Universiteit Leiden

Master Media Technology

From Perception to Action: Translating Audiovisual Crossmodal Correspondences into a Sketch-Based Music Synthesis Tool

Name: Marlinde van den Bosch
Student ID: s4088115
Date: 11/05/2025

Specialisation: Creative Intelligence and Technology

1st supervisor: Rob Saunders
2nd supervisor: Tessa Verhoef

Master's Thesis in Media Technology

Leiden Institute of Advanced Computer Science
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

LEIDEN UNIVERSITY

Leiden Institute of Advanced Computer Science (LIACS)

Creative Intelligence and Technology Master Thesis

From Perception to Action: Translating Audiovisual Crossmodal Correspondences into a Sketch-Based Music Synthesis Tool

An investigation into implementing shape-sound correspondences for intuitive music making

First supervisor:
Dr. R. Saunders

Second supervisor:
Dr. T. Verhoef

Student:
Marlinde van den Bosch
(s4088115)

May 4, 2026

Abstract

Crossmodal correspondences (CMCs) between visual shape and sound are well established, but research has primarily focused on passive recognition tasks. This thesis investigates whether such correspondences can be actively applied in creative interaction. A sketch-based music synthesis tool was developed that allows users to generate music by drawing, lowering the barrier to music-making for users without any musical expertise. The system maps the visual features including size, smoothness, circularity and vertical position to the auditory parameters including loudness, timbre, pitch, and tempo.

A perceptual matching questionnaire (N = 51) showed that participants recognized these mappings significantly above chance (73.7%). An in-depth experiment (N = 5) revealed that users successfully applied robust mappings such as size-loudness and vertical position-pitch, while weaker mappings (e.g., tempo-related) were more challenging. Qualitative feedback indicated that the tool felt intuitive and accessible, requiring no prior musical training.

The findings suggest that empirically grounded CMCs can support intuitive music-making through drawing, contributing both a functional prototype and insight into how people apply audiovisual CMCs beyond passive perception.

Keywords: crossmodal correspondences, sketch-based interaction, music synthesis, intuitive interaction, audiovisual mapping, casual creators

Contents

1. Introduction

Human perception is fundamentally multisensory, relying on the combination of information from different senses. Rather than processing sensory information in isolation, the brain continuously integrates signals from different modalities to form coherent experiences of the world [1, 2]. This integration plays a central role not only in everyday perception, but also in creative practices, such as music-making and visual art [3]. As digital creative tools increasingly support embodied interaction, understanding how people naturally connect information across senses becomes more important.

One well-established phenomenon in multisensory perception is the existence of crossmodal correspondences (CMCs). These are systematic non-idiosyncratic associations between features from different sensory modalities [4, 5, 6]. For example, associations between colour and sound [7, 8], or between specific odours and musical notes or geometric shapes [6]. CMCs are widely shared across individuals and cultures and are thought to arise from a combination of perceptual and statistical learning mechanisms [9, 7]. Unlike synesthesia, which involves idiosyncratic and involuntary cross-sensory experiences, CMCs reflect intuitive tendencies that influence perception and behaviour [4, 10].

Within the auditory-visual domain, a substantial body of research has demonstrated reliable links between sound and visual shape properties [7]. Classic phenomena such as the Bouba–Kiki effect demonstrate strong associations between speech sounds and shape curvature [11]. Later studies extended these findings to a wider range of auditory and visual features, e.g., [6, 12, 13]. Collectively, studies suggest that abstract visual shapes can meaningfully represent auditory features, and that these mappings are perceptually grounded rather than arbitrary [7].

Despite this large body of perceptual research on crossmodal correspondences, many studies have relied on passive matching or classification tasks to demonstrate these effects, e.g., [14, 12]. Participants are typically asked to select which visual stimulus best matches a sound, or vice versa. While some studies have explored more active forms of crossmodal communication and sound production [15, 16], far less attention has been paid to whether people can actively apply these correspondences, particularly in interactive contexts. As a result, there remains a gap between theoretical knowledge about crossmodal correspondences and their practical application.

1.1 Casual Creators

In the field of musical interfaces, many digital instruments and sound synthesis systems rely on abstract controls and parameter mappings that require significant learning. This can make them difficult to use for novice users and limit spontaneous exploration. The concept of Casual Creators describes interactive systems that allow people to explore a limited but meaningful creative space with little effort and immediate results [17]. By pri-

oritizing enjoyment and accessibility over technical mastery, such systems can reduce the learning curve typically associated with digital music tools and help beginners get started more easily. Sketch-based interfaces, in which users draw shapes that directly influence sound, could be a promising example of this approach. Although such interfaces remain relatively unexplored, basing them on empirically established crossmodal correspondences may help create tools that feel more intuitive, expressive, and immediately accessible.

1.2 Research Question

Discovering the gap between perceptual theory and interactive application raises the question of whether the established crossmodal correspondences can serve as usable design principles for a creative tool. If visual-auditory correspondences are consistent and widely shared, they could potentially form the basis for an intuitive interface for sound synthesis.

This thesis explores how established crossmodal correspondences between visual shape and sound can be translated into a sketch-based music synthesis tool, and to what extent this translation supports intuitive and effective interaction. The main research question underlying this thesis is:

> **To what extent can empirically established crossmodal correspondences between visual shape features and sound parameters be implemented in a sketch-based music synthesis tool that supports intuitiveness and meaningful interaction?**

This research question can be explored through two sub-questions:

1. *Do people recognize and agree with the implemented shape-sound mappings?*
2. *Can users actively and intentionally apply these correspondences when creating music through drawing, and how intuitive does this interaction feel in practice?*

To investigate these questions, a custom sketchpad synthesizer was developed in which the features of visual shape (size, smoothness, circularity, and spatial position) are mapped to the features of music (loudness, timbre, pitch, and tempo). These mappings were directly based on findings from crossmodal correspondence research.

The research uses a mixed-methods design. First, a larger-scale online questionnaire assesses whether participants consistently recognize the intended shape-sound correspondences implemented in the tool. Second, an in-depth exploratory study investigates whether users can actively apply these correspondences while interacting with the synthesizer, looking at both quantitative performance and qualitative experiences. By combining questionnaire data with quantitative performance and qualitative insights from the in-depth experiment, the study connects perceptual theory with interactive practice.

The thesis is structured as follows. Chapter 2 reviews relevant literature on crossmodal correspondences and audiovisual mappings. Chapter 3 describes the design and implementation of the sketch-based synthesis tool. Chapter4 presents the methodology used in the study. Chapter 5 reports the results, followed by a discussion of the findings in Chapter 6 and a conclusion in Chapter 7.

2. Literature Review

This chapter reviews the existing research on crossmodal correspondences and their potential application in interactive creative tools. It covers the origins of CMCs, reviews empirical evidence for specific audiovisual CMCs, identifies gaps in the literature, and introduces the Casual Creators framework that guides the design of the tool developed for this thesis.

2.1 Multisensory Perception and Crossmodality

Traditionally, perception has been seen as a set of independent sensory channels processing information in isolation [18]. However, more recent research has contradicted this view by describing perception as a model of continuous and automatic multisensory integration [19]. The brain does not process information from different senses separately, it actively combines signals from different modalities to make the perceptual experiences coherent [1, 20]. For instance, a hand clap involves hearing the sound it produces, seeing the hands being put together, and even having the proprioceptive feeling of the action. Three different senses receive signals, but everything is integrated into perceiving it as one event.

A well-established phenomenon within multisensory research is the existence of crossmodal correspondences (CMCs). These are defined as universal and natural associations where features from one sensory modality are systematically associated with features from another [4, 5], such as corresponding a certain colour with a certain emotion. These correspondences have been demonstrated across a wide range of senses [8], including associations between odours and auditory or visual features [6], or between tastes and sounds [21]. CMCs are characterized by their convergence, they are shared across different individuals and cultures, and are intuitive and nnot based on language [9, 7].

It is important to distinguish CMCs from synesthesia, a condition with which they are often confused. Synesthesia is an idiosyncratic and involuntary phenomenon in which stimulation of one sensory modality triggers a consistent, automatic experience in another, such as seeing colours when hearing music [22, 10]. These associations are specific to the individual and often arbitrary [4]. In contrast, CMCs are consistent and shared across the general population. While a synesthete might see a specific shade of blue when hearing a violin, a non-synesthete might experience a CMC of associating a high-pitched violin sound with a lighter colour or a smaller shape [9].

2.2 Origins of CMCs

The origins of these shared correspondences are still being researched, which has resulted in several proposed explanations. One of these involves statistical learning from environmental regularities [7, 9, 23]. Throughout life, humans are exposed to natural correlations between sensory features. For example, larger objects in the environment tend to pro-

duce louder, lower-frequency sounds when they move or fall, and smaller objects tend to produce quieter, higher-frequency sounds. The brain implicitly learns these correlations, which then shape our expectations and associations, and applies them later in broader contexts.

Another explanation involves structural isomorphism, which proposes that CMCs arise because different sensory modalities process information using similar neural architectures or representational formats [7, 24]. According to this view, the brain detects abstract structural similarities across domains, even without prior associative learning. Ravignani and Sonnweber (2017) tested this hypothesis with chimpanzees by presenting the animals with visual patterns (symmetric versus asymmetric) and sound patterns (repeated tones with symmetric versus asymmetric temporal structure) [25]. Without any training, the chimpanzees spontaneously looked longer at the visual pattern that matched the temporal structure of the sound they heard. This result mirrors findings for humans, and therefore suggests that the ability to recognize certain CMCs is not learned through language or culture, but reflects an evolutionarily ancient mechanism shared across species.

One example of structural isomorphism is A Theory Of Magnitude (ATOM) [26]. ATOM proposes that the brain uses a common generalized system for representing different types of magnitude, such as time, space, quantity, and intensity. From this perspective, features like the size of a visual object, the loudness of a sound, its pitch height, and the tempo of a rhythm are all processed as expressions of 'more' or 'less' along a shared neural metric [26, 27]. This explains why a correspondence such as loudness-size exists [28], as it arises from a common neural representation of magnitude. This theory is particularly relevant for explaining certain correspondences related to size, loudness, pitch, and tempo that, among others, form the basis of this thesis.

A third factor that may influence CMCs is language. Many audiovisual associations are encoded in everyday metaphors. Pitch can be labelled as 'high' and 'low', sounds as 'bright' and 'dark', or timbre as 'warm' and 'cold' [3, 7, 29]. However, these linguistic labels likely reflect rather than create the underlying perceptual associations, as it is found that similar correspondences appear in pre-linguistic infants [30], [31]. This finding is important for this current study, as it suggests that correspondences such as pitch-vertical position and size-loudness are grounded in perceptual mechanisms rather than linguistic convention, making them likely to be robust across languages and cultures.

2.3 Audiovisual CMCs

Now that the origins of CMCs have been discussed, the next step is to look at which audiovisual mappings have actually been established in literature. These mappings later form the basis of the sketch-based synthesis tool described in Chapter 3.

2.3.1 Feature Selection

Before reviewing the specific crossmodal correspondences implemented in the sketch-based synthesis tool, it is necessary to clarify why these particular visual and auditory features were selected. This requires considering what features define music and what features

define shapes.

Music can be described along several perceptual dimensions. The following four are particularly fundamental:

- **Pitch** organizes sound along a spectrum from low to high, determined by the frequency of sound wave vibrations [32, 33]. Pitches arranged sequentially form melodies, whereas stacked simultaneously, they form harmonies [34, 32, 35].
- **Timbre** allows listeners to distinguish between different sound sources, e.g., a piano from a trumpet, even when pitch and loudness are the same [36, 37]. Beyond source identification, timbre is often described as the texture or colour of music [38], and it carries emotional expression [39]. Listeners commonly describe timbre using adjectives, such as warm, harsh, sharp, or bright [38, 13].
- **Loudness** determines the intensity of sound, ranging from soft to loud [40]. Variations in loudness create dynamics in music, including crescendos, accents, and emotional contrast [39, 41].
- **Tempo** organizes time in music and determines how fast or slow a piece is played and giving it a sense of rhythm [42]. It contributes to the overall character of a piece, influencing its energy and emotional expression [39, 41].

Perceptual research has identified geometric properties that define a shape's character [43, 7]. The following four are recognized as key dimensions along which a shape can vary:

- **Smoothness** describes a shape's contour. Shapes with gradually curving lines are considered smooth, whereas shapes with sharp corners are considered angular.
- **Circularity** captures how close a shape is to a perfect circle. Although it is sometimes confused with smoothness, the two are independent. Figure 2.1 illustrates how smoothness and circularity can vary independently. The left pair contrasts circularity while keeping smoothness similar, the right pair contrasts smoothness while keeping circularity similar.
- **Size** is determined by the area the shape covers on the canvas. It ranges from small to large and determines visual weight and scale.
- **Position** describes the location of the shape within a spatial frame. A shape has a position on the horizontal (x) and vertical (y) axes, and together determine its placement.

