Quantum GestART: identifying and applying correlations between mathematics, art, and perceptual organization

Maria Mannone, Federico Favali, Balandino Di Donato & Luca Turchet

Published online: 11 Mar 2020.

Submit your article to this journal

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

Check for updates

Quantum GestART: identifying and applying correlations between mathematics, art, and perceptual organization

Maria Mannone [a]*, Federico Favali [b], Balandino Di Donato [c], and Luca Turchet [d]

ᵃDepartment of Mathematics and Informatics, University of Palermo, Palermo, Italy;
ᵇUniversidad Nacional de Tres de Febrero, Buenos Aires, Argentina;
ᶜInformatics Department, University of Leicester, Leicester, UK;
ᵈDepartment of Information Engineering and Computer Science, University of Trento, Trento, Italy

Mathematics can help analyze the arts and inspire new artwork. Mathematics can also help make transformations from one artistic medium to another, considering exceptions and choices, as well as artists' individual and unique contributions. We propose a method based on diagrammatic thinking and quantum formalism. We exploit decompositions of complex forms into a set of simple shapes, discretization of complex images, and Dirac notation, imagining a world of "prototypes" that can be connected to obtain a fine or coarse-graining approximation of a given visual image. Visual prototypes are exchanged with auditory ones, and the information (position, size) characterizing visual prototypes is connected with the information (onset, duration, loudness, pitch range) characterizing auditory prototypes. The topic is contextualized within a philosophical debate (discreteness and comparison of apparently unrelated objects), it develops through mathematical formalism, and it leads to programming, to spark interdisciplinary thinking and ignite creativity within STEAM.

Keywords: Gestural similarity; Gestalt; diagrams; Dirac notation; sonification

2010 Mathematics Subject Classification: 81-00; 18C10; 42-00

1. Introduction: philosophical background

Understanding things that surround us in daily life also includes the ability to classify and even "simplify" them. This means collecting different things (objects, images, shapes of living beings) within classes or categories. This also means recovering similarities between objects and between their transformational processes. Mathematics, and especially some branches of abstract mathematics such as category theory, provide the necessary formalism to compare objects via transformations, and transformations between transformations. Among the things around us, there are also sounds, and, as a subset of sounds, music. Musical structures also contain objects – notes and recognizable themes – that can be transformed.

Is it possible to find a key, a technique, to connect not only objects that easily appear as similar but also objects whose similarities are hidden? Is it possible to simplify an object into essential lines and to translate these lines from apparently unrelated objects, such as a tree and a symphony, a book of geometry and a painting? Can we imagine a world of "visual prototypes" and a world

*Corresponding author. Email: mariacaterina.mannone@unipa.it, manno012@umn.edu

of "sound prototypes," that can be eventually related within organized structures? Unity within the variety of human thinking can remind us of Leonardo da Vinci and of the Italian Renaissance. While researching shared structures between sounds and images, and sounds and shapes/forms from nature, would it be possible and useful to apply some principles of a cognitive theory that specifically investigates forms and their organization? To this aim, we can consider ideas from Gestalt theory, a theory of the form seen as a whole that is more than the sum of its individual parts. This applies to visuals, but also to sounds (Reybrouck 1997; Lerdhal and Jackendoff 1983). Gestalt is related to perceptual organization. We may wonder if, in the arts, the generic "form" can be an "artistic idea" that is embodied in this or that specific artistic code. For this reason, we can ask: how are some cognitive topics, such as Gestalt, relevant to a mathematical theory of the arts?

The search for hidden connections is reminiscent of contemporary philosophical approaches. French philosopher Gilles Deleuze[1] (1925–1995) developed the concept of *rhizomatic thought* (Hillier and Abrahams 2013). According to Deleuze, as the rhizome has many branches, it is possible to link different topics from different fields and study the "hidden" relations among them. Thus, several concepts can refer to a common concept. In this way, we can link together several seemingly different fields. This is connected to the word, "quantum," from our proposed title which also refers to the discretization of rizhoma; "Gestalt" highlights the need for an organization and an identification of groups of rhizomes, and "ART" highlights the importance of this topic for a new approach towards arts. For this reason, we can coin a new expression, *Quantum GestART.*[2]

Rhizomatic systems are mainly horizontal, as opposed to the verticality of Hegelian categorization. Rhizomas can be considered not only as static objects, but as events in space and time, that can change, can be composed together, and grow. This could give an idea of a comprehensive Gestalt for these topics. Also, creativity lies in the combination of rhizomatic entities, with the creation of new connections and new arrows. In this way, new structures and new hierarchies between elements can appear.

An overall form can be seen as the result of an evolution in space and time – and this is the perspective considered in "On Growth and Form" by D'Arcy Wentworth Thompson (D'Arcy Thompson 1966; Prusinkiewicz and Hanan 1980; Prusinkiewicz and Lindenmayer 2004):

> Organic form itself is found, mathematically speaking, to be a function of time . . . We might call the form of an organism an *event in space-time*, and not merely a *configuration in space*.

Also, according to Hallé, Olderman, and Tomlinson (1978, VIII), quoted in Prusinkiewicz and Lindenmayer (2004). "The idea of the form implicitly contains also the history of such a form" (29). Considering a visual form as (the result of) a collection of events in space and time may lead to its sonification, i.e. translation into sounds (Hermann, Hunt, and Neuhoff 2011). In fact, such a sequence of visual events can be mapped into a sequence of musical events in time. If we consider a small branch (Mannone 2017), we can build it little by little in time as a temporal representation of the growth process. Thus, the sonified little branch would be a sequence of events in the auditory domain rather than the visual, which then aims to convey to the listener the idea of an organic shape that is replicated and grows. One example is a short theme which is repeated and transformed.

Reading an image as a collection of events implies its discretization. Discretization via Deleuzian rhizomas can be metaphorically compared with discreteness proposed in quantum

---

[1] Deleuze's contribution also involves diagrammatic thinking. According to Deleuze (1986, 43): "Le diagramme ne fonctionne jamais pour représenter un monde préexistant, il produit un nouveau type de réalité, un nouveau modèle de vérité," that is, "A diagram doesn't work to represent a world as a pre-existing one; it rather produces a new type of reality, a new model of truth."

[2] The term "GestART" has been used by M. Mannone in a European application (H2020) in Summer/Fall 2017, and then used in this framework.

mechanics. Thus, an original image can be thought of as a continuous form, and its decomposition onto a set of simple shapes. This idea will be described in detail in Section 3. The image discretization is provided by a superposition of rhizomatic entities. Gestalt is given by the collection of these entities – it may be metaphorically compared with their categorical colimit[3] (Mac Lane 1978). Why are we considering the concept of colimit in such a framework, even metaphorically? Intuitively, let us imagine to have several arrows starting from each one of these rhizomatic entities, and reaching a form. A form that is well recognizable, made up by rhizomatic entities as a particular building or a clear form made up of bricks. We may think of several LEGO^{TM} bricks bound up to recreate a specific, easy-to-recognize, form. And this construction is "universal" in the sense that the arrows starting from other bricks can be composed with the other arrows to reach the same form, and that composition of arrows is unique.[4]
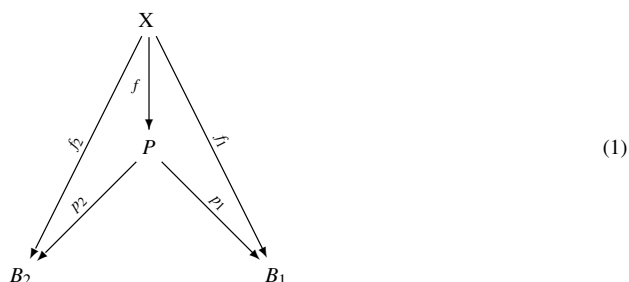
The step of the image discretization via a "base" decomposition is represented by the step F, filter, in the mathematical procedure we propose in Section 3. Then we sonify the decomposition (step T). Finally, we can refine the obtained sequences to compose a complete musical piece (step R), via an interpolation of the obtained musical sequences. In some sense, from the discrete we get back to the continuous.[5]

In this article, we summarize some ideas of recent literature about these questions. Then, we propose a theoretical development: a complex image can be filtered and approximated via the superposition of a set of schematic shapes with opportunely chosen coefficients. A minimal number of simple shapes – which we will call "visual vectors," or "visual kets" in Section 2.4 – is needed to make the image recognizable, as in Gestalt principles. In fact, we use a formalism inspired from the bra-ket Dirac notation used in quantum mechanics (Dirac 1939).

