser and collected a dataset of annotated synthesiser sounds through an online participant study. Participants were presented with a prompt to change a reference sound and then asked to annotate the result. The prompts were presented in the form *make the sound less/more: bright/rough/thick.* The attributes were derived from the luminance-texture-mass (LTM) model (Zacharakis & Pastiadis, 2015, 2016) that suggests that timbre descriptions can sufficiently be explained by these three dimensions for sounds with similar amplitude envelopes.

[1] `https://www.audiocommons.org/2019/01/30/timbral-explorer.html`

### 2.3. Sketch recognition

To statistically evaluate cross-modal associations between sound and shapes both domains have to be described as quantitative features. The extraction of information from hand-drawn sketches by a computer is called sketch recognition. Typical applications include recognition of handwriting or image retrieval from a sketch input. Sketches can be analysed as rasterised images using a computer vision approach. A benchmark task is the classification of handwritten digits with the MNIST dataset that is typically achieved through machine learning with convolutional neural networks (CNNs) producing the best results (LeCun, 1998; Deng, 2012). Representing sketches sequentially either through vectorisation or collecting sketches digitally gives the opportunity for a wider range of analyses. The most notable example is seminal work by Ha and Eck (2017) that introduced the *SketchRNN* architecture that uses a variational autoencoder (VAE) built on recurrent neural networks (RNNs). *SketchRNN* was trained on *Quick, Draw!*[2], a large-scale, open-source dataset with over 50 million sketches divided into categories ranging from simple geometric shapes to complex representations of animals and objects. *Quick, Draw!* inspired a large number of projects by enabling researchers to experiment with pre-train models and (re-)train them for specific tasks.

While *SketchRNN* produces impressive results for sketch classification and generation, algorithmic approaches might be more suitable for describing the shape of a sketch. The *ShortStraw* algorithm (Wolin et al., 2008) provides a simple, effective tool to extract corner points. Xiong and LaViola Jr (2009) extended the algorithm to also recognise curve points. Sezgin (2001) further show that information can not only be extracted from a sketch's shape but also from the sketching speed, for example, corner points can be estimated from a decrease in speed before each point. In the context of sound-shape associations, sketch or movement responses are often interpreted qualitatively by humans rather than computationally as it can be seen in the works by Salgado-Montejo et al. (2016) and Knees and Andersen (2016) discussed in Section 2.1. M. Küssner (2014) proposes a computational approach to extracting features from sound-sketches that, however, expect a representation inside a grid where the x-axis represents time. Engeln et al. (2021) harness advancements in deep learning for sketch recognition to implement a computational method for free-form sketch-based sound retrieval. However, an end-to-end retrieval approach might not reveal which aspects of a sketch are relevant to sound-shape associations.

## 3. Methods and Material

In order to investigate sound-shape associations in free-form sketch representation, two studies were conducted resulting in two sound-sketch datasets and corresponding qualitative and quantitative analyses. In both studies, participants were asked to sketch their associations with sound stimuli using a digital interface. For both studies, quantitative features were extracted from the sound stimuli and sound-sketches to investigate sound-shape connections through statistical correlations. Study 1 was accompanied by qualitative interviews to gain a deeper understanding of how participants chose their representations. This research further provides high-level categories for the broad classification of sound-sketches that were achieved through a card-sorting exercise in Study 1 and an automated, machine-learning approach in Study 2.

[2] `https://quickdraw.withgoogle.com/`

### *3.1. Design of perceptual studies*

As discussed in Section 2.1, relatively little research has been conducted on how humans represent sounds through graphical sketching. The two studies presented in this paper collect sound-sketches by combining existing research that asks participants to match auditory and visual stimuli or undertake specific hand movements while listening to sound with methods of digital sketch collection introduced in Section 2.3. In both studies, participants are presented with a number of different sound stimuli and are asked to sketch their personal association with them. The first study follows an exploratory design with minimal instructions for participants and different categories of sound from abstract synthesiser pads to acoustic instruments to obtain a wide overview of different representational approaches. Informed by the results of Study 1, the second study followed a more controlled design with a further developed interface and a narrower set of synthesised sound stimuli that encouraged simpler, more abstract sketch representations with a stronger link to sound-shape associations. Besides collecting sound-sketches, it is important for this research to understand the reasoning behind different approaches. Qualitative feedback can give a more detailed insight into the thought processes of a participant and might yield information that cannot be captured through quantitative data. For Study 1, qualitative data was collected through behavioural observation and a semi-structured interview that probed participants to reflect on their approaches and the interaction with the digital interface. Study 2 gave the option to provide brief written feedback and asked participants to answer a modified System Usability Scale (SUS) questionnaire (Lewis, 2018). All verbal feedback was analysed through thematic analysis (Braun & Clarke, 2006; Stuckey, 2015). General demographic data including age, gender, occupation and country of origin was collected through survey questions. In addition, the section of the Goldsmiths Music Sophistication Index (Gold MSI) (Müllensiefen et al., 2014) relating to musical training and engagement with music was used to categorise participants by music proficiency.

### *3.2. Material*

Both studies are set entirely in a digital environment and sound stimuli are selected with regard to timbral differences. Participants for Study 1 completed the study on the same laptop and, to keep similar conditions, only laptop or desktop devices were allowed for Study 2.

#### *3.2.1. Study 1*

The study was conducted in person using the touchpad on a 15” MacBook Pro and a pair of Beyerdynamic DT 770 headphones in calm, indoor locations. Ten stimuli with distinct timbral characteristics were crafted in reference to two cross-modal experiments presented in Section 1, which examine musical sounds (Adeli et al., 2014) and synthesized sounds and textures (Grill & Flexer, 2012). This approach aims to broaden the investigation of cross-modal representations beyond more narrowly defined categories to a wider range that could be encountered in a digital music creation environment. The stimuli can be divided into harmonic sounds that include musical instruments (*Piano*, *Strings*, *Electric Guitar*) and synthesised pads (*Telephonic*, *Sub-bass*) and inharmonic sounds that include environmental sounds (*Impact*) and abstract textures (*Noise*, *String Grains*, *Crackles*, *Processed Guitar*). *Piano* and *Strings* were created with virtual instruments in the Kontakt 5 plugin by Native Instruments, *Electric Guitar* was recorded with an audio interface via line-in. *Impact* was taken from the online audio repository freesound.org and *Telephonic*, *Subbass*, *Noise*, *String Grains*, *Crackles* and *Processed Guitar* were designed with built-in plugins in the digital audio workstation Ableton Live 10. Harmonic sounds were pitched to the MIDI note C3 which lies within the frequency range used in the reference experiments. All sound stimuli are monophonic and normalised for equal loudness. They last eight seconds with varying amplitude envelopes and include trailing silence to mark a clear endpoint during looped playback. The perceived base frequency may vary due to prominent harmonics.[3]

#### *3.2.2. Study 2*

The study was conducted entirely online. An initial audio check as well as a mid-study attention test ensured that participants listened to the audio playback either through headphones or speakers. Twenty different FM synthesiser sounds were selected from the dataset by Hayes and Saitis (2020) discussed in Section 2.2. All sounds have a short attack and high sustain to ensure that participants focus on timbre rather than amplitude envelope. The FM synthesiser was implemented with the AudioWorklets interface of the Web Audio API to synthesise sound directly in the browser rather than using audio samples which enabled continuous playback of a sound without looping. All sounds were pitched to the MIDI note A3 and loudness-normalised.[4]

### *3.3. Interfaces*

For both studies, digital sketching interfaces were used that can run in a web browser. This makes it possible to collect sketches directly as sequential data as presented in Section 3.7 and collect data online. As this research investigates sound-shape associations, the interfaces only allow monochromatic sketches with fixed stroke widths. Sketch analysis of Study 1, described in the following Sections, suggests that sound-shape associations more strongly inform simple sketches. To investigate that connection in greater detail, the interface was developed from a generic to a specialised design through an iterative design process that sought to encourage simple sketches between participants without majorly impacting their perceived sense of expressiveness and autonomy.

#### *3.3.1. Study 1*

The requirements for this interface, that is illustrated in Figure 2, were that participants can only express their ideas through monochromatic strokes, but are not otherwise limited or supported. The black canvas was chosen to clearly distinguish the sketching area from the rest of the website. No option to partially or fully erase a sketch was provided so that participants would not revise their original idea. While the goal of Study 1 was to gain insights into the different sketched representations of sound, it also provided feedback on how to develop the interface for sound-sketching tasks. Feedback evaluation showed that, overall, participants were unsatisfied with the

[3] All sounds for Study 1 can be accessed online at `https://bit.ly/3ta6crU` together with the sketches produced by participants.

[4] All sounds for Study 2 can be accessed online at `https://sfrl.github.io/study2_gallery/` together with the sketches produced by participants.


Figure 2.: Interface for Study 1: generic design that allows for fixed-width, white strokes on a black canvas with a fixed size of 750x750 pixels. There are no sketch length limits, stroke simplification and undo or reset options. Sketch points are connected by straight lines.

interface, especially in regards to sketching straight lines or smooth transitions. In addition, the canvas proved to be too small with multiple participants sketching over the edges. The black background received positive comments, but for some participants, it seemed to have artistic rather than purely functional meaning. While the design was successful in forcing participants to stay with their original idea, it did not take into consideration that participants might want to correct technical mistakes rather than change their overall approach.