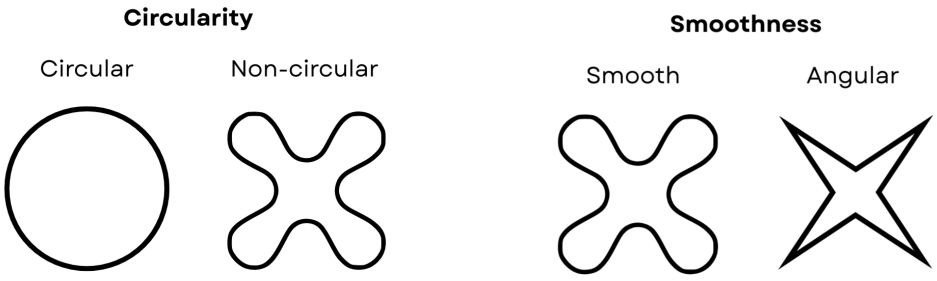


Figure 2.1: Circularity and smoothness as independent dimensions

This thesis builds on existing research showing that these two sets of dimensions can be connected through crossmodal correspondences. The following sections review the empirical evidence for correspondences between these specific features.

2.3.2 Timbre

Of all the auditory parameters, timbre has perhaps the most intuitive visual counterparts. A sound can be warm or harsh, smooth or rough, and these qualities turn out to map onto specific shape properties. The next two subsections look at how timbre relates to smoothness and circularity.

2.3.2.1 Timbre - Smoothness

The most famous example of an audiovisual correspondence is the Bouba-Kiki effect. In this classic example, first observed by Usnadze [44], participants are shown a rounded blob-like shape and a sharp spiky shape and asked which one is called "Bouba" and which is called "Kiki". A strong and consistent majority across cultures associate the rounded shape with the rounded-sounding name "Bouba" and the angular shape with the sharp-sounding name "Kiki" [11, 45]. While this effect originally used speech sounds, later research has extended these findings to musical timbre. Adeli et al. (2014) demonstrated that listeners systematically associate musical timbres with visual shapes [14]. Soft, warm timbres, e.g. piano or marimba, were generally matched with rounded, smooth shapes, whereas harsh, noisy timbres (e.g., cymbal, gong) were matched with jagged, angular shapes. Instruments with mixed timbral qualities (e.g., saxophone), tended to be associated with an intermediate shape. Importantly, the study showed that these shape associations were largely independent of colour, grayscale, musical expertise, age, or self-reported synaesthetic experiences.

2.3.2.2 Timbre - Circularity

In much of the crossmodal correspondence literature concerning timbre, circularity as a distinct visual dimension has received little attention compared to smoothness. The classic Bouba-Kiki effect, for instance, demonstrates a robust mapping between timbre and smoothness [46]. However, as the rounded Bouba shape and the spiky Kiki shape are generally both non-circular, the correspondence being tested is exclusively one of smoothness, not circularity. As discussed earlier in Section 2.3.1, smoothness and circularity are geometrically independent dimensions (see Figure 2.1).

Recognizing this distinction is important for the current work, because circularity is implemented in the sketch-based tool as a simple measure of visual complexity. In the tool designed in this thesis, circularity serves as a computational representation for complexity. Circularity captures how closely a shape approximates a perfect circle, with a perfect circle scoring highest and irregular, star-like, or elliptical shapes scoring lower. While complexity is a multidimensional perceptual feature that is difficult to quantify computationally, circularity gives a mathematical approximation. Shapes with high circularity are perceived as simple, whereas shapes with low circularity are perceived as complex.

The mapping from this complexity dimension to timbre is supported by the con-

cept of affective mediation, which proposes that stimuli from different sensory modalities become associated when they evoke similar emotional responses [8]. Mesz et al. (2023) demonstrated that simple visual shapes are reliably associated with low-arousal, calm musical emotions, whereas complex, irregular shapes are associated with high-arousal, energetic musical emotions [3]. Although Mesz et al. did not directly test timbre, their findings on complexity and arousal provide indirect support for the mapping implemented in the sketchpad-based synthesizing tool. Given that warm timbres are typically perceived as calmer and less arousing, while harsh timbres are perceived as more energetic and arousing [39], shapes with high circularity (low complexity) can therefore be mapped to warm, soft timbres, and shapes with low circularity (high complexity) can be mapped to harsh, bright timbres.

2.3.3 Pitch

Pitch is not only about shape. Research has also found consistent links between pitch and where a shape is located. The following sections will review this correspondence, alongside the shape features size and smoothness.

2.3.3.1 Pitch - Smoothness

In addition to timbre, research has also established a link between pitch height and perceived smoothness. Marks [47] and Parise and Spence [48] showed that high-pitched sounds are more quickly and accurately associated with angular shapes, while low-pitched sounds are associated with rounded shapes. One of Marks' experiments was based on research preceding the Bouba-Kiki effect, where similar results were found when matching an angular and a smooth shape with Maluma (later Bouba) and Takete (later Kiki) [49]. Marks discusses that the consonants /t/ and /k/ in Takete are high-pitched consonants and that the consonants and vowel /m/, /l/, and /a/ in Maluma are low-pitched. This suggests that the Bouba-Kiki effect reflects a broader principle, justifying mapping both timbre and pitch to the visual feature of smoothness.

2.3.3.2 Pitch - Spatial Position

Another robust finding in audiovisual correspondence research is the systematic mapping between pitch and vertical space [5, 50, 12]. In 1930, Pratt [29] observed that listeners conceptualize pitch along a vertical axis, describing high frequencies as "high" and low frequencies as "low". This linguistic metaphor seems to reflect a deeper perceptual mechanism. Roffler and Butler [50] found that listeners localize sounds more accurately when pitch varies, suggesting that pitch inherently carries spatial information. When a sound moves upward in pitch, listeners tend to perceive it as coming from a higher location, even when it is not. This suggests that the auditory system automatically associates frequency with spatial height.

Chiou and Rich (2012) provided additional evidence that this mapping is automatic. In a speeded classification task, participants responded faster when a high-pitched sound accompanied a visual target in a high spatial location, and when a low-pitched sound accompanied a target in a low location [12]. This occurred even though the sounds carried no

spatial information and were not predictive of the target's position. Incongruent pairings (e.g., high pitch with low vertical position) resulted in slower reaction times and reduced accuracy. The strength and direction of this effect depended on relative rather than absolute pitch. When the same tone served as the "high" item in one task and the "low" item in another, it biased attention in opposite spatial directions. This demonstrates that the pitch-vertical position correspondence is driven by higher-level cognitive processes (such as categorical comparison and expectation), rather than fixed sensory mappings [12].

While some research has also explored mappings between pitch and horizontal position (e.g., high pitch = right), these associations appear to depend on musical training and cultural factors such as reading direction [51, 52]. Vertical pitch mappings are therefore the more suitable choice for implementation in the tool, as they remain robust regardless of musical expertise and culture.

2.3.3.3 Pitch - Size

Pitch has also been systematically associated with the size of visual objects. Higher-pitched sounds are consistently linked to smaller visual objects, while lower-pitched sounds are linked to larger ones [53, 54, 5]. Gallace and Spence [53] demonstrated this relationship using a speeded visual size-discrimination task. Participants were faster and more accurate when a high-frequency tone accompanied a small visual disk, or when a low-frequency tone accompanied a large disk, compared to incongruent pairings. This effect occurred even when participants were instructed to ignore the sounds, suggesting the mapping operates automatically.

Evidence that this correspondence emerges early in development comes from Mondloch and Maurer [54], who showed that children as young as 30-36 months reliably matched higher-pitched sounds to smaller objects and lower-pitched sounds to larger ones. This effect occurred even when loudness was varied to prevent intensity-based matching strategies, which confirms that the association is specific to pitch rather than loudness. The researchers argue that pitch-size associations likely originate in early perceptual organization and are due to what remained from the crossmodal neural connections that were present at birth [54].

2.3.4 Loudness

Loudness has only one main visual counterpart and is not as versatile as pitch or timbre. However, this loudness-size correspondence is very well-established, as the following subsection will discuss.

2.3.4.1 Loudness - Size

Within the framework of ATOM [26], one of the most well-established correspondences is mapping between loudness and size. Across multiple directions/branches of studies, louder sounds are consistently associated with larger visual objects, while softer sounds are matched to smaller ones [7, 55]. Marks [47] demonstrated this using speeded classification, showing that participants respond faster when sound intensity and visual size are congruent (loud-large, soft-small) than when they are incongruent. Eitan and Timmers [56]

extended this finding to a musical context, and showed that listeners systematically associate louder passages of music with larger visual forms and broader spatial gestures, even in the absence of real-world sound sources.

Importantly, Smith and Sera [55] showed that even young children reliably associate louder sounds with larger objects. This indicates that the mapping does not depend on formal learning or linguistic metaphor alone, providing further support for Walsh's Theory of Magnitude (ATOM) as the underlying mechanism for this correspondence [26].

2.3.5 Tempo

The vast majority of audiovisual crossmodal correspondence research has used isolated sounds or single tones rather than musical samples. While this approach has led to robust findings for features such as pitch, loudness, and timbre, it has overlooked features that are inherently musical. Tempo is an example of this, as it is not a property of a single sound, but emerges from the temporal structure of music over time. The relative neglect of tempo in literature therefore reflects a broader gap because of the focus on isolated sounds instead of music, rather than indicating that tempo-related correspondences are weak or nonexistent.

Since tempo is a fundamental musical feature, as stated in section 2.3.1, it is important to include it despite the limited research. Of the four shape features, there is no evidence in the literature for linking smoothness and vertical position to tempo. On top of that, there is no clear theoretical reason to expect. Size and circularity, on the other hand, offer plausible connections. The tempo mapping is therefore based on these two features.

2.3.5.1 Tempo - Size

Tempo can be understood not only as the overall speed of a piece, but also in terms of its underlying features: faster tempi typically involve shorter note durations and higher note density (more notes per unit of time), while slower tempi involve longer note durations and lower note density. This broadens the focus from tempo to include duration and density, which have been researched more extensively.

Sound symbolism research has shown that longer vowel durations are associated with perceptions of bigger objects, while shorter durations are associated with smaller objects [57, 58]. This aligns with the ecological intuition that larger objects tend to move more slowly and produce longer-lasting sounds, while smaller objects move quickly and produce shorter sounds. Related to this, Schorn et al. [59] introduced the concept of music event rate (the number of notes per measure) and found that a higher event rate (more notes per measure) led participants to perceive objects as smaller, while a lower event rate led to perceptions of larger objects. Since shorter note durations often result in higher note density, this provides additional support. Faster rhythms (higher density) are associated with smaller size, and slower rhythms (lower density) with larger size. Together, these findings suggest a plausible mapping between size and tempo, based on note duration and density. Larger objects are associated with slower tempi (longer notes, lower density), while smaller objects are associated with faster tempi (shorter notes, higher density).

2.3.5.2 Tempo - Circularity

Unlike the other correspondences reviewed earlier, the link between circularity and tempo is not directly researched. However, the implementation of circularity as a complexity measure, as introduced in section 2.3.2.2, offers a plausible foundation for the mapping.

As established, simple shapes (high circularity) are associated with calmness and low arousal, while complex shapes (low circularity) are assicated with energy and high arousal [3, 60]. Tempo aligns with these same affective qualities. Slower tempo conveys calmness and relaxation, while faster tempo convey energy and excitement [39, 41]. Through affective mediation, simple shapes thus align with slow tempo, and complex shapes with fast tempo [8].

Beyond this affect-based link, there is more direct evidence that circularity is perceptually present in temporal dynamics. Thoret et al. (2016) synthesized friction sounds that conveyed the velocity patterns of circular versus elliptical movements and found that listeners could reliably distinguish between these trajectories by sound alone [61]. When asked to draw circular motion while hearing sounds that implied elliptical movement, participants unconsciously produced elliptical shapes. This suggests an automatic crossmodal coupling between circular form and temporal structure, independent of affective mediation.

In summary, the circularity-tempo mapping is supported by a complexity-affect link and by evidence that circularity is perceivable through sound via velocity patterns. This has led to mapping circular shapes to a slow tempo, and non-circular shapes with a fast tempo. While direct empirical evidence for a circularity-tempo correspondence remains limited, this indirect support provides a defensible foundation. The mapping is treated as an exploratory element in this thesis, contributing to the evaluation of whether users can intuitively apply such a mapping in an interactive creative tool.

2.3.6 Robustness of CMCs

An important observation coming from this review is that not all crossmodal correspondences are equally strong or well researched. Some mappings, such as loudness-size or timbre-smoothness, are highly robust and show automatic effects on perception. Others, particularly those involving tempo or circularity, have weaker foundations and have received less attention in the literature.