An image can be reduced to a collection of vectors with coefficients. Then, the "base" can be changed from visuals to sounds: each considered set of visual vectors is associated with a set of sound vectors, a set that is opportunely chosen according to former studies about music and sound. The new superposition of basic sounds and sound sequences gives a larger musical sequence. With an abuse of notation that we will explain later in this article, we can talk of a visual "basis" (or base) and of a sound "basis" (or base). By changing the basis, we make a transformation from the visual domain to the sound domain. Because a Fourier transform may be seen as a change of basis, we can metaphorically describe this process as a Fourier-like transform

---

[3] *Limits* and *Colimits* are concepts from category theory (Mac Lane 1978). As explained by Mannone (2018a, 12): "Intuitively, *limits* and *colimits* are generalizations of *products* and *coproducts*, respectively. The product is a special case of the limit, with discrete categories. The coproduct (also called *sum*) is the dual of the product, obtained by reversing the arrows. Given an object P and two maps $p_1 : P \to B_1$, $p_2 : P \to B_2$, P is a product of B_1, B_2 if for each object X and for each pair of arrows f_1, f_2 we have only and only one arrow $f : X \to P$ such that $f_1 = p_1 f$, $f_2 = p_2 f$ (Lawvere and Schanuel 2009), see diagram (1)."

$$
\begin{array}{ccc}
 & X & \\
f_2 \swarrow & \downarrow f & \searrow f_1 \\
 & P & \\
p_2 \swarrow & & \searrow p_1 \\
B_2 & & B_1
\end{array}
\qquad (1)
$$

[4] This "unicity" is a rough example of the "universal property" shown by some constructions in category theory, such as colimits and limits, dual of the colimits (Mac Lane 1978).

[5] The idea that a shape may contain its story can be also compared with evolution, creating an ideal connection between the evolution of natural forms and the evolution of musical forms. This fact may be contextualized in a wider framework, within the evolution of Western (artistic) thinking.

(and thus we will be using a *T* symbol) from visuals to sounds. One of the goals of this study is to compose music. For this reason, the final result of the described process can be refined by a (human) composer and used as a complete musical piece.

This strategy can also be contextualized within the framework of sonification studies (Hermann, Hunt, and Neuhoff 2011). The resulting music will have its own Gestalt, that would be the sound rendition of the original visual Gestalt (Mannone 2018b). If the final result can reasonably be brought back to the original image – or at least to an equivalence class of images that contains the original one – we can talk about some degree of *conservation* of the overall Gestalt, in a way that is inspired by conservational and invariance principles in physics. Due to the generality of this strategy, the initial image can be a form from nature, an abstract collection of points and lines on a painting, and so on.

Our study is also deeply related to *crossmodal correspondences*. This phenomenon refers to people's tendency to match information presented in one sensory modality to the information presented in another one (Spence and Parise 2012). Crossmodal correspondences are statistically based experiences, and they are considered different from *synesthesia*,[6] where the association is involuntary and more individual-dependent (Deroy and Spence 2013). In some cases, the association between essential properties of sounds and essential properties of visuals allows the identification of the so-called *audio–visual objects* (Kubovy and Schutz 2010); this seems to be relevant also in an artistic framework, where we may characterize multisensory artworks. In this paper, we will be dealing with the perceptual organization of different sensorial stimuli, that can be thought of as a visual Gestalts and an auditory Gestalt. We will refer to joint studies between Gestalt and crossmodal correspondences (Spence 2015). The analysis of visual and auditory Gestalts can also involve sounds in speaking and visual representations (see Nobile 2013 for Italian language).

The structure of the paper is the following: in Section 2, we contextualize our work within the current interdisciplinary literature. In Section 3, we present the mathematical formalism. In Section 4, we show some sketches of a pseudocode, discussing how it would be possible to implement these ideas and strategies via a machine learning application. In Section 5, we discuss future developments of the research, included the possibility of a cognitive experiment that would give concreteness to the theoretical approach we present here. In the Conclusions, we consider additional philosophical and computational developments. The method we describe in Section 3 allows for translations image-to-music; however, reversing the arrows, music-to-image is also possible. This gives a precise and formal background to former approaches to music-image translations. Images include tridimensional objects whose space coordinates are mapped into time, loudness, and pitch (Mannone 2011).

2.   Background: cognition, physics, and mathematics

2.1.   *Cognition: Gestalt and crossmodal correspondences*

In this section, we focus on examples of crossmodal correspondences between sounds and images, on Gestalt in music, and on Gestalt joined with crossmodal correspondences. According to Spence and Parise (2012):

> Crossmodal correspondences have been defined as tendencies for a certain sensory feature, or attribute, in one modality [. . .] to be matched (or associated) with a sensory feature, or attribute, in another sensory modality. (410)

---

[6] The *synesthesia* concept is familiar from poetry: let us recall the "black scream" in "[. . .] All'urlo nero della madre/che andava incontro al figlio/crocifisso sul palo del telegrafo" from the poem *Alle fronde dei salici*, in *Giorno dopo giorno* (1947) by Salvatore Quasimodo, awarded the Nobel prize for Literature in 1959.

Several examples of crossmodal correspondances have been found between vision and audition. For instance, we can cite the widely shared association between an ascending melodic line and an ascending vertical movement (Spence and Parise 2012; Zbikowski 2017). Literature describes crossmodal correspondences between visual, auditory, tactile, and even olfactory stimuli (Deroy, Crisinel, and Spence 2013). We can wonder about the existence of general or generic sensorial concepts that can be embodied in different, specific senses. Crossmodal correspondence, according to (Deroy, Crisinel, and Spence 2013, 885), "is something that occurs between *concepts*." If we extend this idea to organized stimuli, the question about the "translatability" of an artwork from a sensory field to another would be meaningful. In our framework, translations are made via category-inspired diagrams. In fact, the mathematics we use to investigate these problems, category theory, has a high abstraction power. In particular, the search for abstraction and invariants (see the end of Section 3) is also outlined in cognitive literature such as the work of Reybrouck 1997, which states, "The process of abstraction [. . .] is characterized by the extraction of invariants" (62).

The second topic of cognitive sciences relevant to our study is Gestalt theory. According to Reybrouck (1997), besides its intuitiveness for visuals, Gestalt laws often need an experimental validation to be extended in other areas. Reybrouck also points out "a fundamental isomorphism (in a metaphorical sense) between the structure of the object and the psychophysiological structure (Gestalt-like patterns of excitation in the brain) of the perceiver." This highlights a correspondence between what we can call the "main lines" characterizing the structure of an object, and the "patterns" in the brain that are activated when these main lines match with them. As we will see in Section 3, this idea[7] can support a decomposition of visual things into a superposition of base elements, and Gestalt as their metaphorical colimit.[8]

Because music happens in time, a Gestalt approach to music should be considered as a dynamic process, and music should involve a "description in terms of processes" (Reybrouck 1997) rather than a description of static structure.[9] According to Reybrouck (1997), "In contrast with a geometrical figure, that is described as a whole when looking at it, a musical figure needs a successive presentation" (64). However, if we consider a geometrical shape not – yet – as a whole, but as a "making" of it in terms of movements, we can easily connect it with a sequence of musical gestures (see Paragraph 2.3.2). If we can see images in their making, we can also do the opposite: we can try to discretize music in terms of short sequences that are perceived as a whole. According to Reybrouck (1997, 64):

> The grasping of [music's] meaning, then, is polythetic in that it can be grasped only by hearing the music as it unfolds step by step. This hampers the holding of a musical Gestalt as an immediate and directly experienced whole. Music, however, can be grasped in a monothetic way, if the discursive processes are coded as discrete things as in the case of conceptualization.