#### *3.3.2. Study 2*

The feedback from Study 1 was addressed in the further interface design. Stroke simplification and spline interpolation were implemented to make it easier to sketch straight lines and smooth transitions. The canvas automatically stretches to the size of the browser window to prevent sketches from extending over the edges. A reset button ensures that participants can start over, but an erase or undo function was withheld to prevent participants from focusing on retouching their representations. These design choices are similar to the *Quick, Draw!* interface introduced in Section 2.3 which makes collected sketches compatible with their dataset and deep-learning architecture. A major difference between the two interfaces is the length limit introduced in Study 2. The goal of this research is to collect sketches that encode sound characteristics through shape which typically results in simple, abstract representations. However, the categorisation of Study 1 sketches, as described in detail in Section 5, show a significant number of figurative representations like scenes or objects. As these sketches are on average longer, limiting the length permitted by the interface was expected to reduce those types of representations. This was tested in three small-scale design studies with 10-15 participants each. Different designs were evaluated on how well they guided participants towards simple, abstract representations while maintaining the feeling that the interface allows them to be expressive. This concluded in the setup

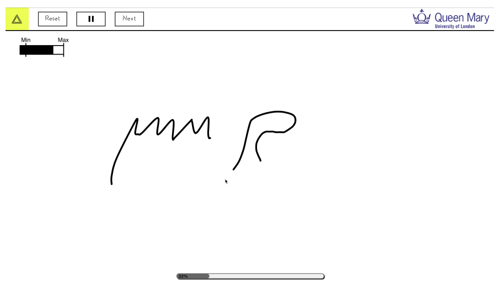



Figure 3.: Interface for Study 1: Specialised design that allows for fixed-width, black strokes on a white canvas that scales to a participant's browser window (a minimal size of 800x600 pixels was enforced). The sketch length is limited to the range of 30 to 150 points. A sketch can only be submitted if it exceeds the lower length limit. If a participant continues sketching after reaching the upper limit, sketch points are erased from the start. A small meter in the top left corner of the canvas indicates the current stroke length. Participants can reset the canvas to start over. The Ramer-Douglas-Peucker algorithm (Douglas & Peucker, 1973) was used for stroke simplification and stroke points were connected with Catmul-Rom splines (Catmull & Rom, 1974).

described in Figure 3.

### *3.4. Participants*

Recruitment was open to all adults over the age of 18. The aim of these studies was to find out about sound-shape associations by sampling from the general population. As engagement with music was expected to have an influence on representations, for Study 1 musicians and non-musicians were recruited in equal parts.

#### *3.4.1. Study 1*

Twenty-eight participants were recruited through mailing lists and in person at the School of Electronic Engineering and Computer Science at Queen Mary University of London. This group was divided equally by gender (14 female, 14 male), 25 were adults below the age of 34 (three between 34 and 49), 22 had a Western background (16 from Europe, 4 from North America, 2 from South America) and 5 an Eastern background (4 from China, 1 from India) with one participant preferring not to disclose this information. A participant was defined as a musician if they responded to having at least one year of formal music education and engaged in musical activity (playing an instrument, producing or composing music) at least once a month. This resulted in 14 musicians and 14 non-musicians.

#### *3.4.2. Study 2*

Eighty-eight participants were recruited through the online platform Prolific.[5] The majority of participants were female (48 female, 38 male, 2 other) and 84 were aged between 18 and 33 (M = 22.5, SD = 4.6). All participants had a Western background (56 from North America, 16 from Europe, 13 from South Africa and 3 from South America) with the majority coming from Mexico (54 participants). Sixty-six were

[5] `https://prolific.co/`

students and 28 stated to have had at least one year of formal music education and actively engage in musical activity at least once a month.

### *3.5. Procedure*

A similar procedure was used for both studies. As Study 2 was conducted online a more robust digital environment was deployed to ensure that participants completed the study successfully.

#### *3.5.1. Study 1*

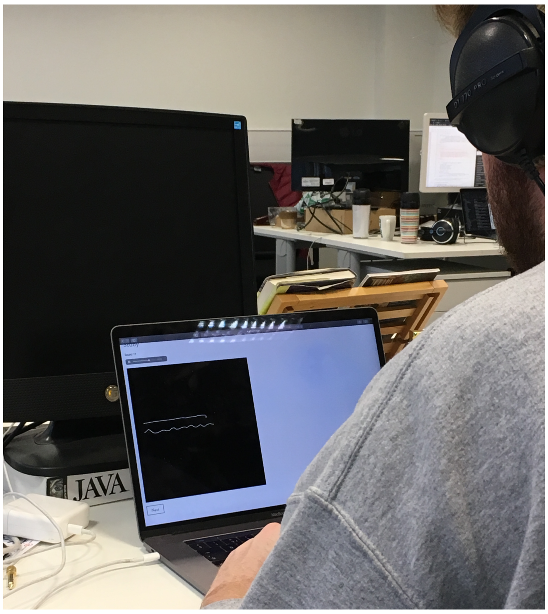



Figure 4.: Participant sketching a sound in Study 1

Participants were first asked to adjust playback audio to a comfortable measure using white noise as a reference. They then completed a questionnaire collecting demographic data and information about their experience with music. Participants were then asked to familiarise themselves with the sketching interface without audio before they were presented with the sound stimuli. The study intended to encourage a spontaneous response, therefore no information about the range of sounds was provided and participants were instructed to sketch what they believed to best represent each sound stimulus. Looped playback started automatically with the option to pause and resume. Each sound was played twice in a randomised order resulting in a total of twenty sketches per participant. After completion, a semi-structured interview was conducted asking participants how they approached the task and whether they found it difficult. No time limit was given and the study typically took twenty to thirty minutes to complete.[6]

[6]The setup for Study 1 can be accessed online at `https://bit.ly/3j3FkVO`.

#### 3.5.2. Study 2

Participants were first presented with a set of information ensuring that they use a laptop or desktop device and are able to listen to sound either through headphones or loudspeakers. This was followed by a short introduction of the study that included guidelines on representing sounds in an abstract rather than figurative way, a short explanation of the sketching interface, a guide to adjust playback volume to a comfortable level and a check that the browser window size was at least 800×600 pixels. Before starting the main task, participants were given the opportunity to familiarise themselves with the interface in two test rounds in which they were asked to sketch their associations with an imagined calm and noisy sound. For the main task, sound stimuli were played back automatically in a randomised order with the instruction *Listen to the sound and draw your association.* An attention test was played back after completing half of the task asking participants to sketch the number four instead of a sound. The study concluded with a survey asking participants about their experience with the study and the interface and collecting demographic data, experience with music and the hardware they used to complete the study.[7]

### 3.6. Sketch categorisation

The first step of the analysis categorises different representational approaches that participants deploy and investigates to what extent sound-shape associations inform them. The sketches produced in Study 1 were categorised in an open card-sorting study by six participants (4 female, 3 musicians) who did not take part in the main study. They were asked to sort the collected sketches into three to ten categories including short written descriptions. The study was completed remotely within three hours on participants' devices following detailed instructions that can be accessed online.[8] Following the methodology of Paea and Baird (2018), the results were encoded and reduced in dimensionality using principal component analysis (PCA). K-means clustering was used together with the silhouette coefficient (Rousseeuw, 1987), a measure of cluster goodness, to find the most suitable number of clusters between three and ten. Clusters were named and described qualitatively based on keywords that participants used in their category descriptions.

Due to the higher participant number, Study 2 produced a considerably larger number of sound-sketches and the categorisation process was automated with the help of machine learning. A variational autoencoder (VAE) using the *SketchRNN* architecture was pre-trained on the *Quick, Draw!* categories *Triangle*, *Square*, *Circle*, *Line*, *Squiggle* and *Zigzag* before feeding in sound-sketches from Study 2. The resulting 128-dimensional latent representation of the dataset was reduced with PCA and the best number of clusters was determined with K-means and the silhouette coefficient as described above. Clusters were named following the findings from Study 1.

Pearson's Chi-squared test was used to test whether sounds are represented equally across sketch categories or if certain types of sound can be connected to a specific category. Similarly, Cochran's Q test was used to assess whether a participant's musical proficiency has an influence on how their sketches are categorised.

[7] The setup for Study 2 can be accessed online at `https://sketching-sounds.web.app/`.

[8] `https://youtu.be/LXTlnaAciWw`

### 3.7. Quantitative analysis of sketch and sound features

All sketches are saved digitally as sequential data in nested arrays. Each stroke is described in a separate array that consists of sketch points described by their x and y positions and timestamps. A sketch is rendered into an image by connecting all points within each stroke array. To describe sketches quantitatively, a number of features can be calculated directly from the data structure and through simple arithmetic operations as demonstrated in Equations 1, 2 and 3, where $N$ is the number of strokes in a sketch and $\overline{L}$, $\overline{T}$ and $\overline{S}$ are their average length, completion time and sketching speed. The number of points in the $k^{th}$ stroke is described by $n_k$. Each point has a position $x_{k_i}$ and timestamp $t_{k_i}$. The Euclidean distance between two points is described by $d(p, q)$.

$$\overline{L} = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=2}^{n_k} d\left(x_{k_i}, x_{k_{i-1}}\right) \tag{1}$$

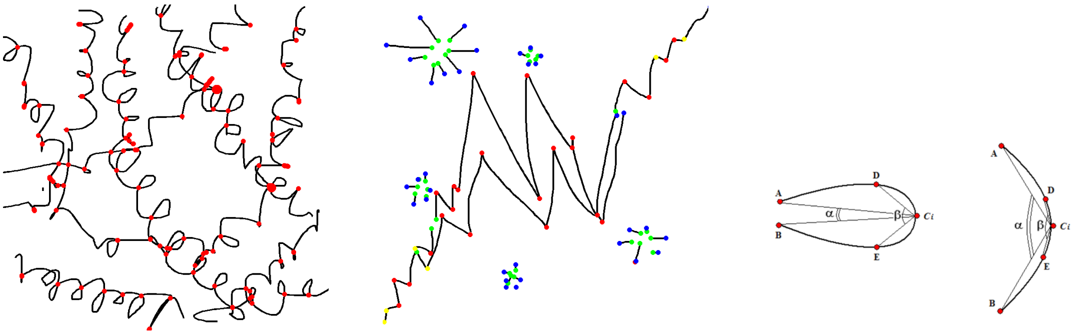
$$\overline{T} = \frac{1}{N} \sum_{k=1}^{N} t_{k_{n_k}} - t_{k_1} \tag{2}$$

$$\overline{S} = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=2}^{n_k} d\left(x_{k_i}, x_{k_{i-1}}\right) \frac{1}{t_{k_{n_k}} - t_{k_1}} \tag{3}$$

Sound-shape associations are usually reported with respect to a shape's contour focusing on their 'jaggedness' or 'roundness' (Adeli et al., 2014; Grill & Flexer, 2012). As illustrated in Figure 5, these attributes were quantified by extracting corner points divided into obtuse, right and acute angles and curve points divided into wide and narrow shape algorithm (Wolin et al., 2008; Xiong & LaViola Jr, 2009). A qualitative review suggested that sketches differ by the number of stroke intersections that can be interpreted as the 'noisiness' of a sketch. The number of intersections was determined using an adaptation of Bresenham's rasterisation algorithm (Bresenham, 1965). Prior to extracting features, the sketch data was cleaned by removing consecutive points with the same position and merging two strokes if a starting point was within a five-pixel distance to an endpoint. The number of intersections, corner and curve points is reported relative to the total stroke length of a sketch.