This difference in robustness has implications for both design and evaluation. Stronger correspondences can be expected to support more intuitive interaction with minimal learning, while weaker mappings may require greater effort or more careful implementation. Recognizing this variability suggests that design choices should be weighted towards the most robust correspondences when prioritizing intuitiveness. The relative robustness of each correspondence will therefore determine how much influence each shape feature has on its corresponding sound parameters, as will be discussed in the next chapter.

2.3.7 From Recognition to Application

The previous subsections have shown that crossmodal correspondences are robust and well-documented. However, nearly all of this research has relied on passive perceptual tasks, such as matching, classification, or rating, where participants select a response rather than generate an outcome. While this approach has established that CMCs exist, it does not tell whether users can actively apply these correspondences in interactive contexts. This thesis therefore shifts focus from passive recognition to active application, investigating whether users can intentionally apply crossmodal correspondences when creating music through drawing.

2.4 Casual Creators

Crossmodal correspondences provide the perceptual foundation for the tool, the concept of Casual Creators provides the interaction design framework. This section introduces the challenges associated with traditional digital music tools, outlines the principles of Casual Creators, and argues that the established crossmodal correspondences can serve as an effective design strategy for lowering the barrier for creative expression.

Many digital music tools, such as software synthesizers (e.g., Serum, Vital) and music production applications (e.g., Ableton Live, FL Studio), are challenging for novice users to use. These systems often rely on abstract parameters, such as oscillators, filters, and envelopes, that require substantial technical musical knowledge to understand and control. Achieving a desired sound is not intuitive for beginners. The steep learning curve can be discouraging and limit spontaneous exploration and creative play. While expert users may appreciate the fine control, this need for prior expertise excludes a large population who might wish to engage with music-making in a more casual and exploratory manner [17].

The concept of Casual Creators, introduced by Compton, describes a genre of interactive systems designed specifically to lower barriers to creative expression [17]. Casual creators are characterised by several key principles:

- **Low barrier to entry:** Users can begin interacting with minimal instruction or prior knowledge.
- **Immediate feedback:** The consequences of user actions are immediately perceptible, supporting learning through exploration.
- **Exploration over precision:** The system encourages playful experimentation rather than demanding precise control.
- **Autotelic experience:** Interaction is rewarding, the process is enjoyable regardless of the output's quality or complexity.
- **Limited but meaningful creative space:** The system constrains possibilities enough to prevent overwhelm, while allowing genuine creative choices.

Casual Creators systems prioritise enjoyment and accessibility over technical mastery, making them particularly well-suited for users who may not identify as musicians or artists but who wish to engage in creative play.

2.4.1 Sketchbased-Interaction

The principles of Casual Creators can be applied to music production tools. Rather than requiring technical expertise, such systems could prioritize enjoyment, accessibility, and spontaneous exploration. One way to achieve this is through sketch-based interaction.

Drawing is a familiar human activity. Sketch-based interaction relies on this familiarity, allowing users to create and manipulate digital content through freehand drawing. This approach has been applied in various domains, from 3D modelling to photo editing, but has not been used much in music creation [62]. For the present work, sketch-based interaction offers several advantages. Drawing requires no specialized musical knowledge. Users can focus on the visual form they wish to create, with sound emerging as the consequence of their drawing. Drawing is expressive, allowing for variations in shapes and their positioning that can be directly translated into audio variation. Furthermore, drawing supports direct manipulation [63]. Users interact with the visual representation of sound parameters, rather than with abstract sliders or numerical inputs. This aligns closely with the principles of Casual Creators.

In summary, this thesis argues that crossmodal correspondences can serve as an effective design strategy for Casual Creators. When the mapping between user action (drawing shapes) and system response (the resulting music) aligns with natural perceptual expectations, the interaction feels intuitive rather than learned. Users do not need to memorize the mappings or follow a manual when first using the tool, instead, they can rely on their intuition of how a particular shape should sound. This supports control and enables users to focus on creative exploration rather than learning how to use the tool. By grounding the sketch-sound mappings in empirically established CMCs, the tool aims to achieve "natural mapping", a concept that refers to controls that feel immediate and intuitive through physical analogies, spatial correspondences, or cultural standards [64]. When successful, such mappings make the system's behaviour seem obvious and effortless, supporting fluent interaction even for first-time users.

2.5 Summary, Gaps, and Contributions

Several key points emerged from this literature review. Crossmodal correspondences between visual shape and auditory features are robust, shared across individuals, and grounded in perceptual mechanisms. Table 2.1 provides an overview of the correspondences discussed in this chapter.

However, despite this rich body of evidence, a few gaps have been identified. Tempo has received far less attention than other sound parameters, largely because research has focused on isolated sounds rather than music. The specific role of circularity, distinct from smoothness, is also underexplored. Most fundamentally, nearly all CMC research has relied on passive perceptual tasks, leaving it unclear whether users can actively apply these correspondences in interactive contexts. At the same time, the framework of Casual Creators offers a design principle for lowering barriers to creative expression. Sketch-based interaction aligns well with this principle, and grounding it in empirical CMCs offers a path toward intuitive, exploratory music-making.

Visual Features				Auditory Parameters			
Circularity	**Position**	**Size**	**Smoothness**	**Loudness**	**Pitch**	**Timbre**	**Tempo**
circular						warm	slow
irregular						harsh	fast
	high				high		
	low				low		
		big		loud	low		slow
		small		soft	high		fast
			smooth		low	warm	
			angular		high	harsh	

Table 2.1: Mapping Crossmodal Correspondences from Visual Features to Auditory Parameters

The current thesis addresses these foundations and gaps in three ways. First, it translates established CMCs into a sketch-based music synthesis tool, weighting each mapping according to the relative robustness of its empirical support. Second, it shifts focus from passive recognition to active application, evaluating not only whether users recognize the mappings, but also whether they can intentionally apply them to achieve desired sounds. Third, it offers an exploratory assessment of less established mappings. Together, these elements inform the design of the sketch-based synthesis tool described in the next chapter. The implementation translates the empirical evidence reviewed here into concrete design decisions, which are then evaluated through the user study discussed in subsequent chapters.

3. Implementation

This chapter describes how the crossmodal correspondences reviewed in Chapter 2 were translated into an interactive, sketch-based sound synthesis tool. The tool allows users to draw shapes on a digital canvas, which are then analyzed and translated into music in real time. First, the design goals are outlined, followed by the rationale for key design decisions, a detailed explanation of the mapping between visual features and sound parameters, the interaction design, and finally the technical implementation.

3.1 Design Goals

The primary goal was to create an interactive tool that allows users to explore music through drawing in a way that feels intuitive, expressive, and accessible. Rather than focusing on precise musical control, the tool is designed to encourage playful exploration, grounded in the audiovisual correspondences described in Chapter 2.

Four main design goals were derived from this primary goal:

- Support intuitive interaction: The system should be based on established audiovisual mappings so that users can understand how their drawings influence sound without instruction.
- Enable sound control through drawing: Shapes and their spatial position should directly influence auditory output in real time, allowing users to discover the mappings as they draw.
- Minimize the need for musical expertise: Users should not require prior knowledge of music theory, notation, or sound synthesis to interact with the tool meaningfully.
- Align with Casual Creator principles: The tool should encourage playful exploration, low commitment, and immediate feedback, rather than requiring technical skill.

3.2 Design Rationale

Several design decisions were made to allow for maximum user freedom and intuitive interaction.

- Users create their own shapes from scratch rather than selecting from a library of predefined forms. This allows for more abstract and creative expression, giving users more control over the resulting sound.
- Drawing was chosen over widgets like buttons or sliders to encourage exploration and discovery. This aligns with the principles of Casual Creators, where the process of interaction is part of the creative experience.
- The translation from visual features to auditory parameters is grounded in the crossmodal correspondences described in Chapter 2. The relative robustness of each cor-

respondence (e.g., strong for smoothness-timbre, exploratory for circularity-timbre) directly determines how much influence a given shape feature has on its associated sound parameter. This ensures the system is both scientifically informed and intuitively predictable.

- The interface is kept intentionally minimal, with visual feedback limited to the drawing canvas itself. This keeps the user's attention focused on what they are doing without distracting them with unnecessary information.

3.3 Technical Implementation

The system consists of two main components: a visual front-end for drawing and shape analysis, and an audio back-end for sound synthesis. They communicate via the Open Sound Control (OSC) protocol, which enables low-latency communication between the two components. Figure 3.1 shows the data flow between the two components, from drawing a shape to playing sound.

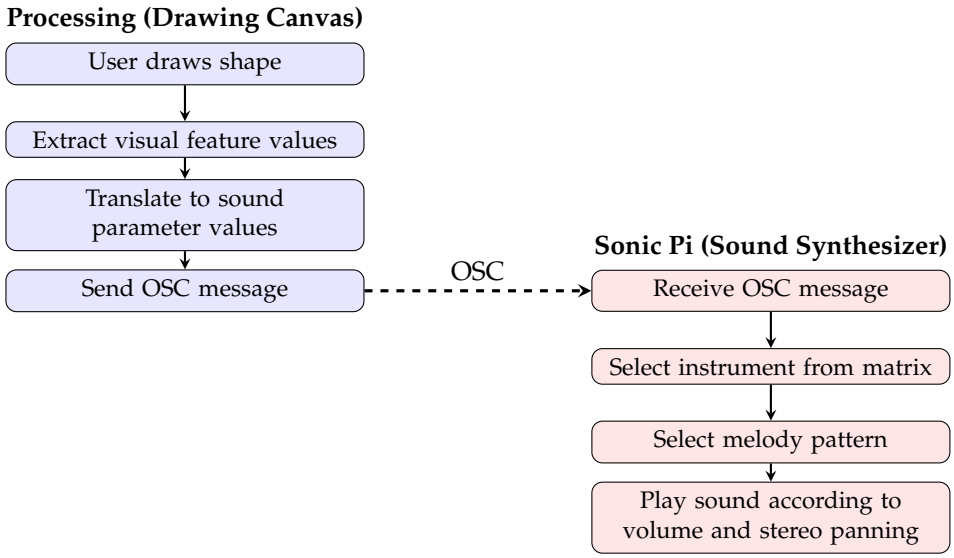


Figure 3.1: Data flow between Processing and Sonic Pi

3.3.1 Visual Interface

The drawing canvas and shape analysis are implemented in Processing, an IDE suited for visual applications. The interface consists of a full-screen drawing canvas with a minimal control panel in the top-right corner. This panel indicates whether the system is in drawing mode or erasing mode and allows users to switch between the two. In drawing mode, users can freely draw shapes on the canvas. Figure 3.2 shows the interface with several example shapes drawn on the canvas.

In erasing mode, clicking on an existing shape removes it, along with its corresponding sound. Shapes are always automatically closed by the system. When a user draws a line, the system records the sequence of points. Upon release, the final point is automatically connected back to the starting point, ensuring that every drawn mark forms a closed

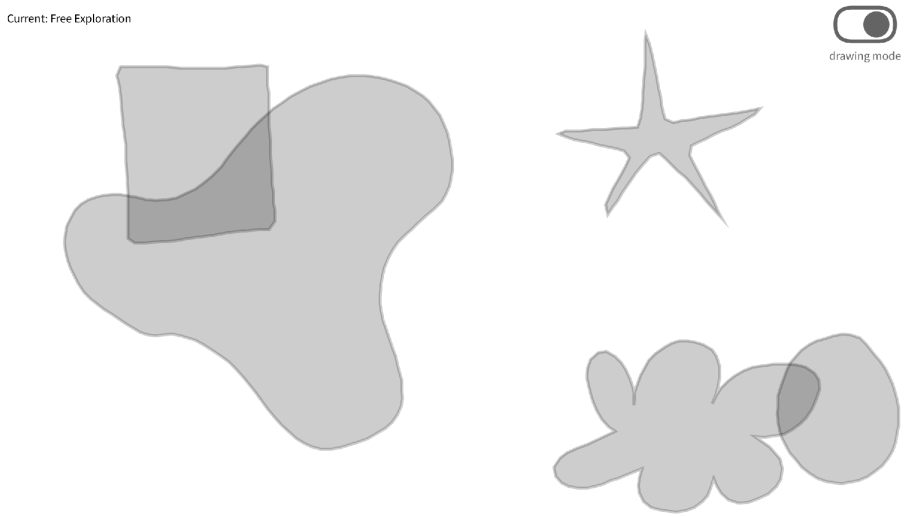


Figure 3.2: The sketch-based interface

polygon. This guarantees that each shape has a well-defined area and perimeter, which are essential for calculating the geometric properties.

When the maximum of ten shapes simultaneously on the screen is reached, a red warning banner appears at the top of the screen, and the system automatically switches to erasing mode until the user removes at least one shape (Figure 3.3). This limit was implemented because, on the computer used for designing the tool and conducting the experiment, the sound synthesizer (Sonic Pi) struggled to process a large number of simultaneous sounds. The maximum ensured stable audio performance while still allowing for sufficiently rich compositions.