This seems consistent with the quantum approach proposed in Section 3, where simple visual units are connected with simple musical units, and groups of visual units are thus mapped into groups of musical units. Groups of visual units are meaningful entities that can be considered as a

---

[7] Curiously, Reybrouck (1997) explicitly refers to algebraic methodologies and cites the *functor* term, but it appears as more in the sense of a function than of a categorical functor. A functor is a morphism between categories (Mac Lane 1978).
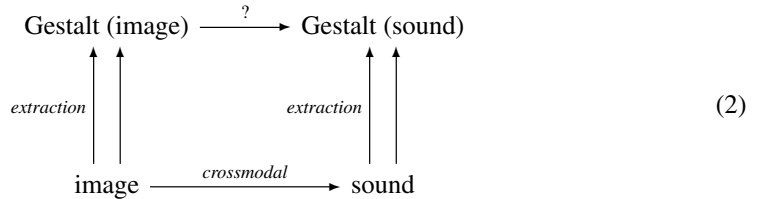
[8] Gestalt as a colimit: all simple elements are connected into a colimit. We are considering here colimits and not limits because arrows are directed from small entities to the complete, recognizable form, the Gestalt. As explained above in our example with LEGO[TM] bricks, let us consider a collection of simple shapes that can be combined together to obtain a recognizable form. Thus, the arrows we are considering are directed from the simple shapes to the 'complete' form. A reviewer thought of a colimit of 'visual things' as patterns in the brain. This is a very interesting remark that can be addressed in future research; also, there are recent studies between categories, brain, and consciousness, see Northoff, Tsuchiya, and Saigo (2019).

[9] In the approach proposed in (Mannone 2017), images are seen as the result of processes, thus the use of Gestalt from images to music in this framework appears as more coherent.

visual Gestalt. Equivalently, groups of musical units can be considered as a *sound Gestalt*. Thus, mapping groups of visual units into groups of musical units can ideally be seen as a mapping from a visual Gestalt to a sound Gestalt. We wonder if such a mapping can be possible and effective.
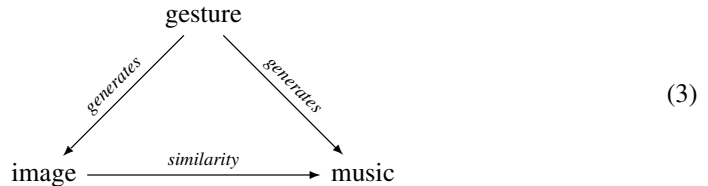
Spence (2015) summarizes recent research on Gestalt and crossmodal correspondences. He also considers open questions such as ambiguities of perception in Gestalt and their possible crossmodal equivalent. Another question about the possible influence of crossmodal perceptual grouping on intramodal grouping, that is the opposite of the first question. Experimental evidence shows that there is an influence of a grouping within a sensory area towards a crossmodal sensory grouping which involves that area and another one (Spence 2015). The first intersensory Gestalt that has been validated is the audiotactile musical meter perception.[10]

Thinking of categories, we can see crossmodal correspondences as arrows between different sensory areas – e.g. visual and auditory areas – and Gestalt as arrows that start from a collection of elements on each sensory area and reach a point (as a categorical colimit), the abstract form. We wonder if we can draw arrows, that is, if we can connect a Gestalt of a sensory area (e.g., Gestalt of image) with a Gestalt of another sensory area (e.g., Gestalt of sound), represented, respectively, as two "final" points in diagram (2), where the question mark indicates a connection not verified yet.

$$
\begin{array}{ccc}
\text{Gestalt (image)} & \xrightarrow{\quad ? \quad} & \text{Gestalt (sound)} \\
\Big\uparrow\Big\uparrow {\scriptstyle extraction} & & \Big\uparrow\Big\uparrow {\scriptstyle extraction} \\
\text{image} & \xrightarrow[\quad crossmodal \quad]{} & \text{sound}
\end{array}
\tag{2}
$$

The reference to Gestalt as an organization principle and crossmodal correspondences as comparisons between different sensorial stimuli make us think of a superior, more general, "suprasensory" organizing principles (Allen and Kolers 1981; Spence 2015), or of trans-modal Gestalts (Kubovy and Yu 2012). In particular, Kubovy and Yu (2012) consider the additivity of groups that is relevant for models with states superposition.

These ideas can be connected with a recent development of the mathematical theory of musical gestures: the concept of *gestural generator* (Mannone 2018a), see Paragraph 2.3.2, that can be in some way compared with these approaches. Intuitively, a simple drawing and a short musical sequence can be seen as *similar* if they appear as being produced (generated) by the same movement, with pencil on paper, or with hands/breath/bow on musical instruments (Mannone 2018a); see diagram (3).

$$
\begin{array}{ccc}
& \text{gesture} & \\
{\scriptstyle generates}\swarrow & & \searrow{\scriptstyle generates} \\
\text{image} & \xrightarrow[\quad similarity \quad]{} & \text{music}
\end{array}
\tag{3}
$$

This conjecture has been experimentally validated for simple shapes and simple musical sequences (Collins et al. 2018; Mannone et al. 2018). Our theoretical framework could be used to contextualize other experiments about sounds and gestures, even with non-musical sounds or sounds not derived from sonification but just noises connected with the drawing production.

---

[10] This can easily be contextualized within studies about music and gestures in terms of physical movements. If we interpret visuals in terms of gestures, our quest for Gestalt translation from visual domain to auditory domain via mathematics and gestures may appear as well-defined. Also, according to Spence (2015, 11), crossmodal correspondences "can be conceptualized in terms of crossmodal grouping by similarity."

In particular, we can refer to a recent work on synchronization of gestures with synthetic friction sounds (Thoret et al. 2014a). Diagrammatic thinking can help investigate mutual influences and biases between audition, gestures, and drawing production. This is the case of sounds variations that may induce people to draw ellipses instead of circles (Thoret et al. 2016), and the perception of biological movements via sound detection (Thoret et al. 2014b).[11] The hypothesis of a connection between different senses and media towards the representation of "higher" and sense-independent concepts may be supported by recent research in the so-called "supramodal brain" (Rosenblum, Dias, and Dorsi 2016). According to Rosenblum, Dias, and Dorsi (2016, 65), within the supramodal conception of the perceptual brain, "task, rather than sensory modality, is the primary organising principle." Also, research focused on speech shows that "supramodal speech mechanism also seems sensitive to haptic and kinaesthetic input" (Rosenblum, Dias, and Dorsi 2016, 73). This may be related to the idea of gestural generator; see Paragraph 2.3.2.

Let us also consider the possibility of translations, transfer of Gestalt from one domain to the other. About crossmodal Gestalt, we may wonder about the existence of some *uncertainty principle*; we may not be able to discern with the same precision all groupings and so on in one medium, e.g., in visuals, and, at the same time, to perfectly recognize groupings on another medium, e.g., sound, if performed simultaneously. Formally, we have the inequality of Equation (4).

$$
\begin{aligned}
&\text{(incertitude in finding groups/similarities in visuals)} \\
&\times \text{ (incertitude in finding groups/similarities in music)} \\
&\geq \text{ (constant)}
\end{aligned}
\tag{4}
$$

This means that either we hear with extreme attention or we make a visual analysis with extreme attention. An intermediate value of attention should be supported by artistic realizations such as *Optical Poem* (1938) by Oskar Fischinger. Already Gabor (1947) proposed an uncertainty principle for sound (only), with a formalism borrowed from physics. The inequality of Figure 4 is not directly inspired of non-commutativity, but of precision levels. However, we can draw a diagram of attention levels starting from diagram (2), and we can experimentally investigate its commutativity. The inequality relationship of Equation (4) implies the use of short temporal windows, equivalent to the sequences of Figure 8 for each small group. In conclusion, the relations between cognition and neural system can be presented more formally by using also a mathematical approach to the neuro-cognitive system, based on "dynamic" category theory (Ehresmann and Vanbremeersch 2007; Kainen 2009). In this model, called "Memory Evolutive Systems," developed by Andrée Ehresmann and Paul Vanbremeersch, the mental objects are colimits of diagrams that represent synchronized neural groupings. According to Kainen (2009, 228): "Its creators, Ehresmann and Vanbremeersch, have proposed a unique synthesis of various ideas and concepts, organized around the notion that life is an evolving hierarchical category with a functorial dynamic." Thus, this idea also connects cognition with natural studies. This approach led to very recent developments about mental representations investigated via categories (Northoff, Tsuchiya, and Saigo 2019).