In order to investigate sound-shape associations through statistical analysis, the sound stimuli also have to be described using quantitative features. This was accomplished by computing the mean values of *Centroid Frequency*, *Spectral Flatness*, *Zero Crossing* and *Root Mean Square Power (RMS)* for each sound using the *Librosa* Python library (McFee et al., 2022) with an FFT window size of 2048 and hop length of 512. In addition, the *timbral models* by Pearce et al. (2019) provided quantified measures of *Hardness*, *Depth*, *Brightness*, *Roughness*, *Warmth*, *Sharpness* and *Boominess* which can more easily be related to human perception of sound. The additional feature *RMS Slope*, describing how continuous or intersected a sound is, was quantified by the slope between prominent extrema in the RMS envelope.

The sketch categories described in Section 5 provide a qualitative description of representational approaches. The quantitative set of features described in this section is used to investigate sound-shape associations statistically. Spearman's rank correlation coefficient is used to find out which, if any, visual features correlate with which audio features. As a linear perceptual relationship between features cannot be assumed and the underlying distribution is unknown, the non-parametric Spearman test was chosen




(a) Intersection detection using an adaptation of Bresenham's rasterisation algorithm (Bresenham, 1965). The size of the red dots corresponds to the number of intersections at that point.

(b) Feature point detection with the *ShortStraw* algorithm (Wolin et al., 2008; Xiong & LaViola Jr, 2009). Stroke start/end points are coloured green/blue and corner/curve points are coloured red/yellow.

(c) Different curve shapes according to Xiong and LaViola (Xiong & LaViola Jr, 2009). For this study, a narrow curve was defined with $\alpha < 90°$ and a wide curve with $\alpha \geq 90°$.

Figure 5.: Sketch feature extraction

as it can find not only linear but monotonic relationships in general. The correlation analysis uses mean values of visual features for each sound. To ensure that averaged features still pose meaningful sketch descriptions the inter-rater reliability was determined using the ICC(2,k) model intraclass correlation coefficient (ICC) (Koo & Li, 2016). The ICC(2,k) measures absolute agreement of average raters by averaging responses of k raters for each subject. In this context, sound stimuli were defined as subjects and sketch features as measurements. Sketch features were first log-transformed to meet the normal distribution assumption of the ICC. This approach was chosen because perceptual data typically includes large variance and patterns emerge more clearly when observing averages. This means that results will provide information about the average sound-sketch representation derived from multiple participants, but might not be applicable for predicting or describing sketches of an individual.

## 4. Qualitative feedback analysis

Qualitative feedback was given in a semi-structured interview in Study 1 which provided a broad picture of participants' approaches and laid the groundwork for the development of Study 2. A part of these results is illustrated in Figure 6. In Study 2, brief qualitative feedback was given in written form and supported by quantitative survey responses. This section first focuses on feedback about the task itself and then summarises feedback about the interaction with the interface.

### *4.1. Sketching task*

Task difficulty was reported as easy by 50%, neutral by 21% and hard by 29% of participants in Study 1 which changed to 84%, 9% and 7% in Study 2. Despite the positive skew in Study 2, mixed responses were recorded to the question of whether it was easy to think about sound in a visual way with 50% agreeing, 27% disagree-

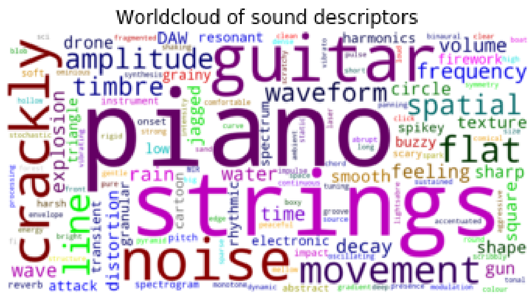



Figure 6.: Reoccurring concepts when describing the sounds in Study 1 extracted from participant interviews.

ing and 23% giving a neutral response. This is well captured in the response 'I have never had to interpret sounds visually, therefore I found it to be kind of a difficult but interesting task.' (P2.70). This was echoed in interviews from Study 1 where some participants found it difficult to 'think of sound in a very visual way' (P1.8) (P1.16). However, for some, it became easier to visualise a sound once the task started, as one participant stated 'I wasn't really expecting to be able to visualize sound, but some of those frequencies were extremely clear to me, as far as how they looked in my brain.' (P2.53). Participants who felt that the task was easy thought that 'there was no right or wrong' (P1.4), 'it was just about being creative' (P1.15), they did not have to 'achieve something' (P1.10), the setup 'allowed the listener space to interpret all sorts of sound visually' (P2.91) or simply found the task 'interesting' (P2.39, P2.41, P2.70, P2.83, P2.91) and 'fun' (P2.20, P2.33, P2.45). For Study 2, some criticized that the 'sounds were very alike' P(2.13) which made it '[...] hard to find an specifically draw[ing] to each sound' P(2.34). On the contrary, in Study 1 some participants struggled with the 'great variety in the sounds' (P1.2) which made it difficult to find a consistent approach. While some participants approached the task as an intuitive, creative activity, others were concerned with establishing a consistent visual language, difficulties arose while deciding which sound characteristics to follow because 'there are too many things to consider' like 'brightness or aggressiveness or how it [timbre] develops over time' (P1.6). Deploying a more systematic rather than intuitive approach appeared more difficult with one participant who did not deviate from their initial concept finding themselves 'going round in circles, and question how valid the whole approach is' (P1.9). Some participants reported that 'complicated ones [sounds] sounded like pictures, and then the simple ones [...] like piano notes were a lot harder to draw' (P1.8) possibly because they 'hear [them] all the time' (P1.1), while other participants thought that 'it's pretty straightforward because I know a piano note more than others' (P1.5). Most participants approached the task by listening to 'the actual sonic qualities of them [the sounds]' (P1.1) and representing them with '[...]abstract patterns, along with patterns that would come from what I thought the instruments were, and it was kind of a mixture of going back and forth between the two.' P(1.5). Familiar sounds like the piano can influence participants to choose figurative representations that include the sound-producing source. However, some participants adopted a figurative approach that does not pay attention to specific sound characteristics but rather extracts general information like an emotion from a sound and then depicts a scene or an object that fits this information. One participant explained their sketch which is shown in Figure 7 as follows: 'I wanted to show the emotion in the sound like the sound was scary so I drew the forest with only one person inside it. If I think the sound is peaceful, I will draw the sea and the sun and the water.' (P1.21). These two different approaches suggest that abstract representations are more strongly informed by sound-shape associations and figurative representations draw more strongly from an emotional response or personal memory.

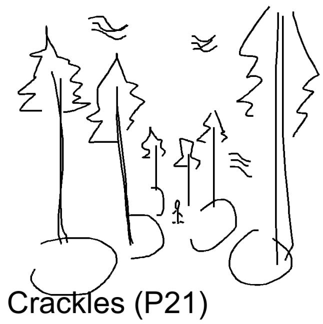



Figure 7.: A participant in Study 1 represented a sound that they perceived as 'scary' with a scene that represented that emotion.

### *4.2. Sketching interface*

Analysing responses for feedback about the interface interaction showed that for Study 1, overall, participants were unsatisfied with the interface with twelve mentioning that they would prefer a pen (either digital or analogue) to be able to sketch more accurately. Six stated that they would have liked to utilise additional visual tools like colour, different strokes and textures. However, responses suggested that while the interface 'could be more expressive [...] for the purpose it was expressive enough' (P1.5). The feedback led to the overhaul of the interface design for Study 2 described in Section 3.3.2. For Study 2, 70 and 75 participants responded with agree or completely agree to the questions *I thought the drawing interface allowed me to be expressive* and *I thought the drawing interface was easy to use.* Three participants still mentioned that they would like to add colours to the interface and one mentioned that the task 'would be easier to do on my phone' (P2.85). A further three participants mentioned that they would like to increase or discard the stroke length limit.

## 5. Sketch Categorisation

From the interview analysis of Study 1 two broad representational approaches can be defined: an abstract approach that is guided at least in part by sound-shape associations and a figurative approach that is strongly influenced by imagery. Sketch categorisation for Study 1 focuses on further formalising these findings and investigating how participants could be guided towards abstract representations. Study 2 focuses on the influence of prominent sound characteristics on abstract representational categories.

### 5.1. Study 1: manual categorisation through card sorting

As described in Section 3.6, sketches were categorised in an open card sorting study. Analysis of the responses returned an optimal number of five categories that were named: *Chaotic/Jagged* (172 sketches), *Radiating/Round* (126), *Lines* (120), *Objects/Scenes* (86) and *Grains* (56). The results are visualised in Figure 8. Descriptive keywords and sketch examples for each category can be found in Table 1. A maximal silhouette coefficient of 0.49 suggests that categories are distinguishable, but not clearly separated which is also reflected by occasionally overlapping keywords. The *Object/Scenes* category consists mainly of figurative representations while the remaining categories include mainly abstract representations. Chi-squared test suggests that non-musicians produce *Objects/Scenes* sketches more often ($\chi^2$(1,N=28)=22.51 p<.0001) while musicians produce *Lines* sketches at a higher rate ($\chi^2$(1,N=28)=7.5 p<.01) possibly because this category contains sketches that appear to reference audio visualisations like envelopes or waveforms. Category counts for *Objects/Scenes* sketches significantly differ between sounds ($\chi^2$(9)=67.07 p<.0001) with post-hoc analysis revealing that *Piano* and *Impact* show significantly higher counts than *Noise*, *String Grains* and *Processed Guitar* (p<.01 for each pair). A possible explanation is that *Piano* and *Impact* have an easily identifiable source that participants attempted to sketch rather than capturing sound characteristics directly. *Noise* and *String Grains* that show high values for the *roughness* audio feature (81 and 56) as displayed in Table A1 also have the largest share of sketches in the *Chaotic/Jagged* category. On the other hand, *Subbass*, *Telephonic* and *Impact* with high values for the *warmth* audio feature (65, 54 and 51) have the highest share of the *Radiating/Round* category. Interestingly, despite high values for *warmth* (54), *Piano* and *Strings* are more frequently represented with *Lines* sketches, possibly because they are comparably spectrally simple sounds which is reflected by the low values for spectral flatness. However, it could also relate to auditory aspects such as pitch and loudness, as it is not feasible to evaluate timbre independently from these attributes. Table 2 shows that *Object/Scenes* has the highest average number of sketch points and the second highest average number of strokes. The highest number of strokes can be found in the *Grains* category which is most prevalent in the sounds *Crackles* and *String Grains*. However, with a low average number of points, this category appears to represent the intersecteness of sound quantified by the audio feature *RMS slope* with multiple short, simple structures. *Chaotic/Jagged* and *Radiating* have a similar average number of points as *Object/Scenes*, but a significantly lower number of average strokes. This analysis indicates that *Object/Scenes* sketches are more likely to be long and complex with multiple components which informed the interface development for Study 2 described in Section 3.3.2.