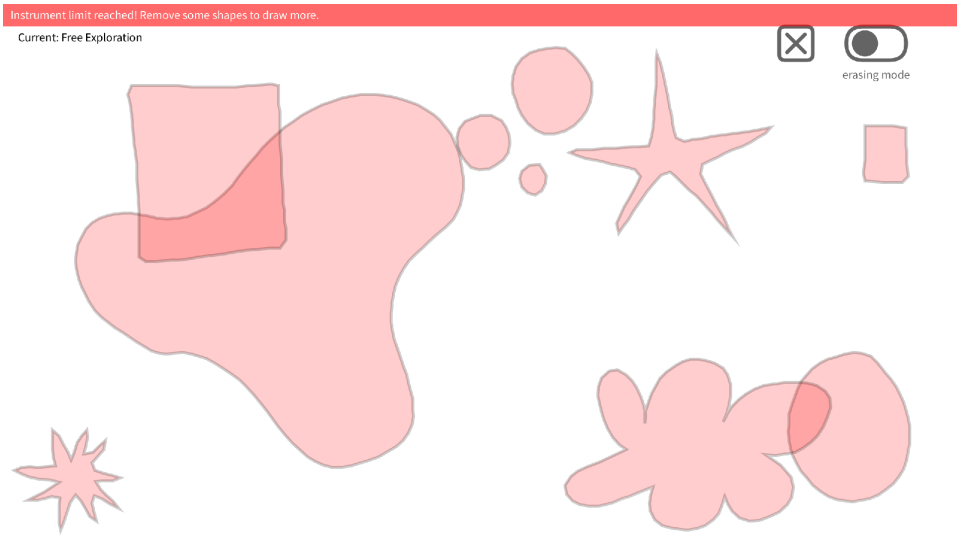


Figure 3.3: Limit reached: red warning banner in erasing mode

When a user completes a shape by releasing the mouse, the system analyzes the shape to extract the four geometric properties. These calculations are performed using the following methods:

- Area is calculated using the shoelace formula, which computes the area of a polygon from the coordinates of its vertices. The raw area is then normalized relative to the canvas size.
- Circularity is calculated using the formula given in Equation 3.1 , which yields a value of 1 for a perfect circle and lower values for shapes that deviate from circularity.

$$circularity = \frac{4\pi \times area}{perimeter^2} \tag{3.1}$$

- Smoothness is calculated by measuring the angular acceleration along the shape's perimeter. The algorithm computes the turning angle at each vertex. It then calculates the angular acceleration as the absolute second difference between consecutive turning angles, reflecting how abruptly the direction changes. These accelerations are summed across all vertices and normalized by the theoretical maximum for a shape with maximally alternating sharp angles ($n \times \pi$, where n is the number of vertices). The final smoothness value ranges between 0, reflecting a maximally angular shape, and 1, reflecting a perfectly smooth shape. The calculation is summarized in Equation 3.2:

$$smoothness = 1 - \frac{\sum_{i=1}^{n} |\Delta^2 \alpha_i|}{n \times \pi} \tag{3.2}$$

  where α_i is the turning angle at vertex i and $\Delta^2 \alpha_i = \alpha_{i+1} - 2\alpha_i + \alpha_{i-1}$ is the second difference (angular acceleration).
- Position is the y-coordinate of the shape's centroid, normalized to the canvas height, with 0 representing the top and 1 representing the bottom.

Once the visual features are extracted, they are translated into four auditory parameters using weighted formulas that reflect the relative robustness of each crossmodal correspondence, as discussed in Section 2.3.6. Figure 3.4 visualizes these weighted mappings.

- Loudness is determined primarily by the area of the shape. Shapes with an area of 2000 pixels or less (<0.26% of the canvas, e.g., a square of 44x44 pixels) receive a loudness value of 0, while shapes exceeding 15% of the canvas (116,640 pixels) receive a value of 100. Areas between these thresholds are normalized linearly.

$$loudness = 100 \times \frac{area - 2000}{116,640 - 2000} \tag{3.3}$$

- Pitch receives contributions from vertical position (weight 0.6), size (weight 0.4), and smoothness (weight 0.1). The weighted sum is first constrained to the range $[-1, 1]$

using the `constrain()` function. This value is then mapped using Equation 3.4, where 0 represents the highest pitch and 4 the lowest.

$$pitch = 4 - \left\lfloor \frac{(-0.4 \times size - 0.1 \times smoothness - 0.6 \times position + 1)}{0.4} \right\rfloor \tag{3.4}$$

- Timbre receives contributions from smoothness (weight 0.8) and circularity (weight 0.2). The weighted sum is mapped to a discrete timbre level using Equation 3.5, where 0 represents the harshest timbre and 4 the warmest.

$$timbre = \begin{cases} 0 & \text{if } s < -0.6 \\ 1 & \text{if } -0.6 \leq s < -0.2 \\ 2 & \text{if } -0.2 \leq s < 0.2 \\ 3 & \text{if } 0.2 \leq s < 0.6 \\ 4 & \text{if } s \geq 0.6 \end{cases} \quad \text{where} \quad s = 0.8 \times smoothness + 0.2 \times circularity \tag{3.5}$$

- Tempo receives contributions from size (weight 0.7) and circularity (weight 0.3). The weighted sum is mapped to one of five tempo levels using Equation 3.6, corresponding to note durations: very slow (whole notes), slow (half notes), medium (quarter notes), fast (eighth notes), and very fast (sixteenth notes).

$$tempo = \begin{cases} 0.25 & \text{if } t < -0.6 \\ 0.5 & \text{if } -0.6 \leq t < -0.2 \\ 1 & \text{if } -0.2 \leq t < 0.2 \\ 2 & \text{if } 0.2 \leq t < 0.6 \\ 4 & \text{if } t \geq 0.6 \end{cases} \quad \text{where} \quad t = 0.7 \times size + 0.3 \times circularity \tag{3.6}$$

- Horizontal spatial position also influences stereo panning. The horizontal position (x-coordinate) of the shape determines the left-right volume balance. Shapes drawn on the left side of the canvas sound louder in the left channel, while shapes drawn on the right side sound louder in the right channel. The pan value ranges from -1 (full left) to 1 (full right), calculated as shown in Equation 3.7.

$$pan = \frac{2x - width}{width} \tag{3.7}$$

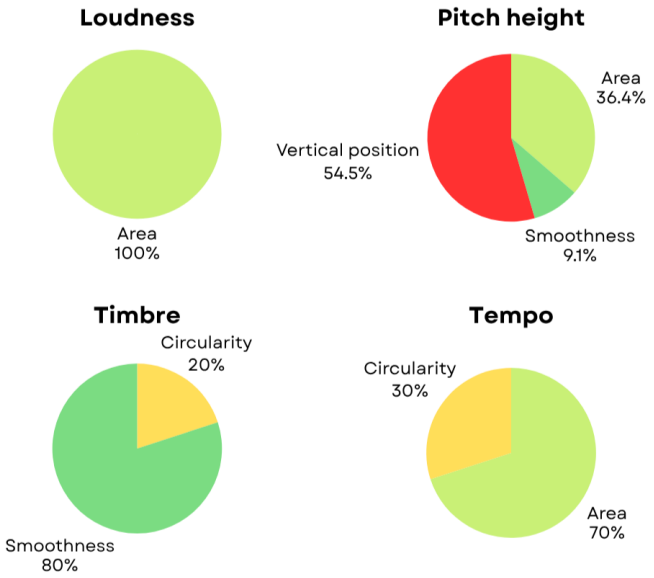


Figure 3.4: Composition of the auditory parameters

The complete source code of the tool is provided in Appendix A.

3.4 Sound Synthesis

The audio back-end is implemented in Sonic Pi[1], a live-coding environment for audio. Sonic Pi runs continuously in the background, listening for OSC messages from the Processing sketch. When it receives an instrument message, it extracts the parameters and synthesizes the corresponding sound.

Instrument selection is handled by a 5×5 matrix that maps combinations of pitch level (0–4) and timbre level (0–4) to specific instrument samples. For example, a high pitch (level 0) combined with a warm timbre (level 4) might select a flute, while a low pitch (level 4) combined with a harsh timbre (level 0) might select a timpani. This matrix allows the system to generate a wide variety of timbres while maintaining a consistent mapping between the visual features and the resulting sound. Table 3.1 shows the complete instrument matrix.

	Timbre Level				
Pitch Level	**0 (Harshest)**	**1**	**2**	**3**	**4 (Warmest)**
0 (Highest)	Chime	Xylophone	Violin	Piccolo	Flute
1	Cornet	Banjo	Ukulele	Alto Sax	Kazoo
2	Trumpet	Harpsichord	Acoustic Guitar	Piano	Harp
3	Trombone	Electric Guitar	Bassoon	Organ	Classical Guitar
4 (Lowest)	Timpani	Acoustic Bass	Cello	Bass Guitar	Tuba

Table 3.1: Instrument selection matrix mapping pitch level and timbre level to specific instrument samples

[1]Sonic Pi is available from `https://sonic-pi.net`

For each instrument in the matrix, there are five samples, one for each tempo level (sixteenth, eighth, quarter, half, and whole note). All samples are in the key of C, the base pitch. When Sonic Pi plays a sample, it uses pitch shifting to transpose the sample to the intended pitch. The tempo of the sample playback is controlled by the sleep time between notes, which is derived from the tempo parameter received from Processing.

To keep the musical output coherent when multiple instruments play simultaneously, all melody patterns are written in the same key. For each tempo level, a set of precomposed melody patterns is defined as sequences of semitone offsets from the root note. When an instrument is created, it randomly selects one of these patterns to use for its entire lifetime. The pattern determines the sequence of pitches played, while the tempo determines the duration of each note. This ensures that different melodies sound coherent when played together.

Stereo panning is implemented by calculating a balance value from the left and right volume parameters received from Processing. A shape positioned on the left side of the canvas results in a negative balance (pan left), while a shape on the right results in a positive balance (pan right). The overall volume is determined by the maximum of the two channel volumes.

Sonic Pi runs each instrument in a separate thread, allowing multiple shapes to sound simultaneously. When the Processing sketch sends a clear message, Sonic Pi stops all active threads and clears its instrument list. When a specific instrument needs to be stopped, because its corresponding shape was erased, Sonic Pi receives a stop message with the instrument ID and terminates only that thread.

3.5 Data Logging for Experiment

To enable analysis of user behavior during the experiment, further explained in Chapter 4, the system includes a data logger. For each drawing session, a DataLogger object is initialized with a unique participant ID and session timestamp. The logger creates a CSV file for the session to log shape data and general events. The CSV file logs the following information for every shape added or removed:

- Timestamp (milliseconds since session start)
- Event type (ADD or REMOVE)
- Shape ID and participant ID
- Current task ID (0 for free exploration, 1–8 for directed tasks)
- Active shape IDs and total number of shapes on the canvas
- All shape metrics (area, circularity, smoothness, centroid x and y)
- All sound parameters (pitch, timbre, tempo, left volume, right volume)
- The complete set of vertices defining the shape

This logging serves two purposes. First, it allows reconstruction of the exact state of the canvas at any moment during the experiment. Second, it enables quantitative analysis of the participants' drawing during the tasks, including how many attempts they made

before achieving a satisfactory result and how shape metrics differed between conditions (e.g., loud vs. soft, fast vs. slow). The logged data formed the basis for the quantitative results presented in Chapter 5.

4. Method

This chapter describes the methodological approach used to investigate the proposed mappings between visual shape features and sound parameters. A mixed-methods approach was chosen, combining a large-scale perceptual matching questionnaire with an in-depth interactive experiment and interview study. This design allowed to examine both how the implemented mappings were perceived and how they functioned during active interaction with the sketch-based sound synthesis tool.

4.1 Study Design Overview

The study consisted of two complementary components. The first component was an online matching questionnaire that tested whether the visual–auditory mappings implemented in the tool were consistently perceived by participants and aligned with the established crossmodal correspondences. Because participants only evaluated stimuli rather than interacting with the tool, this component primarily assessed the perceptual validity of the implemented mappings.

The second component was an in-depth interactive experiment with a smaller group of participants using the sketch-based sound synthesis tool. This experiment examined whether participants could intentionally control sound through drawing, whether the mappings felt intuitive during interaction, and how users experienced the tool in terms of usability, expectation, and creativity.

Together, these components assess both whether the mappings were perceived as intended and whether users could effectively apply them during interaction with the tool.

4.2 Participants

Fifty-one participants completed the online questionnaire. Participants were recruited through convenience sampling. No specific level of musical expertise was required. Demographic information was collected at the start of the questionnaire, including age, gender, native language, musical training, and drawing or design experience. These variables were recorded to assess sample diversity and to explore whether background factors influenced responses. Participants were also asked to indicate whether they had any hearing or vision impairment that could influence their perception of the stimuli. No participants reported such impairment.