2.2. *Physics: quantum mechanics*

References to cognition may be relevant for a mathematical theory of the arts for two reasons. First, as it happens in physics, mathematics can give more solutions than the ones that are really

---

[11] In these works, the $\frac{1}{3}$ law is cited, that is, the connection between timbre variations and velocity profiles. This reminds one of the $\frac{2}{3}$ power law describing the connection between the kinematics of handwriting and the trajectory of a movement, between the curvature of handwriting and the pentip's angular velocity (Lacquaniti, Terzuolo, and Viviani 1984; Plamondon and Guerfali 1998).

coherent with natural phenomena. Experiments are needed to verify the consistency of a theory with the physical reality. Likewise, cognitive experiments can be useful to validate specific approaches or conjectures of mathematical theories of the arts. Second, cognition-relevant topics may give new blood to theoretical advancements, that stimulate the development of new theories or the generalization of existing ones.

The parallel between physics and mathematical approaches to the arts is not only a metaphor. To investigate cognition, visuals, and sounds, there have been attempts to exploit ideas and methods from quantum mechanics, using the main concepts of quantization, destructive measure, and probability (Gabor 1947; Khrennikov 2010; Rocchesso and Mannone 2018). The step from continuity to quantization can sometimes make evaluations easier, and it can allow analysis in a new light. A renowned example is about sound itself. Acoustical waves are often used as a paradigm for quantum waves (Gabor 1947); conversely, if we use the formalism of quantum waves to describe sound, we can also borrow the duality wave-particle and introduce quanta of sound (Gabor 1947). Other examples include analysis of images (Youssry, El-Rafei, and Elramly 2015), psychology (Khrennikov 2010), musical forces and cognition (beim Graben and Blutner 2017, 2019), quantum information, cognition, and music (Dalla Chiara et al. 2015), adaptation to music of techniques to evaluate memory in quantum systems (Mannone and Compagno 2014), and vocal theory based on quantum formalism (Rocchesso and Mannone 2018), where vocal categories are used as probes to describe sounds in general.

2.3.　*Mathematics: category theory and diagrammatic thinking*

Another topic often applied to music is a mathematical theory: category theory. A category is constituted by objects and morphisms between them, verifying associative and identity properties (Mac Lane 1978). This is a general theory that starts with the very basic concepts of objects (visually represented as points), transformations between them (visually represented as arrows), and transformations between transformations (arrows between arrows). We can see an entire category as a point, and we can draw arrows between categories (called *functors*), creating nested structures. Finally, a mathematical tool we will use is a Fourier-inspired transform, which will be defined in Section 3. The use of Fourier transforms not only for sound synthesis but also for music analysis has been widely exploited in music theory, to characterize rhythmic or harmonic distributions (Amiot 2016).

Categories play an important role in attempts to unify different branches of mathematics itself (Caramello 2018). Also, they have been used to investigate human cognition (Phillips and Wilson 2010), neuroscience (Ehresmann and Gomez-Ramirez 2016), aesthetics, art production, and art evaluation processes (Kubota et al. 2017),[12] the aesthetics of mathematics (Mannone 2018b), and physics (Baez and Aaron 2011). Categories have also been applied to music (Jedrzejewski 2006, 2019; Fiore and Noll 2011; Popoff, Andreatta, and Ehresmann 2015; Mannone 2018a; Mazzola 2018). We can describe, with categories, musical structures as elements and transformations between them, musical gestures (Mazzola and Andreatta 2010; Arias 2018), recursive musical gestures (Mannone 2018c), and even musical instruments (Mannone and Favali 2019), and music and dance (Mannone and Turchet 2019). In recent literature, there have been some pioneering connections between quantum mechanics and diagrammatic thinking (Coecke 2006).

Category theory is the general framework of recent studies we are starting from. However, in this paper we will not make use of precise categories; we will rather consider a more generic diagrammatic thinking as a visual aid to describe transformations and transformations between transformations.

---

[12] Curiously, this paper has been submitted on the same day as a paper on categories and composer/conductor/listener, see Mannone (2018a).

2.3.1. *Category theory and nature*

The categorical concept of "transformation between transformations" can be applied to shapes from visual arts, and even to shapes and forms from nature. For example, we can compare the growth of a fish with the growth of another fish via an arrow between arrows (Mannone 2019). Categories in biology have already been applied to nature in terms of input-output processes (Letelier et al. 2006). Other relevant mathematical approaches to shapes from nature are the classic studies by D'Arcy Wentworth Thompson (D'Arcy Thompson 1966) that investigate the geometry of nature, and the studies of Aristid Lindenmayer (Prusinkiewicz and Hanan 1980), that investigate the growth of plants. These studies are relevant in particular for computer graphics applications. D'Arcy W. Thompson and Lindenmayer's studies do not involve category theory, but they can be re-read in its light (Mannone 2019). According to Reybrouck (1997), references to *processes* and *growth*, jointly with the concept of *form*, suggest an immediate connection between Gestalt and forms and growth of natural forms (D'Arcy Thompson 1966). The categorical formalism has been applied to processes in biology (not to morphology) by Rosen (1958), and extended to the emergence of cognition from neurons by Ehresmann and Vanbremeersch (2007). Finally, classification can also be investigated in light of categories, by using the concept of colimit to connect species, genera, and families with the composition of arrows leading to the same "essential idea" (Mannone 2019).

2.3.2. *Gestural generator*

We can intuitively define a "gestural generator" via its action as we will see. The concept can be applied within music, and between music and visuals. Within music, let us consider two performance gestures: a movement to play a note with an unspecified dynamics (loudness level) on a piano, and the movement to play a note with an unspecified dynamics on a vibraphone, for example. To play a note with dynamics *forte*, we have to "deform" these movements. The two *forte* movements, on the piano and on the vibraphone, will then be *similar* with respect to the loudness. *Similar* changes in piano and violin loudness (and timbre) are provoked by a *similar gestural generator*, that operates a similar change in the respective gestures – e.g., acceleration increase for hands/hammers, or of airstream pressure for a wind instrument.[13] Of course, all the changes must be made in the appropriate spaces. Striking a flute with a hammer will not produce a *forte* flute sound: instead, it will just break the flute! A formal definition of gestural similarity is given by Mannone (2018a). We report the Heuristic Conjecture about gestural similarity of musical gestures (Mannone 2018a, 22):

CONJECTURE 2.1 *Two gestures, based on the same skeleton,[14] are similar if and only if they can be connected via a transformation:*

(1) that homotopically transforms one gesture into the other, and
(2) that also leads to similar changes in their respective acoustical spectra.

---

[13] Diagram (5) from Mannone (2018a), shows a mapping from an abstract collection of points and arrows (Δ) to the topological space $\vec{X}$ (containing curves in the X, the space of points in physical space and time). In the diagram, g_ϕ represent the physical gesture (φ) without any specified dynamics, while g_ϕ^F is the gesture with the dynamics *forte* (F). The double arrow F represent the deformation of the unspecified gesture into the "forte" gesture. This is a specific *forte* operator. We can think of the *"forte" gestural generator* as a generic "forte" gesture, to be adapted in the specific space of gestures for each musical instrument that deforms anonymous movements into movements that produce *forte* loudness on musical instruments.

[14] A "skeleton" is an abstract structure of points and arrows; see Δ in diagram (5) as an example. Such a generic structure is then specified into a collection of points and arrows in a topological space. In our case, we consider physical space and time.

Homotopy is a necessary, but not sufficient, condition to get similar gestures.

While dealing with music and images, we can think of a gesture that can either generate an image or a drawing: a staccato sequence is similar to a collection of points, or a legato phrase to a continuous line, and so on. Thus, pairs of images and sounds can share some degree of *gestural similarity* if they appear as being generated by the same movement. See diagram (3) and its development in Section 3, with diagram (19). This approach can help investigate interactions between gestures, images – and in particular, drawing – and sounds – and in particular, music. We can re-read in this light the pioneering work by Alexandre Truslit (Repp 1992) including the comparison of landscapes and melodic profiles of folk songs, and the focus on "inner motion" and movement in music analysis. This analysis can be extended to contextualize experimental studies about gesture production as responses to auditory stimuli and in particular music (Küssner and Caramiaux 2015) and the influence of vision of gestures in music perception (Schutz and Lipscomb 2007).

$$\Delta \quad \overset{g_\phi}{\underset{g_\phi^F}{\Big\downarrow F}} \quad \vec{X} \tag{5}$$

2.4. *The proposed method*

Mathematical formalism can thus be applied to music, images, and nature, and it can even help connect all these areas together. For example, we can first analyze a visual shape, derive a mathematical model, define a mapping strategy, and apply the strategy to make music. Visual shapes do not intrinsically have a temporal dimension. However, we can imagine creating an image through a drawing gesture, and we can see the main lines and points of an image as a sequence of events (sounds) in time. A form that changes through time can be represented as a sequence of changing melodies, as the sonification of a series of pictures (as in chronophotography).