### 5.2. Study 2: automated categorisation using machine learning

As described in Section 3.6 the *SketchRNN* deep learning architecture was used to create a latent representation of the collected sound-sketches. Figure 10 shows a visualisation of the latent space and categorisation through cluster analysis. In contrast to Study 1, three clusters were determined to separate the data best, however, a silhouette score of 0.36 shows that large overlaps exist between them. Three major categories informed by the card sorting annotations from Study 1 were defined: *Lines* (610), *Chaotic/Jagged* (455 sketches), and *Radiating/Round* (694). The categories *Object/Scenes* and *Grains* were no longer present in the dataset. This was expected as the

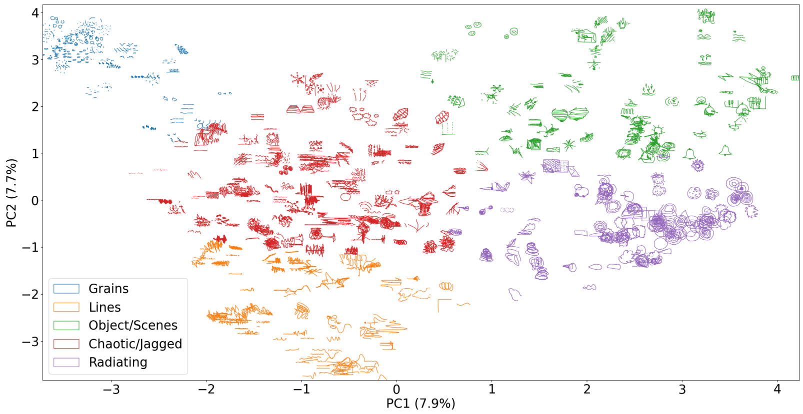



Figure 8.: Study 1 sound sketches organised by the K-Means clusters calculated from the card-sorting study. Colours indicate the different clusters.

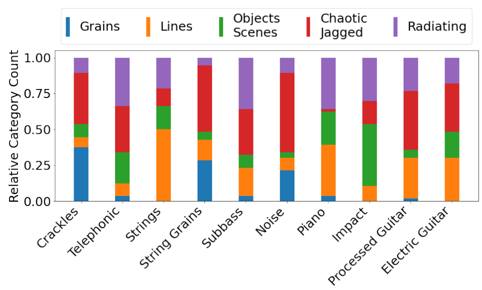



Figure 9.: Categories by sound stimulus in Study 1.

interface design discouraged *Objects/Scenes* sketches and *Grains* sketches were largely found in intersected sounds that were not included in the audio stimuli for Study 2. To compare differences in category distribution, six sound groups were created from sounds from the annotated attributes *bright*, *rough* and *thick* derived from the FM synthesiser sound dataset by Hayes and Saitis (2020) discussed in Section 2.2. Each group contains three sounds with either the highest or lowest value for an attribute. Further analysis showed that category counts for *Chaotic/Jagged* and *Lines* differ significantly between sound groups ($\chi^2(5)$=20.38 p<.01 and $\chi^2(5)$=34.29 p<.00001), but no significant differences could be found for *Radiating/Round* ($\chi^2(5)$=9.59 p>.05). Post-hoc analysis showed prominent differences in category distribution between the most and least rough sounds as illustrated in Figure 11 supporting findings from Study 1 which showed that rough sounds were more frequently represented with chaotic, complex sketches and calmer, less rough sounds with simpler lines. Measurements for roughness which were extracted automatically with *timbral model* by Pearce et al. (2019) and presented in Table A1 were high for the *rough* sound group (70, 63 and 59 for Synth 9,13,19) and low for the *not rough* sound group (29,0,45 for Synth 1,11,14). This further supports the validity of these features as measurements relevant to human perception.

| Grains | Lines | Object/ Scenes | Chaotic/ Jagged | Radiating/ Round |
|---|---|---|---|---|
| *Small, repeated, grainy, spots, multiple components, layers, abstract, distinct* | *round, soft, continuous, jagged, irregular, simple, single, lines* | *real-life objects, environment, actions or feelings, abstract structures* | *chaotic, intense, jagged, multiple layers, single objects* | *round, circular, spiral, sharp, shaking, distinct objects, radiating, natural* |

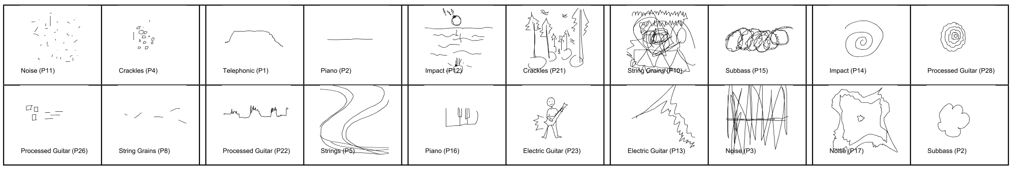



Table 1.: Sketch categories with examples. Category names and keywords were obtained through thematic analysis as described in Section 3.6. *Objects/Scenes* mainly refers to real-world associations while other categories highlight different abstract approaches, but category clusters might overlap with a number of sketches showing characteristics of more than one category. Colours were inverted for better visibility.

| | Grains | | Lines | | Object/ Scenes | | Chaotic/ Jagged | | Radiating/ Round | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| **Points** | 560 | 532 | 534 | 338 | 1068 | 557 | 1060 | 696 | 914 | 722 |
| **Strokes** | 14 | 12 | 2 | 2 | 13 | 9 | 7 | 10 | 4 | 4 |

Table 2.: Average number of points and strokes for each sketch category rounded to the closest integer. *Object/Scenes* shows the highest number of points and second highest number of strokes which implies that this category could be reduced when limiting the size of a sketch.

## 6. Quantitative feature analysis

This analysis was conducted with the aim of statistically investigating to what extent sound-shape associations emerge from sound-sketches. First, inter-rater reliability was determined to confirm sketch features used in this research can measure agreement between participants. This was followed by calculating correlations between individual audio and visual features. As discussed in Section 3.7, the interpretation of the results needs to consider that this analysis uses averaged values

### *6.1. Inter-rater reliability*

The results of the ICC(2,k) inter-rater reliability measures are illustrated in Figure 14. For Study 1, reliability measures were good to excellent for *Intersections* and *Acute Angles*, poor to good for *Average Speed* and moderate to good for all remaining features within the 95% confidence interval (CI). For Study 2, measures were excellent for *Acute Angles* and good to excellent for *Intersections*, *Wide Curves* and *Right Angles* solidifying findings from Study 1 with higher reliability scores and narrower confidence intervals. In contrast to Study 1, *Number of Strokes* and *Average Time* only returned

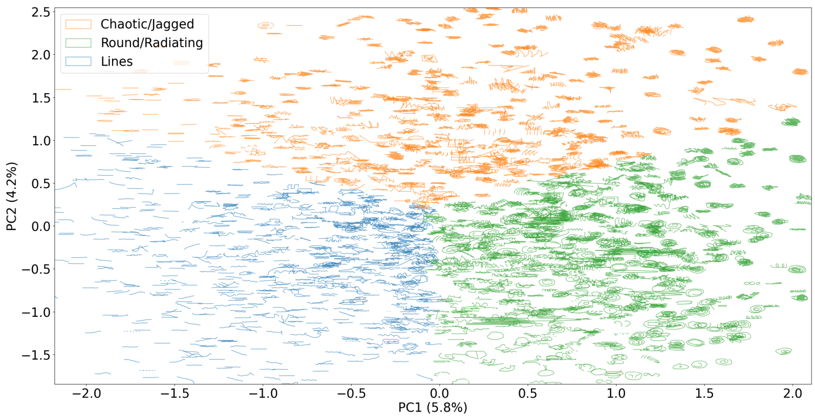



Figure 10.: Study 2 sound sketches organised in the latent space that was reduced to two dimensions with PCA. Colours indicate the different clusters found through K-Means.

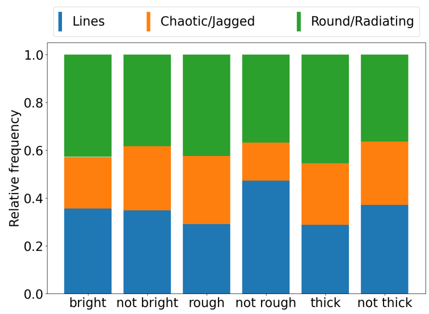
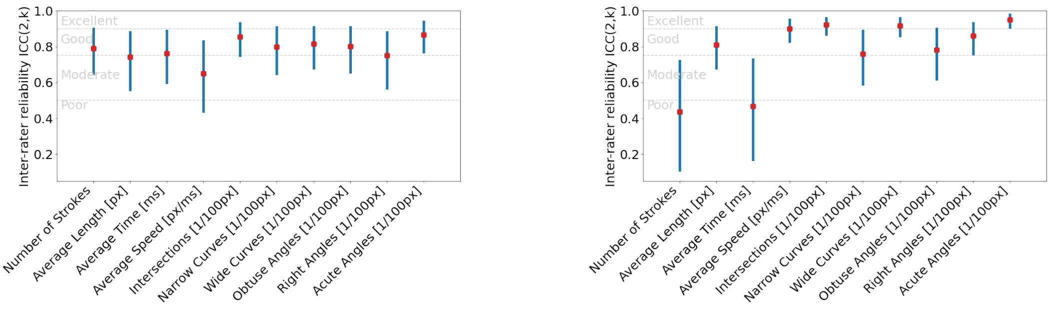



Figure 11.: Categories by sounds grouped by attributes in Study 2.

poor to moderate reliability which can be accredited to the sketch length limit of the interface. *Average Speed* however showed good to excellent reliability compared to poor to good in Study 1 implying that participants might have expressed characteristics through their sketching speed that would have been expressed through longer, more complex sketches with an unrestricted interface. This is supported by a larger variance in average sketching speed compared to Study 1 (SD = 0.23 compared to SD = 0.12). These results suggest that some level of agreement exists between averaged participants on how to represent sounds visually and that it can be measured with the extracted sketch features.