In addition, five participants took part in the in-depth interactive experiment. These participants were recruited separately and participated in individual sessions. The smaller sample size enabled detailed observation, rich qualitative data collection, and in-depth analysis of their interaction behaviour. All participants took part voluntarily and provided informed consent prior to participation.

4.3 Materials

4.3.1 Perceptual Matching Questionnaire

The questionnaire consisted of 32 two-alternative forced-choice (2AFC) tasks designed to test crossmodal correspondences between visual shape properties and auditory features. In each task, participants were instructed to select the option that felt most appropriate based on their intuition, and were explicitly informed that there were no right or wrong answers.

Despite this instruction, responses were analytically classified as correct or incorrect based on whether they aligned with the intended mapping derived from the literature and implemented in the tool. The option that followed the intended correspondence was defined as correct, while the contrasting option was defined as incorrect.

The questionnaire was divided into two sections. In the first section, participants were presented with a single visual image accompanied by two audio stimuli and were asked to select the sound that best matched the image (see Figure 4.1). These items tested the auditory properties loudness, timbral sharpness, tempo, and pitch height. Each property was tested at both ends of its range using semi-contrasting conditions (e.g., maximum sharpness versus medium warmth for timbre and maximum timbre versus medium warmth), and each condition was repeated once, resulting in sixteen items. After completing this section, participants rated their confidence in their answers on a five-point Likert scale.

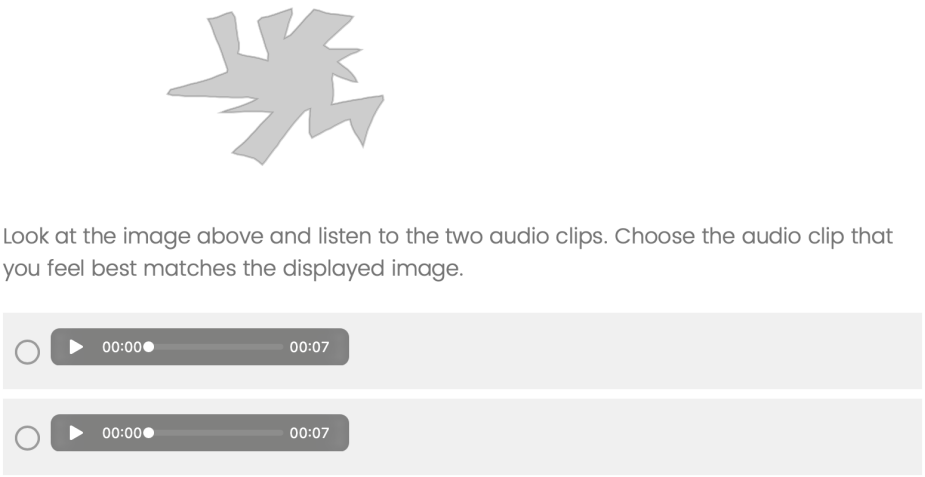


Figure 4.1: Example of a visual-to-audio matching question

In the second section, participants were presented with a single audio stimulus accompanied by two visual images and were asked to select the image that best matched the sound (see Figure 4.2). These items tested visual properties including area, vertical position, circularity, and smoothness. As in the first section, both extremes of each feature were tested and repeated once, resulting in sixteen additional items. The same confidence

question was asked at the end of this section.

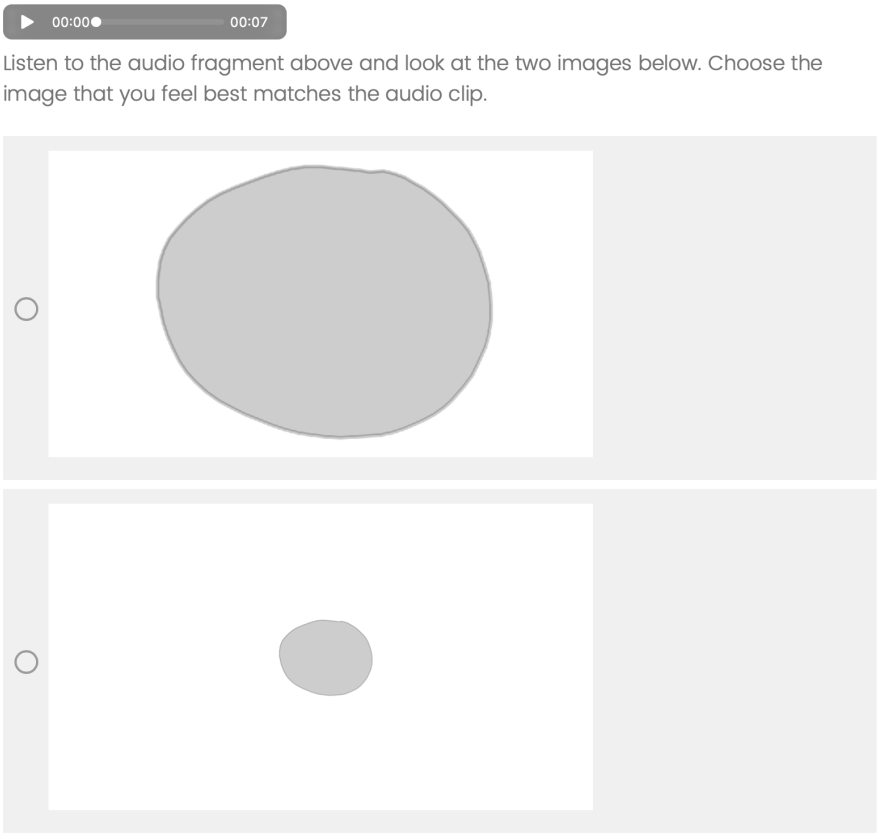


Figure 4.2: Example of an audio-to-visual matching question

All visual and auditory stimuli used in the questionnaire were generated directly from the interactive tool itself, to test the implementation of the correspondences and ensure consistency between the questionnaire and the interactive experiment.

After completing the matching tasks, participants answered four Likert-scale questions (1–5) assessing their perceived intuitiveness of the pairings, and the extent to which images and sounds felt as though they "belonged together". The last two questions examined whether participants relied more on intuition or deliberate reasoning, and how aware they were of specific auditory or visual features during the task. One last open-ended question invited participants to describe which aspects of the stimuli influenced their decisions most strongly.

The full questionnaire is available in Appendix B.

4.3.2 Sketch-Based Sound Synthesis Tool

The in-depth experiment used the sketch-based sound synthesis tool described in Chapter 3. The system allows users to draw shapes on a digital canvas, which are automatically translated into sound through predefined mappings. These correspondences link area, circularity, vertical position, and smoothness to loudness, timbral sharpness, pitch height, and tempo.

Participants interacted with the tool using a standard computer interface with touchscreen and could freely draw and modify shapes while hearing the resulting sound in real time through the headphones.

4.4 Procedure

4.4.1 Questionnaire Procedure

Participants completed the questionnaire online. After providing informed consent and demographic information, they performed the 32 matching tasks in a fixed order. Task order and answer positions were kept consistent across participants to allow users to gradually familiarize themselves with the shapes and sounds and to develop a perceptual baseline for certain features, such as loudness and area. The order of questions was chosen deliberately because some auditory features (e.g., loudness) are perceived more relatively than others and are easier to judge after exposure to multiple stimuli.

Upon completion of the matching tasks, participants answered the Likert-scale and open-ended questions. The questionnaire took approximately ten minutes to complete.

4.4.2 In-depth Experiment Procedure

Each in-depth experiment session followed the same structure. Participants first completed the same perceptual matching questionnaire as the larger group, allowing comparison between perceptual recognition and later interactive behavior.

Participants then received a brief explanation of the experiment with the tool and were informed that there was no correct or incorrect way to use it. This was followed by a free exploration phase, during which participants were encouraged to draw shapes and observe how the sound responded. Participants were asked to think aloud during this phase.

After exploration, participants completed a series of eight directed drawing tasks, each framed as “Draw a shape that sounds...”, targeting loudness (loud/quiet), timbre (shrill/warm), pitch (high/low), and tempo (fast/slow). These tasks were presented in a fixed order and were designed to assess whether participants could intentionally apply the shape–sound correspondences through interaction.

Finally, participants took part in a semi-structured interview focusing on intuitiveness, ease of control, moments of confusion or surprise, expectations, and perceived usability. Each session lasted approximately 20–30 minutes.

4.5 Data Collection and Analysis

Quantitative data from the questionnaire consisted of binary responses to the 2AFC tasks and Likert-scale ratings. For each item, binomial tests were used to determine whether the proportion of participants selecting the intended mapping exceeded chance level (50%). Analyses were conducted both at the item level and aggregated across repeated items testing the same feature to assess overall correspondence strength.

For the in-depth experiment, additional quantitative data were collected automatically, including extracted shape metrics, calculated audio metrics, task identifiers, timestamps, and the number of attempts participants made before being satisfied with a result. These data were analyzed and compared to questionnaire outcomes to identify consistencies and discrepancies between recognition and application.

Qualitative data included open-ended questionnaire responses, screen recordings, audio recordings (with additional consent) of think-aloud comments and interview responses. The analysis of this data focused on recognizing recurring themes and notable or unexpected observations.

5. Results

This chapter reports the results of the questionnaire study and the in-depth experiment. The questionnaire addresses whether participants perceptually recognize the intended shape-sound correspondences, while the in-depth experiment examines whether participants could actively apply these correspondences during interaction with the developed tool. Together, these analyses assess both recognition and use of the implemented mappings.

5.1 Questionnaire Results

5.1.1 Participants

A total of 51 participants completed the questionnaire. No participants reported hearing or vision disabilities that could affect perception of the stimuli.

The sample consisted of 24 Dutch (47%), 18 French (35%), and 9 participants of other nationalities (German, Russian, Latvian, Colombian, Portuguese, Indonesian, and Romanian). 30 participants identified as female and 21 as male. Participants had a mean age of 33.1 years (SD = 16.0), with ages ranging from 19 to 82 years.

Regarding musical experience, 25 participants (49%) reported having at least five years of musical training, 18 (35%) reported no formal musical training, and the remaining 8 reported having had less than five years of training.

5.1.2 Overall Performance

Participants completed 32 forced-choice trials in total: 16 trials in which an image was matched to one of two audio clips, and 16 trials in which an audio clip was matched to one of two images.

Overall performance was well above chance level (50%). On average, participants selected the intended match on 23.6 out of 32 trials, corresponding to 73.7% accuracy. A one-sample t-test against chance performance (16 out of 32 trials) confirmed that this difference was statistically significant, $t(50) = 16.94$, $p < 0.0001$.

Performance was comparable across mapping directions. In the image-to-audio trials, participants selected the expected match on average 11.67 out of 16 trials (72.9%, SD = 2.01), which was significantly above chance (8 out of 16), $t(50) = 13.04$, $p < 0.0001$. In the audio-to-image trials, the mean performance was 11.92 out of 16 (74.5%, SD = 1.89), also significantly above chance, $t(50) = 14.81$, $p < 0.0001$.

5.1.3 Performance by Characteristic

Each correspondence category was tested using four questionnaire items per participant. Accuracy scores at the participant level ranged from 0 to 4 correct responses per correspon-

dence. Table 5.1 summarizes the mean accuracy for each correspondence category, along with standard deviations and the proportion of participants achieving three or more correct responses.

Correspondence	Mean (out of 4)	SD	% Correct	Participants $\geq$ 3 Correct (%)
Shape Features				
Area	3.55	0.61	88.7%	94%
Vertical position	3.10	0.99	77.5%	75%
Smoothness	2.71	0.92	67.6%	63%
Circularity	2.57	0.96	64.2%	51%
Auditory Features				
Timbre	3.20	0.85	79.9%	84%
Loudness	3.16	0.86	78.9%	78%
Pitch	2.75	0.80	68.6%	65%
Tempo	2.57	0.94	64.2%	51%

Table 5.1: Perceptual matching performance for shape and auditory features.

All correspondence categories showed mean performance above chance level (i.e., above 2 out of 4 correct responses). One-sample t-tests were conducted against chance level, and all were found to be significantly above chance ($p<0.0001$). Area-based items showed the highest recognition rates, accompanied by a relatively low variability. Timbre, loudness, and vertical position also demonstrated strong performance, with mean accuracies around 75%. Tempo and circularity showed lower mean accuracy and greater variability.

5.1.4 Subjective Experience and Strategy

Participants reported moderate levels of confidence when performing the matching tasks. On a 5-point scale (1 = not confident at all, 5 = very confident), mean confidence for the audio-to-image task was M = 3.20, SD = 1.08, indicating just above neutral confidence. Confidence was slightly higher for the image-to-audio task (M = 3.41, SD = 1.33), again reflecting moderate confidence overall.