In this paper, we propose to join hints from quantum formalism, diagrammatic thinking, cognition, and current research in mathematical music theory (Mannone 2018a), as well as some insights from Gestalt theory applied to music and visual arts.

Also, Gestalt theory is useful when we look for the minimal decomposition: we can decompose a complex image in terms of simple shapes (points, lines, geometric figures), that we may call here "basis vectors." However, this has a very specific meaning in algebra, being strictly connected with the concept of space dimensions (e.g., in a 2-dimensional space, a basis should contain two linearly independent vectors).[15] However, in the method proposed in this article, we do not need a number of basis vectors equal to the dimension of the space: the number of vectors is instead connected with the precision of the approximation of the given form, and it depends on the complexity of the original image. Thus, we may need another name besides "basis vectors." Because of the reference to Dirac notation, we can speak of "visual kets" while dealing with images, and of "sound kets" while dealing with sounds. We would like to convey the idea of minimal decomposition of a form in terms of simple shapes, that is, in terms of some visual

---

[15] The visual kets of a set can be considered as linearly independent if none of them can be obtained as a linear combination of the other ones; similarly for the sound kets. We can consider two kets $|\phi_i^v\rangle$, $|\phi_j^v\rangle$ as orthogonal if their (gestural) similarity degree is zero (Mannone 2018a), and thus their scalar product is zero: $\langle\phi_i^v|\phi_j^v\rangle = 0$. However, if they are identical, $\langle\phi_i^v|\phi_j^v\rangle = 1$. Two kets with a similarity degree strictly between 0 and 1 have a scalar product between 0 and 1. Similarly for the sound kets. Thus, the scalar product assumes the meaning of a comparison between simple shapes or between simple musical sequences.

kets or of a few combinations of them. We need a minimal number of visual kets to make an image recognizable; and in the definition of some minimum principle, Gestalt theory intervenes. Also, we can define a basis of sounds/sound sequences. Each sound sequence should correspond to a basic image via an intuitive sonification: we can imagine both visuals and sounds as being generated by the same gestural generator. This association is an application of the definition of gestural similarity (Mannone 2018a) and has connections with recent studies in musicology, about the connection between movement and sound (Zbikowski 2017).

The validation of the proposed development requires cognitive experiments to establish what a better basis for images and sounds can be, as well as if, and to what extent, Gestalt can be preserved upon a basis change. "Change of basis" means the transition from visual to sound kets (and vice versa). Formally, we indicate the correspondence between each element of the considered set of visual kets (what can be called a "visual basis" with an abuse of notation) with each element of the considered set of sound kets (what can be imprecisely called a "sound basis"), and each coefficient of the visual kets with each coefficient of the sound kets. In this way, the decomposition of an image (meant as a drawing or a painting, or even a sculpture) in terms of visual kets and visual coefficients, can be translated into a decomposition of sound kets and sound coefficients, constituting a rough musical piece. Such an approximative musical piece can be refined to obtain a real and complete musical composition. In fact, with "basis change" we indicate the transition from the domain of visuals to the domain of sounds. We were inspired by the fact that the Fourier transform from the time domain of a signal into the frequency domain can be seen as a basis change. In the following, we will be using the term "transform" just to indicate this metaphorical analogy.

In Section 3, we describe the mathematical formalism, which joins diagrams from category theory with bra-ket notation typically used in quantum mechanics.

3. Bases, filters, and transforms

Let us give some mathematical details about the described procedure, starting with an intuitive description. The rough idea is the following:

$$\text{musical work} = T(\text{visual work}),$$

where T is a transform, which we will define soon. With "musical artwork," we can just indicate a musical piece, while with "visual artwork" we mean a drawing, a painting, or a sculpture, that can be translated into sound via T.

Formally, we can define a vector space for visuals, a vector space for sounds, and a transition between them. An image can be decomposed into a selection of simple visual shapes, and such a (visual) decomposition can be translated into a sound decomposition, living in the sound vector space. Such a sound decomposition constitutes a rough musical composition.

Let V be the space of visuals (the space of visual shapes), and let K be a field. Given $v_1, v_2, v \in V$ and $\lambda \in K$, we have: $V \times V \to V : (v_1, v_2) \mapsto v_1 + v_2$, and $K \times V \to V : (\lambda, v) \mapsto \lambda V$. Similarly, let S be the space of sounds (in the following examples, the space of musical sequences), and K' another field. The visual vector space is given by the set of visual elements V, an operation of "sum," $+$, assumed to be linear, and an operation of scalar product, $\cdot$. Let u, v be two elements of V and λ a scalar of K; we have $+_V : (u, v) \to u + v \in V$, and $\cdot_V : (\lambda, u) \to \lambda \cdot u \in V$. Similarly, the sound vector space is given by a set of sound elements – here, short musical sequences – S, an operation of sum $+_S$ and an operation of multiplication by scalars of K' indicated with $\cdot_S$. In this framework, the meaning of the addition is to superpose several simple shapes or several musical sequences within their respective spaces; the meaning of the multiplication by a scalar

is to refine resizing and space positioning of visual or sound kets (the latter, in the space of time and sound parameters).

In fact, the elements of the set of visuals are visual shapes, simple or complex. Let us consider original images ($|\psi^v\rangle \in V$) and visual kets as the elements living in the space V (we can substitute $v_1, v_2, v \in V$ with $|\phi_1^v\rangle, |\phi_2^v\rangle, |\phi_3^v\rangle$, and we also have $|\psi^v\rangle \in V$), and the set of sound kets and complete musical pieces in S (that is, $|\phi_1^s\rangle, |\phi_2^s\rangle, |\phi_3^s\rangle, |\psi^s\rangle \in S$, where the last one is a complete musical piece).[16] The (visual) basis elements (that is, a selection of visual kets) are simple shapes, or, better, elementary shapes such as points, segments, and curved segments. Other simple shapes such as triangles and circles[17] can be obtained via a superposition of elementary shapes; e.g., a circle can be obtained with two essential lines: a pair of inverted curved segments, that is, two arches. For simplicity, when needed, in some points of the paper we will indicate as "simple shapes" also these simple superpositions of elementary shapes. Above, we already indicated the (essential) shapes with the visual kets $|\phi_i^v\rangle$. Similarly, we indicated with $|\phi_i^s\rangle$ the sound kets: elementary melodic profiles and articulations. In analogy with Euclidean vectors, visual (and sound) vectors have a magnitude (that we can assume unitary at first) and a direction (that is, in our research, the direction of time). In fact, musical sequences have a direction in time; simple visual shapes have a direction given by the direction of drawing; we can thus see visuals as events in space and time, analogously to what had been discussed about forms from nature (D'Arcy Thompson 1966; Hallé, Olderman, and Tomlinson 1978). The (initially unitary) magnitude and the position in the space of these vectors is modified by their coefficients. In the space of visuals, the coefficients – indicated here with b_i – give information about size and position of the ith basis elements. We can also think of "rich" coefficients telling us also how many kets of the ith type are used. For example, in Figure 2, the image of the palm is simplified via the superposition of a segment and of several curved segments, corresponding to two kets: a segment and an arch. Coefficients b_i, thus, tell us how many kets are needed and where, in which points of the space, they have to be put to filter the original image in a recognizable way. Similarly, sound coefficients – indicated here with a_i – tell us where (in time, thus, with which onset), for how long (with which duration), how loud, and how high in pitch the basic musical sequences have to be put. If we imagine a space of musical parameters with time, loudness, and pitch (Mannone 2011), coefficients a_i are positioning and resizing sound kets within that space. If we think of a generalized musical space with more parameters, we can dare to make more associations, for example, between colors in V and timbre in S, or contour details in V and attack modes in S. For example, a vertical segment in the V space may correspond to a cluster in the S space, an horizontal segment in V to a long, single note in S, and an arch in V to a rising–lowering melody in S. How long the cluster or the melody has to last, and which precise pitches have to be played, with what loudness, is told by sound coefficients. Even a single point has a meaning in V, and it can be translated as S with a *staccato* single sound. The definition of "good" visual and sound kets can involve experiments, see Section 3.2.