### *6.2. Feature correlation*

Several significant correlations were found between sketch and audio features in both studies illustrated in Figure 13. For Study 1, *Acute Angles* (11), *Intersections* (9) and *Number of Strokes* (8) show the highest statistically significant ($p<.05$) number of strong ($r>.6$) and very strong ($r>.8$) correlations with audio features. The strongest correlation overall was found between *RMS Mean* and *Average Time* ($r=.95$, $p<.001$). Opposing audio features like *Warmth* and *Sharpness* showed similar absolute corre-




(a) Study 1 (b) Study 2

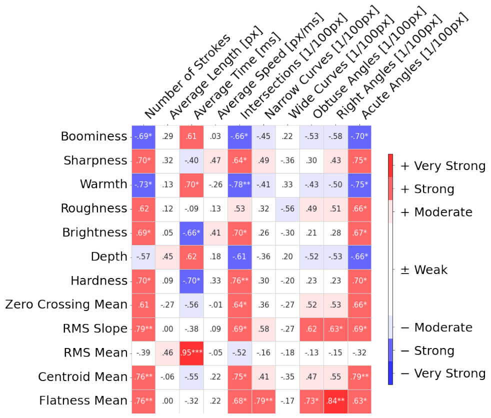
Figure 12.: Mean values and 95% CI of ICC(2,k) inter-rater reliabilities for each sketch feature with evaluation guidelines proposed by Koo and Li (2016) (df1=19, df2=513, p<.01 for all features).

lation values but opposite directions for *Number of Strokes*, *Intersections* and *Acute Angles*. For Study 2, *Wide Curves* (9), *Intersections* (7) and *Acute Angles* (3) show the highest statistically significant (p<.05) number of strong (r>.6) and very strong (r>.8) correlations with audio features. The strongest correlation overall was found between *Intersections* and *Hardness* (r=.81, p<.001). While only 12 strong and very strong correlations were found compared to 19 in Study 1, more results were significant at p<.05 level (49 compared to 37). Similar trends can be observed between both studies with *Acute Angles* and *Intersections* showing negative correlations with audio features *Boominess*, *Warmth* and *Depth* and positive correlations with *Roughness*, *Brightness* and *Hardness*. *Average Speed* shows similar correlations to *Number of Strokes* in Study 1 for example with *Hardness* and *Boominess* further supporting the hypothesis that sketching speed compensated for the sketch length restrictions. In contrast to Study 1, *Wide Curves* shows a large number of significant correlations with audio features that are mirroring *Acute Angles* correlations as expected from known sound-shape associations. As all sound stimuli were created with the same amplitude envelope in Study 2 the features *RMS Slope* and *RMS Mean* did not provide distinguishing descriptions and consequently did not show any significant correlations with sketch features.

## 7. Discussion

### *7.1. Abstract and figurative representations of sound stimuli*

The sound-sketches collected in this research were categorised through a rigorous, human-centred process consisting of participant interviews and a card-sorting study. On the highest level, sound-sketches were divided into two groups: *abstract* and *figurative*. Abstract representations appear to be strongly informed by cross-modal associations and show many similarities to visual stimuli used in matching tasks. Figurative representations, on the other hand, depict objects or scenes associated with a sound that may depict the sound source directly, for example in form of a musical instrument, or may be informed by an emotional response or memory, for example, a sound perceived as scary might be represented with a scene similar to the sketch in Figure 7 that is informed by the memory of a movie scene that corresponds to that emotion. Which representational approach a participant took was influenced by the




(a) Study 1

(b) Study 2

Figure 13.: Spearman's rank correlation coefficients between sketch and audio features with annotated p-values: p<.05 (*), .01 (**), .001 (***)

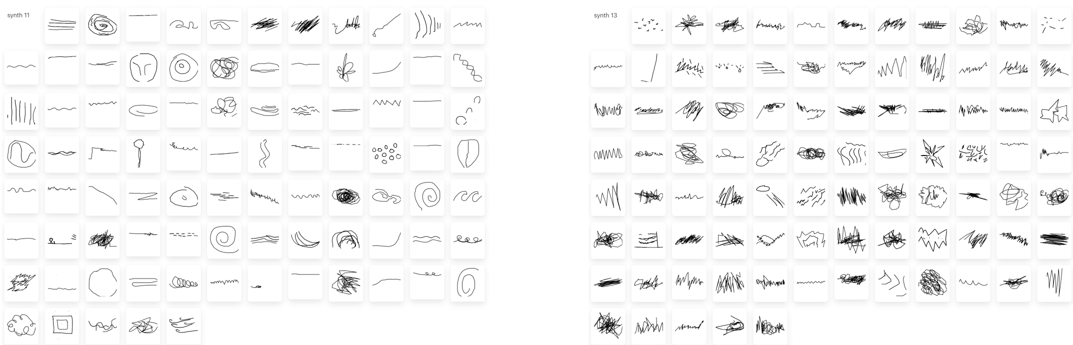
sound type or a participant's experience. Figurative representations were found to be more prevalent among non-musicians and for sounds with an easily identifiable sound source like a musical instrument. This could be interpreted as a difference in listening modes: coined by Schaeffer (2017) and further explored by Chion (2019), the *reduced listening* mode describes a focus on the sound itself as opposed to *semantic listening* which focuses on the source or meaning of a sound. Described as hardly natural by Chion (2019), *reduced listening* can be more demanding for the untrained ear, especially for sounds from a familiar source. This analysis suggests that two different approaches would be needed when computationally mapping sketches to sound. For abstract sketches that encode information about sound characteristics directly in their form, extracting quantitative sketch features like the ones described in Section 3.7 appears to be a viable approach. For figurative sketches, a more suitable approach might have to include object recognition for sketches of musical instruments and other sound-producing objects or sentiment analysis for emotionally informed scenes.

Section 6 shows that for Study 1 multiple significant correlations were found between audio and sketch features despite including abstract and figurative sketches. With 84% of sketches classified as abstract, figurative sketches did not appear to impact these correlations. Because of the small sample size, differences in correlations between abstract and figurative sketches could not be investigated in a meaningful way. Figurative representations might still be indirectly influenced by cross-modal associations, for example, an uncomfortable noisy and dissonant sound might be represented with a scene similar to the design *Landscape of Thorns* by Moisey (2017) that aims to communicate discomfort and danger through sharp and jagged structures. On the other hand, abstract representations might include references to symbolic representations of sound that are not based on cross-modal associations. The feedback analysis in Section 4 reveals that some sketches, in particular in the *Lines* category, are informed by audio representations like waveforms, spectrograms or amplitude envelopes.

As described in Section 2.1, cross-modal associations are not fixed; they emerge in different forms depending on the stimulus and situation and sketch representations might incorporate multiple associations or references to objects, scenes or symbols. Without explicitly asking a participant what they had in mind, it may not be possible to exactly determine their underlying motivation. Generally, only a few participants strictly adhered to one representational approach. While dominated by a primary approach, most participants switched between or mixed representational styles throughout the sketching task.

### *7.2. Cross-modal associations in free-form sketches*

Compared to the figurative category *Object/Scenes* the abstract categories *Chaotic/-Jagged*,*Grains*,*Lines* and *Radiating/Rounds* appear to be primarily influenced by cross-modal associations. This research aimed to investigate sound-shape associations specifically, but it might be useful to define the terminology to better interpret the results. The visual stimuli used in sound-shape matching tasks like Adeli et al. (2014) focus very strictly on shape meaning that stimuli only show the outlining *form* through a non-intersecting connected series of lines. Looking at Figures 8 and 10 it is obvious that participants did not only use *form* in their abstract sketches. Knees and Andersen (2016) made similar findings and, in their sound-sketch prototype, included tools to create outlines of simple forms like circles and triangles that can be given more *complexity* through free-form lines and filled with *textures*.



(a) Synth 11 (b) Synth 13

Figure 14.: Sound-sketches taken from Study 2 for the stimuli Synth 11, characterized by high values for *Warmth* and low values for *Roughness* and *Spectral Flatness*, and Synth 13, which exhibits contrasting values for these attributes. There is a difference in form, with Synth 11 more frequently depicted as *Radiating/Round* and Synth 13 as *Chaotic/Jagged* in sketches. In addition, Synth 11 tends to be represented by less complex sketches that fall into the *Lines* category.

In this research, the abstract categories *Chaotic/Jagged* and *Radiating/Round* appear to encode sound characteristics through their *form* with the latter found more frequently for warm sounds in Study 1 and the former for rough sounds in both studies which aligns with existing sound-shape research (Köhler, 1929; Adeli et al., 2014). However, *Chaotic/Jagged* appears to be also contrasted by *Lines* which is more frequently found in the least rough sounds. Sketches in the *Lines* category have considerably fewer sharp angles than sketches in *Chaotic/Jagged* which could be interpreted as a difference in *form*, but *Lines* also exhibits considerably fewer stroke intersections than *Chaotic/Jagged* which can be interpreted as a difference in *complexity*. The remaining abstract category *Grains* was particularly prevalent for the intersected sounds *Crackles* and *String Grains* in Study 1 and is described to consist of small, repetitive components which can be interpreted as *texture* and was quantified by the average number of strokes and the average stroke length. Strictly speaking, sound-shape associations are not the only cross-modal associations that influence abstract representations, however, in this context, the definition of *shape* could be extended to describe any 2-dimensional structure composed of straight or curved lines. For quantitative analysis, this broader definition of shape might be sufficient as the sketch features described in Section 3.7 capture more than just the *form* of a sketch.

The results of the feature correlation analysis in Section 6.2 support that typical sound-shape associations influenced sketch representations with acute angles positively correlating with attributes like *Sharpness*, *Roughness* and *Brightness* and negatively with *Warmth* and *Boominess*. The sketch feature *Wide Curves* which was expected to quantify roundness did not show any significant correlations in Study 1, but multiple significant correlations for Study 2 that point in the opposite direction of *acute angles*. This might be due to noise introduced by higher variance in sketch approaches in Study 1 but could suggest that the extracted features are not a good measure for the overall roundness of a sketch. *Intersections* proved to be meaningful in describing cross-modal associations showing multiple significant correlations with audio features across both studies. The direction of these correlations is the same for acute angles, leading to the belief that the opposing pairs *rough* and *soft* for sound might not only be visualised through jaggedness or roundness but also through complexity and simplicity. Overall, the roughness of a sound appears to have a strong influence on how sketches are represented which is clearly visualised in Figure 11. The figure also suggests that the *Lines* category is more prevalent for *thin* sounds, but the differences found in this study did not prove to be significant. As the sketch and audio features do not directly quantify thinness, feature correlation could not provide any additional information. Further work could focus on this attribute of sound which would provide an additional timbral dimension encoded in sound-sketch representations.