Participants also rated the extent to which the images and sounds "belonged together" at a moderate level (M = 3.49, SD = 1.08, where 1 = not at all and 5 = very much). This suggests that on average, the crossmodal correspondences were perceived reasonably natural.

Regarding strategy, participants reported relying more on intuition than on deliberate reasoning. Intuition ratings (1 = full intuition, 5 = full reasoning) resulted in a mean of M = 2.63, SD = 0.82, indicating a slight tendency towards responding intuitively. At the same time, participants reported relatively high awareness of specific features guiding their decisions (M = 1.88, SD = 0.91, where lower values indicate more awareness).

Open-ended responses to the question "What aspect of the sound or image do you feel influenced your choices the most?" showed that pitch, loudness, and sharpness were most frequently mentioned. Many participants explicitly described mapping rules, such as larger shapes corresponding to louder sounds or higher vertical position corresponding to

higher pitch.

To examine whether subjective confidence was related to objective performance, mean accuracy scores were compared across confidence levels for both parts of the questionnaire separately and combined. No systematic differences in accuracy were observed across confidence categories. Similarly, no relationship was found between total accuracy and reported intuition ratings.

Finally, no differences in performance were observed when grouping participants based on demographic variables, including musical training, drawing or design experience, nationality, age, or gender. This suggests that successfull use of the tool was not dependent on prior expertise or participant background.

5.2 In-Depth Experiment Results

Five participants took part in the in-depth experiment. They were asked to draw shapes that intentionally produced specific sounds (loud, soft, high pitch, low pitch, shrill, warm, fast, slow).

5.2.1 Quantitative Data

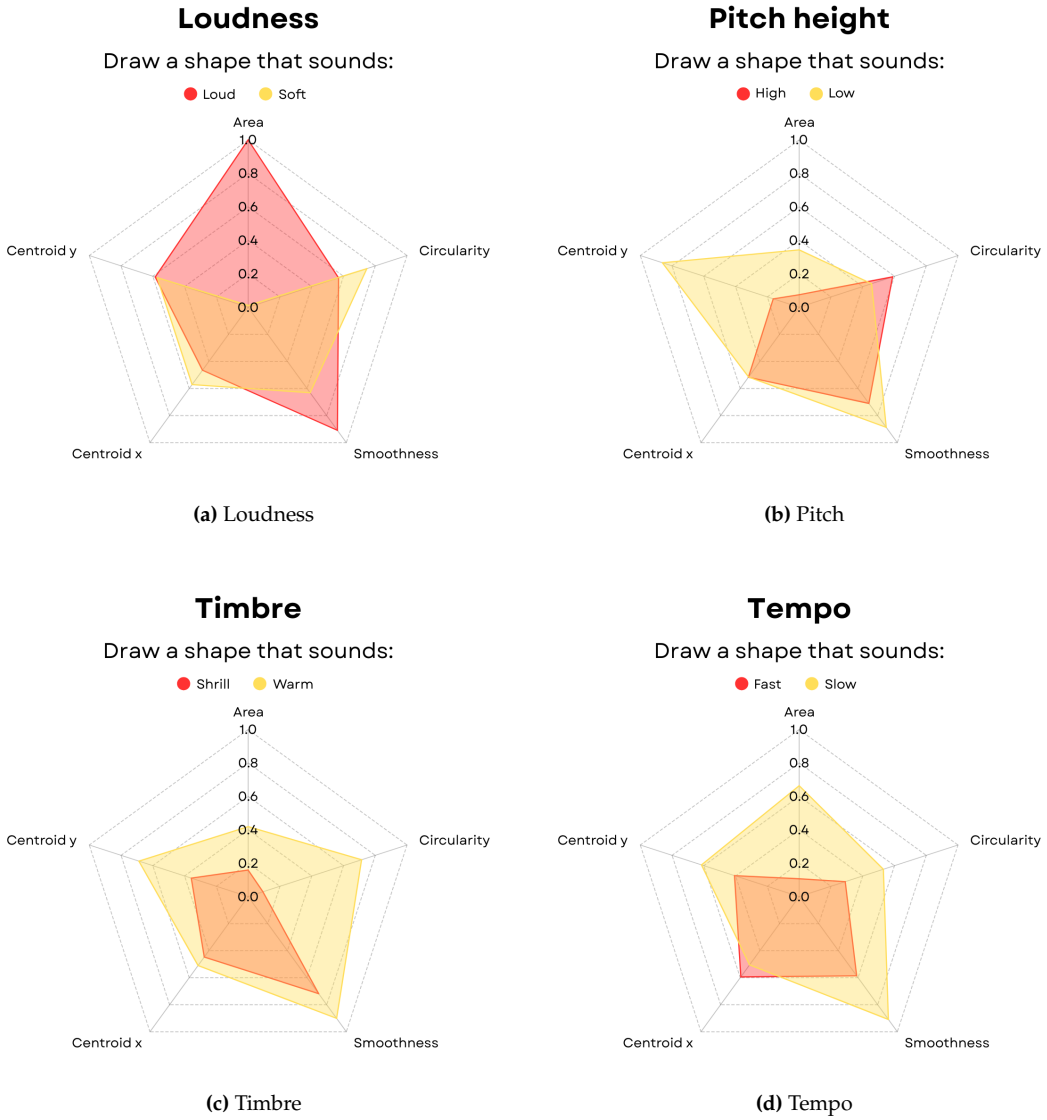
5.2.1.1 Shape Metrics

Table 5.2 presents the mean values of shape metrics per task (averaged across the five participants). Higher values indicate larger area, more circularity, smoother angles, or more extreme centroid positions, depending on the metric (see legend below the table for details).

Task	Area	Circularity	Smoothness	Centroid X	Centroid Y
Loud	**1.00**	0.57	**0.91**	0.47	0.59
Soft	**0.01**	**0.75**	0.63	0.57	0.57
High pitch	0.08	0.59	0.71	0.52	**0.16**
Low pitch	0.34	0.46	0.89	0.52	**0.86**
Shrill	0.16	**0.09**	0.72	**0.45**	0.36
Warm	0.42	0.71	0.90	0.51	0.69
Fast	0.11	0.29	**0.59**	**0.59**	0.41
Slow	0.66	0.53	**0.91**	0.51	0.62

Table 5.2: Mean values of shape metrics for each task. Area, Circularity, and Smoothness are scaled from 0 (small/non-circular/angular) to 1 (big/circular/smooth). Centroid X ranges from 0 (left) to 1 (right), Centroid Y from 0 (high) to 1 (low). Bold values indicate the highest and lowest value per feature.

Figures 5.1a–5.1d visualize these results as radar charts for each sound parameter, contrasting the extreme conditions (e.g., loud vs. soft, fast vs. slow pitch) to show which shape metrics varied the most.



(a) Loudness

(b) Pitch

(c) Timbre

(d) Tempo

Figure 5.1: Radar charts visualizing mean shape metrics for each task per sound parameter.

Paired-sample t-tests were conducted to determine which differences between conditions were statistically significant:

- **Loud versus soft:** Significant differences were found for area ($p = 0.0238$), while circularity, smoothness, and centroid positions did not differ significantly (Figure 5.1a).
- **High versus low pitch:** Significant differences were observed for area ($p = 0.0269$), circularity ($p = 0.0059$), and centroid Y ($p < 0.0001$), with a minimal effect for smoothness ($p = 0.0846$; Figure 5.1b).
- **Shrill versus warm timbre:** Area ($p = 0.0280$), circularity ($p = 0.0014$), and smoothness ($p = 0.0207$) differed significantly. Centroid Y showed a minimal effect ($p = 0.0762$; Figure 5.1c).
- **Fast versus slow tempo:** Only smoothness differed significantly ($p = 0.0350$; Figure 5.1d), while other metrics remained comparable.

5.2.1.2 Sound Metrics

The sound metrics for each task were derived from the recorded shape features, sometimes combining multiple shape features into a single auditory metric. These metrics were converted into five levels (0–4) to send from the canvas (Processing) to the synthesizer (Sonic Pi) for sound generation. Table 5.3 presents the mean values per task, along with the average number of attempts participants required to achieve the intended sound.

Task	Pitch	Timbre	Tempo	Loudness	# of tries
Loud (volume = max. 84)	3.2	3.4	3.6	**80.8**	1.8
Soft (volume = 0)	1.2	1.8	0.6	**1.4**	**1.2**
High pitch (pitch = 0)	**0**	1.8	0.6	6.6	1.4
Low pitch (pitch = 4)	**2.8**	3.0	1.6	33.4	1.8
Shrill (timbre = 0)	0.8	**1.4**	0.2	14.8	2.8
Warm (timbre = 4)	2.6	**3.6**	2.0	38.6	**3.6**
Fast (tempo = 0)	0.8	1.0	**0.4**	10.0	1.8
Slow (tempo = 4)	3.0	3.4	**2.6**	59.2	1.8

Table 5.3: Mean audio metrics and number of attempts per task. Bold values indicate the highest and lowest value per feature.

5.2.2 Qualitative Data

Qualitative data were collected by using the think-aloud method during interaction with the tool and a short semi-structured interview, conducted after the tasks were completed. The analysis focused on first time use interaction during free exploration of the tool, strategies and difficulties during the sound-generation tasks, and participants' subjective experience of intuitiveness and usability.

5.2.2.1 Free Exploration Phase

Before the structured sound-generation tasks, participants were asked to freely explore the sketchpad synthesizer tool for a few minutes. They were instructed to draw without specific goals, told they could delete shapes at any time, and were asked to express their thoughts and expectations and explain their actions while interacting with the tool.

All participants were able to use the basic functionality of the tool immediately. No clarification was required regarding drawing or deleting shape or the resulting sound. However, exploratory behaviour was generally conservative. Although participants were informed that multiple shapes could be drawn simultaneously, some initially drew only one, or sometimes two, shapes at a time. As a result, they did not spontaneously explore how multiple shapes combined into a more complex musical output. When they were suggested to experiment with drawing several shapes simultaneously to hear more instruments together, participants adjusted their behaviour accordingly.

Similarly, participants did not always explore the extremities of the shape features. Differences between shapes were sometimes subtle (e.g., slightly smoother versus slightly sharper), resulting in moderate rather than extreme sound variations. Consequently, some participants never fully encountered the slowest possible tempo or the lowest possible pitch during exploration.

5.2.2.2 Sound-Generation Tasks

All participants completed the eight directed sound-generation tasks. Most tasks required between one and three attempts before the participants were satisfied with the outcome. The tool does not allow adjustments to the shape after drawing. Therefore, participants had to delete and redraw shapes when they were not satisfied with the outcome.

During the think aloud phase, loudness and pitch height were consistently mentioned as the easiest features to control. Participants reported that the tasks that involved loudness and pitch were generally straightforward. They found the relationship between shape size and loudness, as well as between vertical position and pitch, easy and intuitive. These mappings were typically achieved in one or two attempts. Timbre and tempo were found to be more challenging. Two participants struggled particularly with producing specific timbral qualities (shrill or warm), and several participants found it difficult to achieve a sufficiently slow tempo. In some cases, participants reached a sound that was slower than previous attempts but not as slow as possible in the tool. Although further extremes were technically possible, participants did not always discover them and settled with a less extreme outcome.

In three cases across all tasks and participants, the participant moved on to the next task without being fully satisfied, but uncertain about how to further manipulate the shape to achieve the intended sound. This occurred twice for slow tempo, and once for shrill timbre. One participant (without musical experience) initially struggled with the concept of "shrill" timbre and, as a result, had difficulty producing the intended sound. After a brief clarification of the term, the participant was immediately able to achieve a satisfactory result.

5.2.2.3 Observed Interaction Strategies

Across participants, consistent strategies emerged for achieving the target sounds, aligning with the previous findings:

- Loud/soft sounds were primarily controlled by increasing or decreasing shape size.
- High/low pitch was controlled mainly through vertical positioning on the canvas.
- Shrill sounds were associated with angular or spiky shapes.
- Warm sounds were represented by smoother and more circular shapes.
- Fast/slow tempo was often approached through changes in size, and for some participants, through vertical positioning.

Several participants placed slower-sounding shapes lower on the canvas than faster ones. However, a vertical mapping of tempo was not implemented, thus did not influence sound.

A common behavioural pattern was perceptual calibration. Participants frequently began with an exaggerated shape, listened to the resulting sound, and then adjusted accordingly. Subsequent attempts were either more exaggerated or structurally different. This pattern shows refinement with control rather than random trial-and-error exploration.

5.2.2.4 Post-Tasks Interview

The semi-structured interviews confirmed many of the observational findings. Participants consistently reported that the tool felt intuitive and easy to use, no prior musical knowledge was required, and that the relationships between shape and sound were generally clear.

Loudness and pitch were most often mentioned as clearly controllable dimensions. Timbre and tempo were described as slightly more difficult, particularly when multiple visual properties seemed to influence the resulting sound simultaneously.