About the use of kets on a real field K (while in quantum mechanics K is complex), see Sparavigna and Marazzato (2016).[18]

---

[16] Formally, kets would require an Hilbert space, but here the imaginary component is zero or it is not considered. It could be identified with the mind component, that is embodied into the real part, that is, an artwork, during the artistic creation process (Mazzola 2018).

[17] In $\mathbb{R}^n$ a circle is not a vector but a set of such vectors.

[18] Sparavigna and Marazzato (2016) consider kets as living in a subset of the vector space that can be of interest also for our study, both for the space of visuals and the space of sounds. It is required that: there is a zero vector $|0\rangle$; given $|a\rangle$ and $|b\rangle$ their sum $|a\rangle + |b\rangle$ is in the subset; the product by a scalar is also in the subset; for each $|a\rangle$ in the subset there is the opposite element $-|a\rangle$. In our study, the sum can be the superposition of shapes or musical sequences, and the zero vector the absence of a visual form (i.e. the empty visual space) and of musical sequences (i.e. silence). The opposite $-|a\rangle$ can be the time-reverse of an image $|a\rangle$ – that is, imagine starting with the finished simple drawing and going back to the empty space, or playing backward in time and going back to the silence. The operation of time reversing can be

As a general remark, even if we are speaking of basis elements, there is no specific relation – at least in this starting model – with the dimension of the space. In fact, the more simple the original image ψ^v, the less visual kets ϕ_i^y are needed to unambiguously represent its Gestalt. Thus, the number of kets is connected with what we can call "Gestalt-precision." See Figure 7 for an explanation.

Let us use the Greek letter ψ^v to indicate the original image in V, and ψ^s to indicate a complete, well-refined, musical piece. The basis elements are, however, indicated with ϕ_i^y, ϕ_i^s, where i indicates the ith element in the considered decomposition. We are not using capital letters Ψ or Φ here.

Let us start from the concept of superposition to develop a basic formalism, and from a complete image, e.g., a drawing or a picture.

First, the initial image (the "visual work") can be approximated via a superposition of (the least number of) simple shapes creating a sketch, such as a skeleton, and an overall envelope. Such a superposition makes the original shape recognizable, see Equation (6), where $|\phi^v\rangle$ is the original, initial image, b_i the coefficients for the image (visuals), and $|\phi_i^v\rangle$ the vectors for the visuals.

$$|\psi^v\rangle \sim \sum_i b_i|\phi_i^v\rangle. \tag{6}$$

Mapping from visuals to sounds is made via a basis change, see Equation (7) for coefficients and vectors (a_i are the coefficients for the sound and $|\phi_i^s\rangle$ the vectors for the sound). We use index v for visuals, and s for sounds.

$$a_i|\phi_i^{sound}\rangle \iff b_i|\phi_i^{visual}\rangle. \tag{7}$$

Such a correspondence is used to evaluate transform T in Equation (8). The correspondence between vectors for visuals and vectors for sound is established *a priori* (see gestural similarity for the ideas behind), while the value of coefficients and the number of basis terms involved differs case by case. In Equation (8), $|\psi^v\rangle$ indicates the original image, and $|\phi_i^s\rangle$ the sound kets, obtained via a mapping from the visual kets $|\phi_i^v\rangle$ used to filter $|\psi^v\rangle$. We might assume that T is linear: $T(b_i|\phi_i^v\rangle + b_j|\phi_j^v\rangle) = T(b_i|\phi_i^v\rangle) + T(b_j|\phi_j^v\rangle)$. However, $T(b_i|\phi_i^v\rangle) = a_i|\phi_i^s\rangle$, while linearity would require $T(b_i|\phi_i^v\rangle) = b_iT|\phi_i^v\rangle = b_i|\phi_i^s\rangle \neq a_i|\phi_i^s\rangle$. Thus, we can talk of *semilinearity*[19] of T. Also, we can suppose that T can be represented by a diagonal matrix: $T_{ij} \neq 0 \iff i = j$. In this way, we have $T(b_i|\phi_i^v\rangle) + T(b_j|\phi_j^v\rangle) = a_i|\phi_i^s\rangle + a_j|\phi_j^s\rangle$, and we can keep the correspondence between each visual ket with each sound ket. Of course, the condition of diagonality is a strong simplification, that can be useful for a first implementation.

$$|\psi^s\rangle \sim \sum_i a_i|\phi_i^s\rangle = T\left(\sum_i b_i|\phi_i^v\rangle\right) \sim T(|\psi^v\rangle). \tag{8}$$

---

seen as some deleting operation, or, more interestingly, as just a time direction change: playing a melody from the latest to the first note, or drawing a simple shape from the last point to the first one.

[19] Given two vector spaces V, S over fields K, K' respectively, $f : V \to S$ is a semilinear map if there exists a homomorphism $\sigma : K \to K'$ such that, for all $u, v \in V$ and $\lambda \in K$, we have $f(u, v) = f(u) + f(v)$ and $f(\lambda u) = \sigma(\lambda)f(u)$.
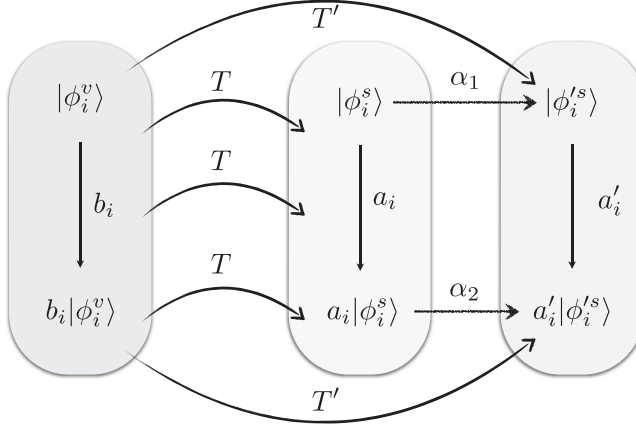
Figure 1. Transform T connects simplified images with sonified simplified images. A different sonification strategy, represented by T', leads to different sonified images. The comparison between the action of T and the action of T' is indicated by transformations α_1, α_2. The visual objects are the simple images indicated as kets $|\phi_i^v\rangle$, and their coefficients b_i "center" each simple image in a point of the tridimensional space with a given size. Similarly, sound coefficients a_i tell us in which point of sound space a sound ket $|\phi_i^s\rangle$ has to be put – i.e., which time, loudness, duration we shall give it. Even if we are using diagrams and a category-inspired way to represent ideas, we are not precisely using a categorical formalism here: in fact, abstract kets $|\phi_i^v\rangle$, $|\phi_i^s\rangle$ are ontologically different from the same kets put in specific points of visual or sound space respectively, on a specific time with a given size or duration and loudness, that is, kets with coefficients: $b_i|\phi_i^v\rangle$, $a_i|\phi_i^s\rangle$.

Diagram (9), a commutative diagram, contextualizes the whole procedure in a category-inspired framework.

$$\sum_i b_i|\phi_i^v\rangle \xrightarrow{\sim} |\psi^v\rangle$$

$$T\Big(\sum_i b_i|\phi_i^v\rangle\Big) \xrightarrow{\sim} T(|\psi^v\rangle) \tag{9}$$

$$=\qquad\qquad =$$

$$\sum_i a_i|\phi_i^s\rangle \xrightarrow{\sim} |\psi^s\rangle.$$

Let us consider visuals and sounds (Figure 1). To each visual object (a simple shape as shown in Figure 8), we associate a coefficient. The superposition of simple visual shapes with their coefficients constitutes the filtered visual images. Visual coefficients tell us *where* an image is within a space and which size it has.[20] Thus, the objects represent *which element* (e.g. a circle, a point, a straight line) we need, and the arrows tell us *where, how big*, and possibly also *how many* simple shapes of a specific kind we need (e.g., where, how big, and how many circles we need). Let us now consider sounds, where the objects are the simple musical sequences (obtained as sonified filtered images). To each simple musical sequence, we can associate a coefficient. Sound coefficients tell us *when* a musical sequence is played, how loud/long it is. Coefficients give a

---

[20] We could also talk of 'dimension' here, however, specifying that this is about the size of each ket in the space, not about the number of elements in a basis.

structure to the set of images or sounds, giving them an organization. Each (visual) coefficient represents the action of taking an image from the set of all kets and putting it into a point of the space with a certain size. Thus, visual coefficients, whose action can be indicated via an arrow, act as positioning and resizing. Similarly, auditive arrows act as positioning in time and resizing in loudness. We have a correspondence between coefficients about the position in tridimensional space and size, and coefficients about position in temporal space and loudness. The transformation (indicated here as the transform T) from the domain of visuals to the domain of sounds associates images with musical sequences, and visual coefficients with sound coefficients. If we consider two different sonification strategies, say T and T', for example, based on slightly different choices of vectors, we can define transformations from musical pieces obtained via T and musical pieces obtained via T'. This is equivalent to comparing two musical pieces, one obtained via a given technique, and the other obtained via another one; see Figure 1.