A general problem that emerged in both studies is that it cannot be clearly determined which sound characteristic participants focused on. As a participant in Study 1 stated, there are a lot of different aspects of sound and it is difficult to represent them all in a simple sketch. This might be made even more difficult with the sketch length limit that was introduced in Study 2. While it succeeded in guiding participants to represent sounds in more abstract ways, it does prevent them from elaborating or overlaying multiple approaches that might capture more aspects of a sound. It might be possible that thinness or thickness can be encoded through sketching, but participants mainly focused on roughness when representing sound. For future research, it could be considered to also ask participants to rate which attributes of the sound they focused on with a design similar to Hayes et al. (2021). An alternative method might involve prompting participants to envision a sound possessing specific attributes instead of exposing them to an actual auditory stimulus. This approach was employed during the test session in Study 2, where participants were tasked with sketching a noisy and calm sound. However, understanding how participants mentally conceptualize a sound remains somewhat ambiguous, particularly considering that certain common auditory descriptors, such as ”sharpness“ or ”roundness“ may inherently include visual elements. In addition to verbal feedback, allowing participants to select a sound that most accurately captures their mental representation of a descriptor could offer valuable insights into their attribute assignments.

### 7.3. Implications for development of a sketch-based sound synthesiser

Study 2 was designed with a future implementation of a sketch-based sound synthesiser in mind. The results can shed light on which mapping architecture is more appropriate for such a system:

- a regression approach where a change in timbre is induced by an incremental change in the corresponding sketch feature. For example, a sound would become noisier with an increase in sharp angles.
- a classification approach where the overall timbre category is determined. In this scenario, a sketch could represent multiple categories for example either *rough* or *soft* in combination with either *thick* or *thin*.

The feature correlations in Section 6.2 describe a monotonic relationship between audio and sketch features which can lead to the conclusion that a gradual increase in an audio feature like roughness coincides with a gradual increase in a sketch feature like *acute angles*. However, it has to be remembered that these sketch features represent the average sketch of the participants meaning that a single participant might not follow a change of timbre in such an incremental fashion. Rather participants could think of sounds in categories and with rising roughness, an increasing number of participants switch from round to jagged shapes. Similarly, statistical agreement for sketch representations was only determined for an average participant in Section 6.1 and, in this research, significant results could not be obtained when looking at individual participants. In addition, if a specific, computed audio feature increases linearly for a series of sounds, it does not mean that humans would perceive this as a linear change in timbre. When thinking about a sound that is neither particularly rough nor soft, it is likely that another more prominent feature, for example, *thick* or *thin* would inform the perception of this sound and hence the sketch representation of it. In fact, Hayes et al. (2022b) found that relatively small, localised clusters emerge when participants were asked to define sound within a timbre space according to descriptors like *rough* or *soft*. A linear interpolation between the parameters of two FM synthesiser sounds might not be perceived as a linear transition in sound characteristics, but rather different sound environments that emerge along this path. Given this analysis, a categorisation approach appears to be more viable when designing a sketch-based sound synthesiser. It should also be added that sketch to synthesiser mappings do not have to be limited to timbre. An advantage of this design is that multiple parameters can be represented with a single input. Revisiting the research of Salgado-Montejo et al. (2016) and M. B. Küssner et al. (2014) shows that sketch representations can also serve to represent pitch or amplitude envelopes. Future work can explore how a sketch input could manipulate these parameters at the same time, potentially using a combination of mappings that are established through a data-driven machine learning approach and hard-coded mappings that are based on findings from empirical studies. Given that colour and visual texture have been identified as common associations with timbre in previous studies (Ward et al., 2006; Adeli et al., 2014; Gurman et al., 2021; Grill & Flexer, 2012), the exploration of polychromatic colour palettes, different stroke widths and brush styles could be considered in further development to facilitate multi-dimensional mappings.

## 8. Conclusion

In two participant studies, 2320 free-form sound-sketches (560 from Study 1 and 1760 from Study 2) were collected with simple, monochromatic digital sketching interfaces. Through a rigorous human-centred analysis sketches were categorised by their primary representational approach. Several significant correlations between quantitative audio and sketch features were found that align with findings from cross-modal matching tasks. The results show that while sound-shape associations play a significant role in sketched representations, humans incorporate other visual aspects like structural complexity or texture as well or choose figurative representations of emotions or sound-producing objects. Some level of agreement on how to represent sounds could be found between participants which appears strongest for sounds that are dominated by one sound characteristic. This research provides useful insights and suggestions for designing a sketch-based sound synthesiser.

## 9. Acknowledgments

EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1).

## 10. Disclosure statement

All studies that contributed to this research were approved by the Queen Mary University of London Ethics Committee (Reference numbers QMREC2341 for Study 1 and QMERC2517a the card sorting exercise, QMERC20.478 for participant evaluation of the interface development and QMERC20.009 for Study 2). All participants gave informed consent. No monetary compensation was offered except for Study 2 which was conducted via Prolific rewarding £2.50 for 20-minute participation. The authors report there are no competing interests to declare.

## References

Adeli, M., Rouat, J., & Molotchnikoff, S. (2014). Audiovisual correspondence between musical timbre and visual shapes. *Frontiers in human neuroscience*, *8*, 352. doi:

Albertazzi, L., Canal, L., & Micciolo, R. (2015). Cross-modal associations between materic painting and classical spanish music. *Frontiers in Psychology*, *6*, 424. doi:

ANSI. (1994). Timbre. *American National Standards Institute. Psychoacoustic terminology.*

Baron-Cohen, S., Burt, L., Smith-Laittan, F., Harrison, J., & Bolton, P. (1996). Synaesthesia: prevalence and familiality. *Perception*, *25*(9), 1073–1079.

Blake, D. K. (2012, June). Timbre as Differentiation in Indie Music. *Music Theory Online*, *18*(2). doi:

Bottini, R., Barilari, M., & Collignon, O. (2019). Sound symbolism in sighted and blind. the role of vision and orthography in sound-shape correspondences. *Cognition*, *185*, 62–70. doi:

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. doi:

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? a remote culture make similar shape–sound matches, but different shape–taste matches to westerners. *Cognition*, *126*(2), 165–172. doi:

Bresenham, J. E. (1965). Algorithm for computer control of a digital plotter. *IBM Systems journal*, *4*(1), 25–30.

Bruford, F., Barthet, M., McDonald, S., & Sandler, M. B. (2019). Groove explorer: An intelligent visual interface for drum loop library navigation. In *Proceedings of the acm iui workshops.* Los Angeles, USA: CEUR-WS.org.

Catmull, E., & Rom, R. (1974). A class of local interpolating splines. In *Computer aided geometric design* (pp. 317–326). Elsevier. doi:

Chion, M. (2019). 2. the three listening modes. In *Audio-vision: Sound on screen* (pp. 22–34). New York Chichester, West Sussex: Columbia University Press. doi:

Clemente, A., Vila-Vidal, M., Pearce, M. T., Aguiló, G., Corradi, G., & Nadal, M. (2020). A set of 200 musical stimuli varying in balance, contour, symmetry, and complexity: Behavioral and computational assessments. *Behavior Research Methods*, *52*(4), 1491–1509.

Cramer, J., Wu, H.-H., Salamon, J., & Bello, J. P. (2019, May). Look, listen and learn more: Design choices for deep audio embeddings. In *Ieee int. conf. on acoustics, speech and signal processing (icassp)* (pp. 3852–3856). Brighton, UK.

Cuskley, C., Dingemanse, M., Kirby, S., & Van Leeuwen, T. M. (2019). Cross-modal associations and synesthesia: Categorical perception and structure in vowel–color mappings in a large online sample. *Behavior Research Methods*, *51*, 1651–1675. doi:

Davis, R. (1961). The Fitness of Names to Drawings. a Cross-Cultural Study in Tanganyika. *British Journal of Psychology*, *52*(3), 259–268. doi:

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, *29*(6), 141–142.

Douglas, D. H., & Peucker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, *10*(2), 112–122. doi:

Engeln, L., & Groh, R. (2020). Cohearence of audible shapes—a qualitative user study for coherent visual audio design with resynthesized shapes. *Personal and Ubiquitous Computing*, 1–11. doi:

Engeln, L., Le, N. L., McGinity, M., & Groh, R. (2021). Similarity analysis of visual sketch-based search for sounds. In *Proceedings of audio mostly 2021* (pp. 101–108). Trento, Italy: Association for Computing Machinery. doi:

Fried, O., Jin, Z., Finkelstein, A., & Oda, R. (2014). *AudioQuilt: 2D Arrangements of Audio Samples using Metric Learning and Kernelized Sorting.*

Garber, L., y Ciencia, M. A., Ciccola, T., & Amusategui, J. C. (2021). Audiostellar, an open source corpus-based musical instrument for latent sound structure discovery and sonic experimentation. In *Proceedings of international computer music conference* (pp. 86–91). Santiago, Chile. Retrieved from `https://hdl.handle.net/2027/fulcrum.t435gg568`

Giannakis, K. (2006). A comparative evaluation of auditory-visual mappings for sound visualisation. *Organised Sound*, *11*(3), 297–307.

Giannakis, K., & Smith, M. (2000). Towards a theoretical framework for sound synthesis based on auditoryvisual associations.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *the Journal of the Acoustical Society of America*, *61*(5), 1270–1277.

Grill, T., & Flexer, A. (2012). Visualization of Perceptual Qualities in Textural Sounds. In *Proceedings of international computer music conference* (pp. 589–596). Ljubljana, Slovenia: Michigan Publishing Services. Retrieved from `http://hdl.handle.net/2027/spo.bbp2372.2012.110`

Gurman, D., McCormick, C. R., & Klein, R. M. (2021). Crossmodal correspondence between auditory timbre and visual shape. *Multisensory Research*, *35*(3), 221–241.

Ha, D., & Eck, D. (2017). A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*.

Hayes, B., & Saitis, C. (2020). There's more to timbre than musical instruments: semantic dimensions of FM sounds. In *Proceedings of international conference on timbre.* Thessaloniki, Greece: Timbre 2020.