When asked whether the shapes and sounds felt naturally connected or learned, participants described the mappings as intuitive rather than arbitrary. One musically trained participant (10+ years of musical experience) indicated that the tool would likely also be accessible to users without musical background and could serve as an introductory step into music-making. One participant mentioned that the tool occasionally felt resistant when attempting to reach more extreme parameter values, particularly for slow tempo. Even though the participant reported being able to produce a relatively slow sound, it was not as slow as desired.

6. Discussion

This chapter interprets how participants perceived and applied the shape–sound correspondences. It highlights which mappings were most intuitive and where challenges arose.

6.1 Recognition versus Application of CMCs

The questionnaire results demonstrated high perceptual recognition of the implemented correspondences. Overall performance (73.7% accuracy) was significantly above chance level, and all correspondence categories performed above chance individually. Among the visual features, mappings involving shape area showed the highest recognition, followed by vertical position, smoothness, and circularity. For sound parameters, timbre and loudness showed the highest recognition rates, while pitch and particularly tempo showed lower scores. Importantly, performance was comparable across the two mapping directions (image-to-audio and audio-to-image), which shows that the correspondences are bidirectional. Participants were able both to connect a sound to a visual shape and to connect a shape to a sound, indicating that the relationships between the visual and auditory features were perceived as meaningful and clear.

The in-depth experiment further showed that participants were also able to actively apply several of these correspondences when interacting with the sketch-based synthesizer. Loudness and pitch were found to be the easiest to control and typically required only a small number of attempts to draw the shape. Shape size was clearly used to influence the loudness of a sound, and vertical position was consistently used to distinguish high- and low-pitched sounds. These findings align with both the questionnaire results and existing literature on crossmodal correspondences. Next to this, participants also used other correspondences consistent with theoretical expectations. For example, shrill timbres were generally used to create more angular or spiky shapes, whereas warmer timbres were associated with smoother and more circular forms. This indicates that participants successfully applied several of the expected mappings when intuitively producing shapes for specific sounds, which shows that certain correspondences are not only perceptually recognized but can also be actively used in an interactive system.

However, perceptual recognition does not automatically imply ease of application. While some correspondences were applied easily, others proved more difficult during active interaction. One outstanding example of this difference concerns timbre.

6.2 Timbre

In the questionnaire, timbre correspondences were recognized relatively strongly (79.9% accuracy). In the in-depth experiment, however, participants often required more attempts to produce a shrill or warm sound they were satisfied with. Interestingly, this difficulty

does not appear to be caused by a lack of intuitive understanding. Participants generally drew shapes that matched theoretical expectations for timbral correspondences. For example, shrill sounds were typically represented by sharper and more angular shapes, while warm sounds were drawn as smoother and more rounded forms. As this is according to results from previous research, this indicates that participants understood the intended relationship between shape and timbre. However, the sounds produced by the tool did not always match participants' expectations, suggesting that the difficulty arose primarily from the implementation of the mapping rather than from the perceptual correspondence itself.

In the tool, timbre was controlled by two visual features: smoothness (weight = 0.8) and circularity (weight = 0.2). While circularity clearly differed between the shapes representing a shrill versus a warm sound, smoothness did not show as big of a difference as would be expected. Smoothness was calculated by measuring how gradually the turning angles of the shape changed along the drawing line. If consecutive angles changed slowly, the shape was considered smooth, whereas abrupt changes in angle resulted in lower smoothness values. The smoothness value was obtained by summing the acceleration of angular change along the drawing line and normalizing the result between 0 (very sharp) and 1 (very smooth).

Although this method provides a good mathematical measure of smoothness, it did not always correspond well with how participants visually perceived the shapes they drew. In several cases, shapes that appeared clearly more angular than others produced only small differences in the calculated smoothness value. This indicates that participants attempted to manipulate smoothness, but the metric did not capture these variations strongly enough. As a result, participants sometimes needed to exaggerate their shapes more than expected in order to produce the intended timbre. Relatively extreme shapes were required before big changes in the sound occurred. This explains why participants' drawing strategies often aligned with theoretical expectations while the resulting sound still did not fully satisfy them.

This finding highlights that perceptual correspondences can only support intuitive interaction when the computational implementation reflects how users visually perceive the features they manipulate. When the calculated metric does not align with the perceived feature, the mapping becomes more difficult to control, even if the underlying correspondence is conceptually correct. One possible improvement would be to increase the sensitivity of the system to changes in smoothness. Rather than normalizing the metric across the full theoretical range between 0 and 1, a smaller, more relevant segment of this range could be normalized instead, with values outside this segment clamped to 0 or 1. This would reduce the need for participants to draw exaggerated shapes and allow smaller variations in smoothness to be detected more easily and translated more clearly into musical output.

6.3 Tempo

Tempo showed more mixed results than the other sound parameters. In the questionnaire, tempo correspondences showed relatively lower recognition (64.2%) compared to other mappings. In the in-depth experiment, while participants generally did not struggle with producing a fast sound, a very slow tempo proved particularly challenging. In two cases

participants even moved on to the next task without reaching a fully satisfactory result. The analysis of the drawn shapes showed that when the average shape metrics were compared across the fast and slow tasks, several visual features appeared to differ, including area, circularity, smoothness, and vertical position.

However, a closer inspection at the participant level showed variation in strategy. All participants used shape size to influence tempo, drawing larger shapes for slower sounds and smaller shapes for faster sounds. This aligns with the implemented mapping, where tempo was primarily influenced by area. Four out of five participants also varied circularity, which was the second feature used in the implementation. Beyond these two expected metrics, participants used different additional strategies. Two participants additionally used smoothness, vertical and even horizontal position of the shape to manipulate tempo, even though these parameters were not linked to tempo in the system. One particularly interesting observation was that these two participants placed their fast-tempo shapes in the top-right area of the canvas. One participant explicitly mentioned expecting faster sounds to be located further to the right. While this spatial association has not been found in the literature, it suggests that participants may intuitively expect additional relationships between visual features and tempo. Overall, the strategies used in the tempo task were more diverse than for other sound parameters. This variability likely reflects the relatively weak and less well-established nature of shape–tempo correspondences in existing research. Most crossmodal correspondence studies focus on isolated sounds rather than musical structures, meaning that tempo has received comparatively little attention. Nevertheless, participants generally moved in the expected direction when attempting to produce slower or faster sounds. The main difficulty was reaching the extreme values of the tempo range within the tool.

Another factor that likely contributed to this limitation is that producing the slowest possible tempo required shapes that relied on values at the ends of the parameter range (a very large, circular shape). Participants rarely explored these ends during both the free exploration phase and the structured tasks. Drawing shapes very close to the edges of the canvas or large shapes that cover a big part of the canvas may feel visually unnatural, which can discourage users from approaching these limits.

Thus, the difficulty observed in the tempo task appears to result from a combination of two factors: the relatively weak perceptual correspondence between shape and tempo and participants' limited exploration of the extreme parameter values. Despite this, the inclusion of tempo in the tool remains important. Tempo is a fundamental musical parameter, and excluding it would significantly limit the musical expressiveness of the system. Although the correspondence may be less intuitive than others, participants were still able to influence tempo in the intended direction, suggesting that its inclusion does not undermine the usability of the tool.

6.4 Interaction Behaviour

The qualitative observations revealed a consistent interaction pattern that can be described as perceptual calibration. Participants typically began with an exaggerated shape, listened to the resulting sound, and then refined the shape based on auditory feedback. These

adjustments were systematic rather than random, indicating deliberate exploration of the mapping. During the free exploration phase, however, participants were generally slightly careful and mindful in exploring the tool. Most participants initially had generally only one or two shapes on the canvas at the same time and rarely experimented with more exaggerated shapes that reached the ends of the parameters. Even though they were informed that multiple shapes could be combined, this behaviour was not immediately discovered by all participants. This hesitant exploration suggests that users initially focus on understanding the workings of the tool and how their shapes influence the sound, before experimenting with more extreme or complex shapes and compositions. As a result, during the drawing tasks, participants often evaluated sounds relative to previously heard outputs rather than relative to the full possible range of the system.

Despite this conservative exploration behaviour, participants consistently reported that the tool felt intuitive and easy to use. Even participants without or with less musical experience were able to understand the mappings and complete the tasks within a small number of attempts. This supports the idea that perceptually grounded mappings can reduce the cognitive barrier typically associated with digital sound synthesis tools. At the same time, the results indicate that the discoverability of the tool's full expressive range could be improved. While users quickly understood the basic interaction, they did not always explore the full set of possibilities available.

Future iterations of the tool could address this by providing clearer guidance during initial use. For example, the interface could include example compositions or demonstration shapes that showcase the range of possible sounds. This would give users a clearer understanding of the potential of the system and provide inspiration for their own creations. Another improvement would be to allow users to modify shapes after drawing them. In the current version of the tool, shapes must be deleted and redrawn to change their properties. Allowing users to adjust parameters such as size, position, smoothness, or circularity directly would make it easier to reach extreme values and refine sounds more precisely. Such functionality could be implemented through simple interactive controls or sliders that appear when a shape is selected. This would enhance usability without fundamentally increasing the complexity of the interface.

6.5 Casual Creators

The findings align well with the concept of casual creators, which emphasizes accessibility, playfulness, and immediate usability. Participants were able to interact with the tool with minimal prior instruction, and the mappings between shapes and sounds were generally described as intuitive rather than learned. Several aspects of the results support this interpretation. First, the basic interaction required no explanation, and participants were able to start drawing and producing sound immediately. Second, the most robust correspondences were easily understood and applied (e.g., size-loudness and vertical position-pitch). Third, even participants without musical training can achieve the intended sounds within a small number of attempts. An especially relevant finding for the casual creator perspective was that performance did not differ meaningfully across demographic groups or prior experience levels. Users with and without musical or design backgrounds performed similarly,

suggesting that successful interaction with the tool relied more on intuitive perception than on learned expertise.

At the same time, the study shows the importance of the tool being discoverable alongside being intuitive. While the core mappings were generally easy for users to understand, participants did not always explore the full range of possibilities offered by the system. Future versions of the tool could therefore benefit from design improvements that encourage experimentation without reducing its accessibility. For example, this could include enabling adjustments to the shapes after drawing, visual feedback indicating how close a shape is to the parameter extremes, or the inclusion of example shapes and demonstrations to inspire users when they first use the tool, or even during later interactions.

7. Conclusions

This study attempted to answer the following main research question:

> **To what extent can empirically established crossmodal correspondences between visual shape features and sound parameters be implemented in a sketch-based music synthesis tool that supports intuitiveness and meaningful interaction?**

This question was addressed through two sub-questions:

1. *Do people recognize and agree with the implemented shape-sound mappings?*
2. *Can users actively and intentionally apply these correspondences when creating music through drawing, and how intuitive does this interaction feel in practice?*

To address this question, a sketch-based music synthesis tool was designed and implemented that translates visual shapes into musical sounds through crossmodal mappings derived from existing literature. The tool was evaluated through two complementary experiments: a questionnaire examining whether participants perceptually recognized the implemented correspondences, and an in-depth experiment investigating whether users could actively apply these mappings when interacting with the tool.

The results demonstrate empirically established crossmodal correspondences can be successfully implemented in an interactive system and can support intuitive interaction with the system. In the questionnaire study, participants recognized all implemented correspondences significantly above chance level, indicating that the mappings between visual and auditory features were perceptually meaningful. The findings align with previous research on CMCs and suggest that the mappings implemented in the tool reflect these established associations. The in-depth experiment further showed that participants were able to actively apply these correspondences when creating music through drawing. Participants consistently manipulated shape size to influence loudness and vertical position to control pitch, often requiring only a small number of attempts to achieve the desired sound. These were the most robust mappings, but overall participants generally approached all tasks in the expected direction. This indicates that certain CMCs can function not only as perceptual associations but also as mechanisms for interaction in creative systems.

At the same time, the findings highlight that successful implementation of CMCs depends strongly on how the visual features are represented and translated into sound parameters. The results show the importance of aligning feature measurements with how users visually perceive shapes. For example, although participants intuitively attempted to influence timbre through changes in shape smoothness, the implemented smoothness metric did not always capture these visual differences strongly enough. This resulted in sounds that did not always fully match participants' expectations. Similarly, the tempo mapping showed greater variation in participants' strategies, partly reflecting the relatively weaker and less extensively studied correspondences between shape features and musical

tempo. These observations show that the perceptual strength of a correspondence does not automatically guarantee intuitive interaction if the implementation in the system does not adequately reflect how users visually interpret the manipulated features. Careful design of how shape features are computationally measured and translated into sound parameters is thus essential for creating an effective crossmodal interaction system.

Beyond evaluating individual correspondences, this study contributes a functional prototype of a sketch-based music synthesis tool that demonstrates how empirically established mappings can support accessible music creation. Participants were able to interact with the tool without prior instruction and reported that the system felt intuitive and enjoyable to use, even for users without musical experience. This suggests that crossmodal correspondences can help make digital music tools more approachable and support creative interaction without requiring technical musical knowledge.