While decomposing the original image, we can distinguish between a pattern (e.g. if we consider a picture of a fish, the fish's "skeleton") and an envelope (e.g. the overall shape of the considered fish). The envelope is represented by the $0, \ldots, j$th coefficients, and the pattern by the other $j + 1, \ldots, N$th coefficients. If we start from 0, formally the last index should be $N - 1$ for the Nth coefficient. We have: $|\phi_i^v\rangle = \{(|\phi_1^v\rangle, |\phi_2^v\rangle, \ldots, |\phi_j^v\rangle), (|\phi_{(j+1)}^v\rangle, |\phi_{(j+2)}^v\rangle, \ldots, |\phi_N^v\rangle)\}$. The same applies to sound: $|\phi_i^s\rangle = \{(|\phi_1^v\rangle, |\phi_2^s\rangle, \ldots, |\phi_j^s\rangle), (|\phi_{(j+1)}^s\rangle, |\phi_{(j+2)}^s\rangle, \ldots, |\phi_N^s\rangle)\}$. If we consider together the two notations, we can run a double index: $|\phi_i^\mu\rangle$, where $\mu = 1$ stands for v (visuals), and $\mu = 2$ stands for s (sounds).

After having defined our bases, we can talk about the measurement idea. In quantum mechanics, a measurement of a state implies a destruction of part of the initial information (destructive measure). The reason is that a quantum state, described by a wave function, is usually a superposition of eigenstates, that are reduced to a single state after the measurement. The information that is preserved depends on the measurement apparatus. Intuitively, let us consider two states, with spin-up and spin-down respectively, each one with a probability density of $\frac{1}{\sqrt{2}}$, and thus, probability $\frac{1}{2}$. If the experimental apparatus is prepared to detect a specific spin, say, spin-down, after the measurement the spin-up component will be deleted, and the wave function will *collapse* to spin-down. Thus, after the measurement the state is only constituted by a spin-down with probability 1, and all future measurements with the same measurement apparatus will give spin-down with probability 1. A practical application of this idea is shown by the Stern–Gerlach experiment (Cohen-Tannoudji, Diu, and Laloë 2005).

Here, we are not dealing with spin but we can borrow the essential idea about quantum measurement. In the framework we proposed, our ideal "measurement instrument" is constituted of a selection of visual (and sound) kets. For example, if we have the image of a fish and a few kets' superpositions giving an ellipse, a triangle, and a circle, we will have lost all the information about the original fish image except the contour that is schematized via an ellipse, the tail with a triangle, and the shape of the eye via a circle, and maybe the fins with some more triangles. In this sense, the image has been filtered via the an ellipse, some triangles, and a circle. If we make other measurements of the same filtered fish, we will get the same image, because it has already been filtered via that filter. The circle, the ellipse, and the triangle(s) will have a specific coefficient that describe in which spatial point the simple shape is put with which size and so on. For this reason, measuring in this framework is like a filtering process; more precisely, expressing something in terms of (a finite number of) basis coefficients. If we want to sonify our filtered fish, we have first to use a "translation dictionary" (that verifies gestural similarity condition) to see how to sonify each of the simple shapes used, and then we have to superimpose these musical sequences. A gestural similarity condition also applies to the position and size of each one of the selected visual kets. Summarizing, we measure an image, and then we take its transform into sound.

The first rough approximations of the original shape are just in terms of essential lines. This is the envelope part:

$$T\left(\sum_{i=0}^{j}|\phi_i^v\rangle\right) = \sum_{i=0}^{j}|\phi_i^s\rangle \tag{10}$$

However, if we map only the patterns, will just have the terms,

$$T\left(\sum_{i=j+1}^{N}|\phi_i^v\rangle\right) = \sum_{i=j+1}^{N}|\phi_i^s\rangle \tag{11}$$

This requires that matrix T has some block form, so we exclude index permutations. We can get two partial sonifications: a sonification of the envelope and a sonification of the pattern. The complete musical rendition would require summing up (superposing) these terms:

$$T\left(\sum_{i=j+1}^{N}|\phi_i^v\rangle\right) + T\left(\sum_{i=0}^{j}|\phi_i^v\rangle\right). \tag{12}$$

$$|\psi^v\rangle \sim \sum_{i=0}^{N}b_i|\phi_i^v\rangle = \sum_{i=0}^{j}b_i|\phi_i^v\rangle + \sum_{i=j+1}^{N}b_i|\phi_i^v\rangle \Rightarrow T\left(\sum_{i=0}^{j}b_i|\phi_i^v\rangle\right) + T\left(\sum_{i=j+1}^{N}b_i|\phi_i^v\rangle\right)$$

$$= T\left(\sum_{i=0}^{j}b_i|\phi_i^v\rangle + \sum_{i=j+1}^{N}b_i|\phi_i^v\rangle\right) = T\left(\sum_{i=0}^{N}b_i|\phi_i^v\rangle\right) \sim T(|\psi^v\rangle) \sim |\psi^s\rangle. \tag{13}$$

Thus: $T|\psi^v\rangle \sim |\psi^s\rangle$, and $T : [\mathbf{v}] \to [\mathbf{s}]$, where the new notation with square brackets is meant to highlight the basis change. While the original image can be seen as the product of envelope and pattern, the final result after filtering and base decomposition (that is, our selection of visual kets to decompose a given image) may be seen as the *sum* of coefficients, as terms within a matrix. The sum is direct because there is no overlap between the meaning expressed by coefficients. Thus, the external product between overall shape and pattern is transformed into a direct sum: (overall shape $\otimes$ patterns) $\Rightarrow$ ((0 $\to$ j coefficients) $\oplus$ (($j + 1$) $\to$ N coefficients)). The external product becomes a direct sum of visual kets combined via suitable coefficients. In fact, we can represent the basis as follows: $|\phi_0^v\rangle, \ldots, |\phi_j^v\rangle \oplus |\phi_{j+1}^v\rangle, \ldots |\phi_N^v\rangle$ This basis can be represented as a matrix. The sonification operator T acts on coefficients from 0 to j and from $j + 1$ to N, getting $|\phi_0^s\rangle, \ldots, |\phi_j^s\rangle \oplus |\phi_{j+1}^s\rangle, \ldots, |\phi_N^s\rangle$.

Summarizing, we can decompose the initial image via a basis of visual vectors: $|\psi^v\rangle \sim \sum_i^N b_i|\phi_i^v\rangle$, and then, we can map this decomposition into sounds: $T(F|\psi^v\rangle) = T\left(\sum_i^N b_i|\phi_i^v\rangle\right) = \sum_i^N a_i|\phi_i^s\rangle \sim |\psi^s\rangle$, provided that each element of the visual basis corresponds to one element of the sound basis.