Hayes, B., Saitis, C., & Fazekas, G. (2021). Perceptual and semantic scaling of fm synthesis timbres: Common dimensions and the role of expertise. *ICMPC-ESCOM*.

Hayes, B., Saitis, C., & Fazekas, G. (2022a). Disembodied timbres: A study on semantically prompted fm synthesis. *Journal of the Audio Engineering Society*, *70*(5), 373–391.

Hayes, B., Saitis, C., & Fazekas, G. (2022b). timbre.fun: A gamified interactive system for crowdsourcing a timbre semantic vocabulary. In *Proceedings of the 24th international congress on acoustics* (p. 10).

Iverson, P., & Krumhansl, C. (1993). Isolating the dynamic attributes of musical timbrea. *The Journal of the acoustical society of America*, *94*(5), 2595–2603. doi:

Knees, P., & Andersen, K. (2016). Searching for audio by sketching mental images of sound: A brave new idea for audio retrieval in creative music production. In *Proceedings of international conference on multimedia retrieval* (pp. 95–102). New York, USA: Association for Computing Machinery. doi:

Köhler, W. (1929). Gestalt psychology. *Liveright*.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155–163.

Küssner, M. (2014). *Shape, drawing and gesture: Cross-modal mappings of sound and music* (Unpublished doctoral dissertation). King's College London (University of London).

Küssner, M. B., Tidhar, D., Prior, H. M., & Leech-Wilkinson, D. (2014). Musicians are more consistent: Gestural cross-modal mappings of pitch, loudness and tempo in real-time. *Frontiers in Psychology*, *5*. doi:

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Lewis, J. R. (2018, July). The System Usability Scale: Past, Present, and Future. *International Journal of Human–Computer Interaction*, *34*(7), 577–590. doi:

Löbbers, S., Barthet, M., & Fazekas, G. (2021). Sketching sounds: an exploratory study on sound-shape associations. In *Proceedings of international computer music conference* (pp. 299–304). Santiago, Chile: Michigan Publishing Services. Retrieved from `https://hdl.handle.net/2027/fulcrum.t435gg568`

Löbbers, S., & Fazekas, G. (2022). Seeing sounds, hearing shapes: a gamified study to evaluate sound-sketches. In *Proceedings international computer music conference* (pp. 174–179). Limerick, Ireland: Michigan Publishing Services. Retrieved from https://hdl.handle.net/2027/fulcrum.nk322g689

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & psychophysics*, *24*(3), 253–257. doi:

Marin, M. M., Schober, R., Gingras, B., & Leder, H. (2017). Misattribution of musical arousal increases sexual attraction towards opposite-sex faces in females. *PLoS One*, *12*(9), e0183531. doi:

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound–shape correspondences in toddlers and adults. *Developmental Science*, *9*(3), 316–322. doi:

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological research*, *58*, 177–192.

McFee, B., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., ... Thassilo (2022, February). *librosa/librosa: 0.9.1.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.6097378 doi:

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (Vol. 8, pp. 18–25).

Moisey, A. (2017). Permanent negative value: The waste isolation pilot plant. *Critical Inquiry*, *43*(4), 861–892. doi:

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS one*, *9*(2), e89642.

Nielsen, A. K. S., & Rendall, D. (2013). Parsing the role of consonants versus vowels in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *67*(2), 153–163. doi:

Paea, S., & Baird, R. (2018). Information architecture (ia): Using multidimensional scaling (mds) and k-means clustering algorithm for analysis of card sorting data. *Journal of Usability Studies*, *13*(3).

Pearce, A., Brookes, T., & Mason, R. (2019). Modelling Timbral Hardness. *Applied Sciences*, *9*(3), 466. doi:

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the cuidado project. *CUIDADO Ist Project Report*, *54*(0), 1–25.

Provenzano, C. (2018). *Auto-Tune, Labor, and the Pop-Music Voice.* Oxford University Press.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia–a window into perception, thought and language. *Journal of consciousness studies*, *8*(12), 3–34.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Rucsanda, M. D., et al. (2019). Aspects of the relationship between music and painting and their influence on schoenberg and kandinsky. *Bulletin of the Transilvania University of Braşov, Series VIII: Performing Arts*, *12*(2), 91–100.

Saitis, C., & Weinzierl, S. (2019). The semantics of timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, perception,* *and cognition* (pp. 119–149). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-14832-4_5 doi:

Saitis, C., Weinzierl, S., von Kriegstein, K., Ystad, S., & Cuskley, C. (2020). Timbre semantics through the lens of crossmodal correspondences: A new way of asking old questions. *Acoustical Science and Technology*, *41*(1), 365–368.

Salgado-Montejo, A., Marmolejo-Ramos, F., Alvarado, J. A., Arboleda, J. C., Suarez, D. R., & Spence, C. (2016). Drawing sounds: representing tones and chords spatially. *Experimental Brain Research*, *234*, 3509–3522.

Schaeffer, P. (2017). *Treatise on musical objects: An essay across disciplines* (Vol. 20). Univ of California Press.

Seago, A. (2013). A new interaction strategy for musical timbre design. In S. Holland, K. Wilkie, P. Mulholland, & A. Seago (Eds.), *Music and human-computer interaction* (pp. 153–169). London: Springer London. Retrieved from https://doi.org/10.1007/978-1-4471-2990-5_9 doi:

Sezgin, T. M. (2001). *Feature point detection and curve approximation for early processing of free-hand sketches* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Shinohara, K., Yamauchi, N., Kawahara, S., & Tanaka, H. (2016, September). Takete and Maluma in Action: A Cross-Modal Relationship between Gestures and Sounds. *PLOS ONE*, *11*(9), e0163525. doi:

Sidhu, D. M., & Pexman, P. M. (2018). Five mechanisms of sound symbolic association. *Psychonomic bulletin & review*, *25*, 1619–1643.

Siedenburg, K., Fujinaga, I., & McAdams, S. (2016). A comparison of approaches to timbre descriptors in music information retrieval and music psychology. *Journal of New Music Research*, *45*(1), 27–41.

Siedenburg, K., & McAdams, S. (2017, October). Four Distinctions for the Auditory "Wastebasket" of Timbre1. *Frontiers in Psychology*, *8*, 1747. doi:

Soraghan, S., Faire, F., Renaud, A., & Supper, B. (2018). A new timbre visualization technique based on semantic descriptors. *Computer Music Journal*, *42*(1), 23–36.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*, 971–995.

Stuckey, H. L. (2015, June). The second step in data analysis: Coding qualitative research data. *Journal of Social Health and Diabetes*, *03*(1), 7–10. doi:

Taylor, I. K., & Taylor, M. M. (1962). Phonetic symbolism in four unrelated languages. *Canadian Journal of Psychology*, *16*(4), 344–356. doi:

Thoret, E., Aramaki, M., Bringoux, L., Ystad, S., & Kronland-Martinet, R. (2016, April). Seeing Circles and Drawing Ellipses: When Sound Biases Reproduction of Visual Motion. *PLOS ONE*, *11*(4), e0154475. doi:

Van Doorn, G. H., Wuillemin, D., & Spence, C. (2014). Does the colour of the mug influence the taste of the coffee? *Flavour*, *3*(1), 1–7.

Wallmark, Z., Frank, R. J., & Nghiem, L. (2019). Creating novel tones from adjectives: An exploratory study using fm synthesis. *Psychomusicology: Music, Mind, and Brain*, *29*(4), 188.

Wallmark, Z., & Kendall, R. A. (2021, 10). 578579C23Describing Sound: The Cognitive Linguistics of Timbre. In *The Oxford Handbook of Timbre.* Oxford University Press. Retrieved from https://doi.org/10.1093/oxfordhb/9780190637224.013.14 doi:

Ward, J., Huckstep, B., & Tsakanikos, E. (2006). Sound-colour synaesthesia: to what extent does it use cross-modal mechanisms common to us all? *Cortex*, *42*(2),

264–280.
Wolin, A., Eoff, B., & Hammond, T. (2008). ShortStraw: A Simple and Effective Corner Finder for Polylines. In *Proceedings of eurographics workshop on sketch-based interfaces and modeling.* Annecy, France: The Eurographics Association. doi:
Xiong, Y., & LaViola Jr, J. J. (2009). Revisiting shortstraw: improving corner finding in sketch-based interfaces. In *Proceedings of eurographics symposium on sketch-based interfaces and modeling* (pp. 101–108). New Orleans USA: Association for Computing Machinery. doi:
Zacharakis, A., & Pastiadis, K. (2015). A confirmatory approach of the luminance-texture-mass model for musical timbre semantics. In *Proceedings of the audio mostly 2015 on interaction with sound.* New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/2814895.2814898` doi:
Zacharakis, A., & Pastiadis, K. (2016). Revisiting the luminance-texture-mass model for musical timbre semantics: A confirmatory approach and perspectives of extension. *Journal of the Audio Engineering Society*, *64*(9), 636–645.

## Appendix A. Sound and sketch feature extraction

Table A1 shows the values of extracted audio features for Study 1 and Study 2.