Nevertheless, several opportunities remain for further development of the tool. Future work could focus on refining the implementation of certain shape metrics, particularly smoothness, to better align computational measurements with perceived visual differences. Allowing users to manipulate shapes after drawing them, for example by adjusting parameters through interactive controls, could further improve usability and help users explore the full range of possible sounds.

Figure 7.1: Conceptual future interface with controls for shape and colour.

Another interesting direction for future development is the integration of colour as an additional visual dimension. Colour is strongly linked to sound perception in many crossmodal correspondence studies and has been shown to influence how people associate visual stimuli with musical features. Incorporating colour into the sketch-based interface could therefore significantly expand the expressive possibilities of the tool and enable richer mappings between visual and auditory domains. Figure 7.1 shows a conceptual design for a future version of the interface, with added functionality for shape modification and colour control. While this extension was beyond the scope of the current project, it is a promising direction for future research and design. Further evaluation with larger partic-

ipant groups would also provide deeper insight into how users interact with crossmodal music tools and how such systems can support different creative practices. As the current in-depth experiment involved a relatively small number of participants, additional studies would be valuable when developing the tool into a more mature system.

In conclusion, the results of this study demonstrate that empirically established CMCs between visual shapes and sound can be meaningfully implemented in a sketch-based music synthesis tool. Participants recognized and agreed with the implemented shape-sound mappings, and were able to actively apply these correspondences when creating music through drawing shapes that correspond intuitively to sound.

By translating findings from crossmodal perception research into an interactive system, this study shows the potential of CMCs as a foundation for designing intuitive and accessible music creation interfaces. More broadly, it bridges perceptual research and interactive system design, showing how insights from perception can support the design of tools for creative expression in digital media.

Appendices

Appendix A: Source Code

The complete source code for the sketch-based synthesis tool is available on GitHub at the following repository: **github.com/marlindevdb/visualsynth**

Appendix B: Questionnaire

The online questionnaire used in this study is available on Qualtrics at the following link: **Qualtrics Questionnaire**

Bibliography

[1] Y.-C. Chen and C. Spence. "Assessing the role of the 'unity assumption' on multisensory integration: A review". In: *Frontiers in psychology* 8 (2017), p. 445.

[2] Y.-C. Chen and P.-C. Huang. "Examining the automaticity and symmetry of sound–shape correspondences". In: *Frontiers in Psychology* 14 (2023), p. 1172946.

[3] B. Mesz et al. "Marble melancholy: using crossmodal correspondences of shapes, materials, and music to predict music-induced emotions". In: *Frontiers in psychology* 14 (2023), p. 1168258.

[4] S. Lacey et al. "Synesthesia strengthens sound-symbolic cross-modal correspondences". In: *European Journal of Neuroscience* 44.9 (2016), pp. 2716–2721.

[5] K. K. Evans and A. Treisman. "Natural cross-modal mappings between visual and auditory features". In: *Journal of vision* 10.1 (2010), pp. 6–6.

[6] O. Deroy, A.-S. Crisinel, and C. Spence. "Crossmodal correspondences between odors and contingent features: odors, musical notes, and geometrical shapes". In: *Psychonomic bulletin & review* 20.5 (2013), pp. 878–896.

[7] C. Spence. "Crossmodal correspondences: A tutorial review". In: *Attention, Perception, & Psychophysics* 73.4 (2011), pp. 971–995.

[8] C. Spence. "Simple and complex crossmodal correspondences involving audition". In: *Acoustical Science and Technology* 41.1 (2020), pp. 6–12.

[9] C. V. Parise. "Crossmodal correspondences: Standing issues and experimental guidelines". In: *Multisensory research* 29.1-3 (2016), pp. 7–28.

[10] N. Sagiv and J. Ward. "Crossmodal interactions: lessons from synesthesia". In: *Progress in brain research* 155 (2006), pp. 259–271.

[11] A. Ćwiek et al. "The bouba/kiki effect is robust across cultures and writing systems". In: *Philosophical Transactions of the Royal Society B* 377.1841 (2022), p. 20200390.

[12] R. Chiou and A. N. Rich. "Cross-modality correspondence between pitch and spatial location modulates attentional orienting". In: *Perception* 41.3 (2012), pp. 339–353.

[13] L. Reymore and D. T. Lindsey. "Color and tone color: audiovisual crossmodal correspondences with musical instrument timbre". In: *Frontiers in Psychology* 15 (2025), p. 1520131.

[14] M. Adeli, J. Rouat, and S. Molotchnikoff. "Audiovisual correspondence between musical timbre and visual shapes". In: *Frontiers in human neuroscience* 8 (2014), p. 352.

[15] M. Perlman, R. Dale, and G. Lupyan. "Iconicity can ground the creation of vocal symbols". In: *Royal Society open science* 2.8 (2015).

[16] M. Perlman and G. Lupyan. "People can create iconic vocalizations to communicate various meanings to naıve listeners". In: *Scientific reports* 8.1 (2018), p. 2634.

[17] K. Compton. "Casual creators: Defining a genre of autotelic creativity support systems". PhD thesis. University of California, Santa Cruz, 2019.

[18] M. M. Murray and M. T. Wallace. *The neural bases of multisensory processes*. CRC press, 2011.

[19] B. De Gelder and P. Bertelson. "Multisensory integration, perception and ecological validity". In: *Trends in cognitive sciences* 7.10 (2003), pp. 460–467.

[20] C. Spence and S. Squire. "Multisensory integration: maintaining the perception of synchrony". In: *Current Biology* 13.13 (2003), R519–R521.

[21] K. Knöferle and C. Spence. "Crossmodal correspondences between sounds and tastes". In: *Psychonomic bulletin & review* (2012), pp. 1–15.

[22] L. Ginsberg. "A case of synaesthesia". In: *The American Journal of Psychology* (1923), pp. 582–589.

[23] A. Glicksohn and A. Cohen. "The role of cross-modal associations in statistical learning". In: *Psychonomic Bulletin & Review* 20.6 (2013), pp. 1161–1169.
[24] N. Di Stefano and C. Spence. "Perceptual similarity: Insights from crossmodal correspondences". In: *Review of Philosophy and Psychology* 15.3 (2024), pp. 997–1026.
[25] A. Ravignani and R. Sonnweber. "Chimpanzees process structural isomorphisms across sensory modalities". In: *Cognition* 161 (2017), pp. 74–79.
[26] V. Walsh. "A theory of magnitude: common cortical metrics of time, space and quantity". In: *Trends in cognitive sciences* 7.11 (2003), pp. 483–488.
[27] L. Puigcerver et al. "Vertical mapping of auditory loudness: Loud is high, but quiet is not always low". In: *Psicológica Journal* 40.2 (2019), pp. 85–104.
[28] L. E. Marks. "On associations of light and sound: The mediation of brightness, pitch, and loudness". In: *The American journal of psychology* (1974), pp. 173–188.
[29] C. C. Pratt. "The spatial character of high and low tones." In: *Journal of Experimental Psychology* 13.3 (1930), p. 278.
[30] P. Walker et al. "Preverbal infants' sensitivity to synaesthetic cross-modality correspondences". In: *Psychological science* 21.1 (2010), pp. 21–25.
[31] S. Dolscheid et al. "The sound of thickness: Prelinguistic infants' associations of space and pitch". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 34. 34. 2012.
[32] G. M. Bidelman and A. Krishnan. "Neural correlates of consonance, dissonance, and the hierarchy of musical pitch in the human brainstem". In: *Journal of Neuroscience* 29.42 (2009), pp. 13165–13171.
[33] S. S. Stevens and J. Volkmann. "The relation of pitch to frequency: A revised scale". In: *The American Journal of Psychology* 53.3 (1940), pp. 329–353.
[34] C. Palmer and S. Holleran. "Harmonic, melodic, and frequency height influences in the perception of multivoiced music". In: *Perception & psychophysics* 56.3 (1994), pp. 301–312.
[35] P. Boomsliter and W. Creel. "The long pattern hypothesis in harmony and hearing". In: *Journal of Music Theory* 5.1 (1961), pp. 2–31.
[36] S. McAdams. "The perceptual representation of timbre". In: *Timbre: Acoustics, perception, and cognition*. Springer, 2019, pp. 23–57.
[37] S. Handel. "Timbre perception and auditory object identification". In: *Hearing* 2 (1995), pp. 425–461.
[38] A. Zacharakis. "The poetry of senses: exploring semantic mediation in timbre-aroma correspondences". In: *Frontiers in Psychology* 16 (2025), p. 1520046.
[39] Z. Eitan and R. Y. Granot. "How music moves:: Musical parameters and listeners images of motion". In: *Music perception* 23.3 (2006), pp. 221–248.
[40] H. Fastl and E. Zwicker. *Psychoacoustics: facts and models*. Vol. 22. Springer Science & Business Media, 2006.
[41] P. N. Juslin and P. Laukka. "Communication of emotions in vocal expression and music performance: Different channels, same code?" In: *Psychological bulletin* 129.5 (2003), p. 770.
[42] J. London. *Hearing in time: Psychological aspects of musical meter*. Oxford University Press, 2012.
[43] Z. Pizlo. *3D shape: Its unique place in visual perception*. Mit Press, 2010.
[44] D. Usnadze. "Ein experimenteller Beitrag zum Problem der psychologischen Grundlagen der Namengebung". In: *Psychologische Forschung* 5 (1924), pp. 24–43.
[45] M. Fort, A. Martin, and S. Peperkamp. "Consonants are more important than vowels in the bouba-kiki effect". In: *Language and speech* 58.2 (2015), pp. 247–266.
[46] D. Maurer, T. Pathman, and C. J. Mondloch. "The shape of boubas: Sound–shape correspondences in toddlers and adults". In: *Developmental science* 9.3 (2006), pp. 316–322.
[47] L. E. Marks. "On cross-modal similarity: Auditory–visual interactions in speeded discrimination." In: *Journal of experimental psychology: Human perception and performance* 13.3 (1987), p. 384.

[48] C. V. Parise and C. Spence. "Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test". In: *Experimental Brain Research* 220.3 (2012), pp. 319–333.
[49] W. Köhler. *Gestalt psychology*. G. Bell, 1929.
[50] S. K. Roffler and R. A. Butler. "Localization of tonal stimuli in the vertical plane". In: *The Journal of the Acoustical Society of America* 43.6 (1968), pp. 1260–1266.
[51] P. Lidji et al. "Spatial associations for musical stimuli: A piano in the head?" In: *Journal of Experimental Psychology: human perception and performance* 33.5 (2007), p. 1189.
[52] E. Rusconi et al. "Spatial representation of pitch height: the SMARC effect". In: *Cognition* 99.2 (2006), pp. 113–129.
[53] A. Gallace and C. Spence. "Multisensory synesthetic interactions in the speeded classification of visual size". In: *Perception & psychophysics* 68.7 (2006), pp. 1191–1203.
[54] C. J. Mondloch and D. Maurer. "Do small white balls squeak? Pitch-object correspondences in young children". In: *Cognitive, Affective, & Behavioral Neuroscience* 4.2 (2004), pp. 133–136.
[55] L. B. Smith and M. D. Sera. "A developmental analysis of the polar structure of dimensions". In: *Cognitive psychology* 24.1 (1992), pp. 99–142.
[56] Z. Eitan and R. Timmers. "Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context". In: *Cognition* 114.3 (2010), pp. 405–422.
[57] A. Rojczyk. "Sound symbolism in vowels: Vowel quality, duration and pitch in sound-to-size correspondence". In: *Poznań Studies in Contemporary Linguistics PSiCL* 47.3 (2011), pp. 602–615.
[58] K. Knoeferle et al. "What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings". In: *Scientific reports* 7.1 (2017), p. 5562.
[59] R. Schorn, D. Abfalter, and A. Brunner-Sperdin. "It's Relative! The Cross-Modal Effects of Music Density on Perception of Product Size". In: *SAGE Open* 14.4 (2024), p. 21582440241292714.
[60] A. Salgado-Montejo et al. "The sweetest thing: The influence of angularity, symmetry, and the number of elements on shape-valence and shape-taste matches". In: *Frontiers in Psychology* 6 (2015), p. 1382.
[61] E. Thoret et al. "Seeing circles and drawing ellipses: when sound biases reproduction of visual motion". In: *PloS one* 11.4 (2016), e0154475.
[62] B. Buxton. *Sketching user experiences: getting the design right and the right design*. Morgan kaufmann, 2010.
[63] G. Kurtenbach and W. Buxton. "Issues in combining marking and direct manipulation techniques". In: *Proceedings of the 4th annual ACM symposium on User interface software and technology*. 1991, pp. 137–144.
[64] A. Norman Donald. *The design of everyday things*. MIT Press, 2013.