Table A2 show the values of extracted sketch features for Study 1 and Study 2.

| | Flatness Mean | Centroid Mean | RMS Mean | RMS Slope | Zero Cross. Mean | Hard-ness | Depth | Bright-ness | Rough-ness | Warmth | Sharp-ness | Boomi-ness |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Crackles | $4.1 * 10^{-2}$ | 1728 | 0.032 | 21 | 0.035 | 55 | 28 | 60 | 52 | 41 | 48 | 17 |
| Telephonic | $1.0 * 10^{-3}$ | 543 | 0.157 | 5 | 0.013 | 34 | 67 | 43 | 50 | 54 | 28 | 41 |
| Strings | $3.9 * 10^{-5}$ | 1042 | 0.151 | 10 | 0.020 | 42 | 57 | 52 | 52 | 54 | 35 | 33 |
| String Grains | $1.9 * 10^{-3}$ | 1177 | 0.107 | 22 | 0.031 | 52 | 40 | 56 | 56 | 49 | 37 | 25 |
| Subbass | $1.5 * 10^{-8}$ | 206 | 0.416 | 1 | 0.002 | 34 | 77 | 36 | 47 | 65 | 34 | 47 |
| Noise | $2.9 * 10^{-1}$ | 7915 | 0.086 | 34 | 0.144 | 81 | 44 | 81 | 81 | 29 | 68 | 22 |
| Piano | $4.5 * 10^{-7}$ | 542 | 0.051 | 1 | 0.013 | 42 | 65 | 49 | 40 | 54 | 31 | 42 |
| Impact | $2.8 * 10^{-4}$ | 1047 | 0.097 | 8 | 0.011 | 65 | 77 | 60 | 49 | 51 | 49 | 39 |
| Processed Guitar | $1.7 * 10^{-3}$ | 229 | 0.344 | 3 | 0.002 | 30 | 80 | 39 | 40 | 63 | 37 | 46 |
| Electric Guitar | $4.5 * 10^{-6}$ | 1295 | 0.305 | 5 | 0.022 | 47 | 56 | 56 | 56 | 50 | 42 | 34 |

(a) Study 1

| | Flatness Mean | Centroid Mean | RMS Mean | RMS Slope | Zero Cross. Mean | Hard-ness | Depth | Bright-ness | Rough-ness | Warmth | Sharp-ness | Boomi-ness |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Synth 1 | $4.9 * 10^{-8}$ | 562 | 0.0429 | 0 | 0.0199 | 34 | 42 | 39 | 29 | 39 | 29 | 21 |
| Synth 2 | $1.4 * 10^{-7}$ | 794 | 0.0614 | 0 | 0.0239 | 52 | 65 | 54 | 53 | 50 | 37 | 34 |
| Synth 3 | $3.8 * 10^{-5}$ | 2889 | 0.0266 | 0 | 0.113 | 67 | 42 | 75 | 72 | 30 | 61 | 9 |
| Synth 4 | $5.5 * 10^{-3}$ | 9670 | 0.0529 | 0 | 0.437 | 73 | 31 | 86 | 83 | 17 | 77 | -6 |
| Synth 5 | $1.6 * 10^{-8}$ | 247 | 0.0731 | 0 | 0.01 | 32 | 68 | 30 | 41 | 53 | 25 | 43 |
| Synth 6 | $5.9 * 10^{-3}$ | 8607 | 0.0374 | 0 | 0.355 | 76 | 34 | 84 | 77 | 20 | 76 | 12 |
| Synth 7 | $6.0 * 10^{-6}$ | 2671 | 0.116 | 0 | 0.11 | 64 | 43 | 75 | 72 | 31 | 59 | 13 |
| Synth 8 | $2.7 * 10^{-6}$ | 327 | 0.0893 | 0 | 0.0116 | 11 | 62 | 30 | 42 | 48 | 23 | 40 |
| Synth 9 | $8.4 * 10^{-7}$ | 1693 | 0.104 | 0 | 0.0583 | 63 | 46 | 67 | 70 | 39 | 48 | 25 |
| Synth 10 | $5.9 * 10^{-7}$ | 1259 | 0.057 | 0 | 0.0511 | 48 | 45 | 65 | 68 | 40 | 40 | 23 |
| Synth 11 | $1.5 * 10^{-7}$ | 259 | 0.1 | 0 | 0.00995 | 3 | 64 | 27 | 0 | 56 | 26 | 44 |
| Synth 12 | $3.5 * 10^{-5}$ | 5493 | 0.0577 | 0 | 0.192 | 50 | 33 | 79 | 57 | 20 | 72 | 4 |
| Synth 13 | $7.8 * 10^{-4}$ | 8715 | 0.0539 | 0 | 0.357 | 61 | 30 | 84 | 63 | 31 | 80 | 25 |
| Synth 14 | $4.7 * 10^{-8}$ | 432 | 0.0466 | 0 | 0.0157 | 33 | 55 | 36 | 45 | 46 | 25 | 32 |
| Synth 15 | $1.8 * 10^{-7}$ | 515 | 0.0578 | 0 | 0.0194 | 39 | 61 | 46 | 56 | 52 | 29 | 33 |
| Synth 16 | $1.5 * 10^{-5}$ | 2768 | 0.0521 | 0 | 0.121 | 55 | 37 | 73 | 56 | 30 | 62 | 4 |
| Synth 17 | $9.6 * 10^{-3}$ | 9683 | 0.124 | 0 | 0.416 | 77 | 38 | 86 | 82 | 17 | 76 | -1 |
| Synth 18 | $4.1 * 10^{-8}$ | 613 | 0.0758 | 0 | 0.0187 | 35 | 56 | 48 | 51 | 45 | 30 | 33 |
| Synth 19 | $8.2 * 10^{-7}$ | 569 | 0.0574 | 0 | 0.02 | 54 | 60 | 48 | 59 | 52 | 29 | 33 |
| Synth 20 | $2.6 * 10^{-3}$ | 6691 | 0.0564 | 0 | 0.259 | 69 | 25 | 82 | 65 | 25 | 74 | 15 |

(b) Study 2

Table A1.: Sound features extracted from sound stimuli of both studies. For Librosa features, the mean values of all windows are reported.

| | Number of Strokes | Average Length [px] | Average Time [ms] | Average Speed [px/ms] | Inter-sections [1/100px] | Narrow Curves [1/100px] | Wide Curves [1/100px] | Obtuse Angles [1/100px] | Right Angles [1/100px] | Acute Angles [1/100px] |
|---|---|---|---|---|---|---|---|---|---|---|
| Crackles | 10.5 | 471 | 2796 | 0.22 | 1.78 | 0.55 | 0.60 | 0.94 | 0.22 | 0.67 |
| Telephonic | 6.0 | 1091 | 4613 | 0.27 | 1.05 | 0.16 | 0.37 | 0.55 | 0.08 | 0.26 |
| Strings | 3.9 | 900 | 4379 | 0.26 | 0.31 | 0.10 | 0.23 | 0.28 | 0.06 | 0.09 |
| String Grains | 10.1 | 914 | 3340 | 0.28 | 1.48 | 0.24 | 0.21 | 0.58 | 0.15 | 0.54 |
| Subbass | 4.5 | 1328 | 6154 | 0.23 | 0.92 | 0.17 | 0.43 | 0.51 | 0.08 | 0.30 |
| Noise | 12.8 | 1816 | 3540 | 0.57 | 2.37 | 0.32 | 0.22 | 0.61 | 0.16 | 0.56 |
| Piano | 3.8 | 586 | 3020 | 0.29 | 0.61 | 0.06 | 0.31 | 0.20 | 0.02 | 0.06 |
| Impact | 7.2 | 1239 | 3313 | 0.51 | 1.12 | 0.14 | 0.28 | 0.24 | 0.05 | 0.31 |
| Processed Guitar | 4.3 | 1124 | 4707 | 0.33 | 0.43 | 0.17 | 0.23 | 0.49 | 0.09 | 0.22 |
| Electric Guitar | 5.3 | 1121 | 4561 | 0.33 | 0.93 | 0.12 | 0.12 | 0.35 | 0.08 | 0.40 |

(a) Study 1

| | Number of Strokes | Average Length [px] | Average Time [ms] | Average Speed [px/ms] | Inter-sections [1/100px] | Narrow Curves [1/100px] | Wide Curves [1/100px] | Obtuse Angles [1/100px] | Right Angles [1/100px] | Acute Angles [1/100px] |
|---|---|---|---|---|---|---|---|---|---|---|
| Synth 1 | 3.4 | 1002 | 3021 | 0.44 | 1.8 | 0.3 | 0.84 | 1.1 | 0.13 | 0.31 |
| Synth 2 | 3.1 | 876 | 2923 | 0.4 | 2.6 | 0.55 | 0.96 | 1.1 | 0.24 | 0.53 |
| Synth 3 | 2.9 | 1186 | 3274 | 0.46 | 2.0 | 0.42 | 0.88 | 1.1 | 0.079 | 1.2 |
| Synth 4 | 3.9 | 2125 | 3178 | 0.74 | 5.4 | 0.32 | 0.39 | 0.86 | 0.23 | 1.4 |
| Synth 5 | 3.0 | 1636 | 3239 | 0.53 | 1.2 | 0.9 | 1.2 | 1.7 | 0.53 | 0.62 |
| Synth 6 | 4.3 | 1444 | 3344 | 0.49 | 3.7 | 0.75 | 0.67 | 1.4 | 0.19 | 1.6 |
| Synth 7 | 3.9 | 3644 | 2725 | 1.2 | 5.8 | 0.5 | 0.54 | 1.0 | 0.19 | 1.5 |
| Synth 8 | 2.03 | 936 | 3666 | 0.31 | 0.74 | 0.37 | 0.84 | 0.58 | 0.096 | 0.21 |
| Synth 9 | 5.2 | 1815 | 2752 | 0.79 | 5.7 | 0.42 | 0.98 | 0.95 | 0.19 | 0.7 |
| Synth 10 | 3.9 | 1555 | 3008 | 0.56 | 2.2 | 0.45 | 1.0 | 1.6 | 0.18 | 0.48 |
| Synth 11 | 2.1 | 834 | 3682 | 0.27 | 0.6 | 0.19 | 1.0 | 0.92 | 0.082 | 0.12 |
| Synth 12 | 2.6 | 1352 | 3533 | 0.46 | 2.9 | 0.34 | 0.61 | 0.68 | 0.12 | 0.63 |
| Synth 13 | 3.8 | 1239 | 3125 | 0.48 | 3.2 | 0.41 | 0.68 | 0.81 | 0.14 | 0.6 |
| Synth 14 | 2.5 | 995 | 3181 | 0.37 | 0.97 | 0.5 | 0.9 | 0.78 | 0.045 | 0.2 |
| Synth 15 | 2.8 | 1543 | 2984 | 0.55 | 3.3 | 0.62 | 0.74 | 1.5 | 0.31 | 1.4 |
| Synth 16 | 2.7 | 1104 | 2816 | 0.48 | 3.1 | 0.63 | 0.78 | 1.1 | 0.18 | 0.72 |
| Synth 17 | 2.3 | 3055 | 3171 | 1.1 | 6.4 | 0.34 | 0.24 | 0.58 | 0.15 | 1.2 |
| Synth 18 | 2.6 | 1269 | 3555 | 0.39 | 2.0 | 0.35 | 0.68 | 0.72 | 0.11 | 0.34 |
| Synth 19 | 3.3 | 1854 | 3279 | 0.54 | 3.8 | 0.83 | 0.89 | 1.6 | 0.31 | 1.4 |
| Synth 20 | 4.1 | 1993 | 2961 | 0.67 | 3.2 | 0.26 | 0.54 | 0.71 | 0.12 | 0.59 |

(b) Study 2

Table A2.: Sketch features for both studies. The mean value from all participants is presented for each sound stimulus.