If the final music is refined (operator R; i.e., well-orchestrated, with motives and chords completed, and so on), we have: $R(T(F|\psi^v\rangle)) = |\psi^s\rangle$, with $|\psi^s\rangle$ a complete, ready-to-perform piece of music. Thus, we have Equation (14); see on the right a diagrammatic rendition of the same equation. Summarizing, a possible protocol to make a complete musical piece out of an image is filter–transform–refine, or F-T-R. Equation (14) contains the operators in inverse order because
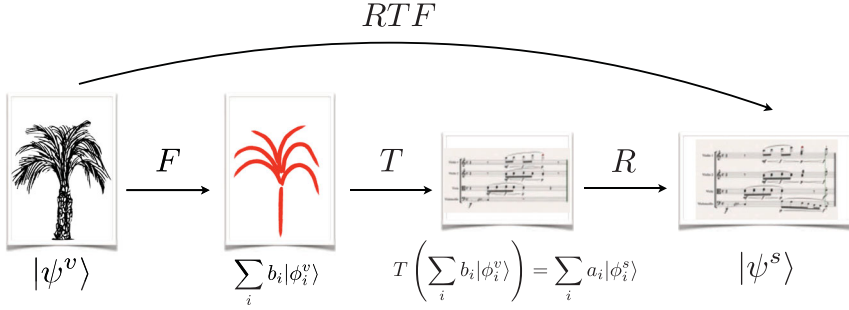
Figure 2. The scheme RTF with the possible filtering (F), transform (T), and refinement (R) from a palm (*Butia capitata*) image. The diagram also shows the composition of the three operations.

they are applied from right to left, according to the (mathematical) composition convention also used in category theory while composing arrows. See also Figure 2.

$$|\psi^s\rangle = RTF|\psi^v\rangle, \quad |\psi^v\rangle \xrightarrow{RTF} |\psi^s\rangle. \tag{14}$$

The refinement action is not unique; it depends on a choice – likewise the basis R maps the superposition of sound vector basis within a complete piece. If we consider all possible (reasonable) choices for R, we grasp an intuitive idea of the equivalence class of musical pieces that can be composed via a refinement of the same starting material. Differences between these pieces are less evident if we neglect R and we just consider several, possible bases for the first step, visual decomposition. These pieces are the sonification(s) *up to an isomorphism* of the initial shape. In fact, if we evaluate the distance between different sonification/musical renditions of the same simplified shape, we expect to find a value η, Equation (15), that is different from the distance between the sonification of a shape and the sonification a completely different shape. This leaves room for experimental investigations of shape differences and values of musical distances. These last ones can be evaluated through matrix distance evaluations, as done in Mannone and Compagno (2014). Equation (15) shows the information we would like to obtain from a suitable distance criterion, where α and β are two visual decompositions. These decompositions are *minimal* for Gestalt: in both cases there is only the minimum number of simple shapes that, combined together, make the image recognizable. The quantity ϵ is a positive, real number. If we apply T, we get the criterion for musical distances, see the second part of Equation (15), with η a real and positive number (we do not use δ here to avoid confusion with the Kronecker delta).

$$\mathcal{D}\left(\sum_i b_i^\alpha |\phi_i^{\alpha,v}\rangle, \sum_i b_i^\beta |\phi_i^{\beta,v}\rangle\right) < \epsilon, \quad \mathcal{D}\left(T\left(\sum_i b_i^\alpha |\phi_i^{\alpha,v}\rangle\right), T\left(\sum_i b_i^\beta |\phi_i^{\beta,v}\rangle\right)\right) < \eta. \tag{15}$$

The distance $\mathcal{D}$ defined in Equation (15) can be formulated in terms of cosine similarity, drawing a parallel with studies on (musical) tonal attraction via quantum formalism (beim Graben and Blutner 2017). With cosine similarity, the distance $\mathcal{D}$ is 0 if the two basis decompositions (and their sonifications) are completely unrelated, and 1 when they are identical. In this framework, ϵ and η should be positive real numbers much smaller than 1.

Equation (15) may suggest that T is a continuous function. However, the general idea is that different, but equivalent, base decompositions of an image should have a small distance between them, and their transform T should keep such a distance small. This means that, within the variety, the listener should be able to connect equivalent filtered images with equivalent sonifications, where equivalence is evaluated within the limits of ϵ and η. The value of ϵ is characteristic of the given original image, and it can be evaluated experimentally via the cognitive experiment

described in the following section. Ambiguity leads to higher values of ϵ: that is, if two initial images are quite close or they show some visual ambiguities, so will their decompositions.

Coefficients a_i, b_i indicate where the corresponding sound or visual ket should be put. In fact, the same kets can create different images if recombined in different positions or with different relative size. To simplify our analysis here, however, we will just consider a_i and b_i as scalars, and thus, as real numbers.

$$\langle \phi_i^v | \phi_j^v \rangle = \delta_{i,j}, \tag{16}$$

where $\delta_{i,j}$ is a Kronecker delta, with values 0 or 1. With vectors of different basis:

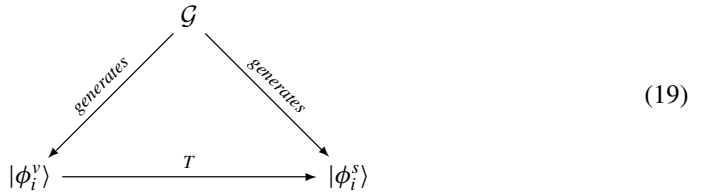$$\langle \phi_i^v | \phi_j^{'v} \rangle = \epsilon_{i,j}, \tag{17}$$

where $\epsilon_{i,j}$ is a positive number between 0 and 1 that indicates the distance between corresponding elements of the two bases. Thus, $\delta_{i,j} \in \{0, 1\}$ and $\epsilon_{i,j} \in [0, 1]$. For the same initial image, two different but *equivalent* decompositions will either have the same visual kets with slightly different coefficients, or also different visual kets but with a negligible difference. In fact, this $\epsilon_{i,j}$ is a local version of the overall ϵ as described above, that is supposed to be much smaller than the distance between the decompositions of different images. The same applies to sonification, and thus to the selected kets with index s. If we compare a selection of (sonified) visual kets, that is, sound kets of a sonified visual basis and sound kets of another sound basis:

$$(T^* \langle \phi_i^v |) | \phi_j^{'s} \rangle = \langle \phi_i^s | \phi_j^s \rangle = \eta_{i,j}, \tag{18}$$

where T^* is the adjoint of T acting on the bra vector. A comparison between the original image and the filtered image will give an idea of the accuracy of filtering, and of the amount of information we can ignore while still leaving the original shape easy to recognize. In the same way, once we have a refined musical piece, we can compare it with the superposition of sound kets, and we can evaluate the amount of personalization put in by the composer, and what are the notes and musical sequences he or she probably started from.

A filtering operation, represented by operator F, can be compared with a quantum measure, that is a destructive measure. This would be one more reason to justify the use of a quantum-like formalism adopted here.

The gestural similarity conjecture (see Section 1) applies to the mapping of each element of the visual basis to each element of the sound basis, see diagram (19), where $\mathcal{G}$ is the common gestural generator that produces an element of the visual basis and its corresponding element of the sound basis, and $\forall i = 0, \ldots, N$.

$$
\begin{array}{ccc}
 & \mathcal{G} & \\
 & \text{generates} \quad \text{generates} & \\
|\phi_i^v\rangle & \xrightarrow{\quad T \quad} & |\phi_i^s\rangle
\end{array}
\tag{19}
$$

This approach can include temporal evolution. In music, $|\phi_i^s\rangle$ already contains time: in fact, each simple shape is seen as a collection of sound events in time. Thus, there is a time where each sound sequence/melody/superposition of short melodies lives. If the initial shape is changing over time, we will have a collection of sound sequences, each one for each time instant. To make sonification practical and possible, we can consider snapshots of the changing image at given time intervals. Thus, we can obtain a finite sequence of sound sequences per shot. This would correspond to the sonification of a chronophotography of the changing original image. Thus, $|\phi_i^s(t)\rangle$ is the sound sequence at each step t, where t indicates when the shot has been taken. This

means that the transformation of $|\phi_i^v(0)\rangle \rightarrow |\phi_i^v(T)\rangle$ of the image, from time $t = 0$ to time $t = T$, correspond to a sound transformation $|\phi_i^s(0)\rangle \rightarrow |\phi_i^s(T)\rangle$. Temporal operator $\mathcal{T}$ acts as

$$\mathcal{T}|\phi_i^s(0)\rangle = |\phi_i^s(t)\rangle.$$

More precisely, we can split the action of temporal operator between sounds and visuals: $\mathcal{T}_s \otimes \mathcal{T}_v$, where $\mathcal{T}$ can be freely compared with a tensor (or monoidal) functor. Temporarily neglecting coefficients, we have,

$$T((\mathcal{T}_v \otimes \mathcal{T}_s)|\phi_i^v\rangle) = T(\mathcal{T}_v|\phi_i^v\rangle \otimes \mathcal{T}_s(1)) = T(\mathcal{T}_v|\phi_i^v\rangle) = T(\mathcal{T}_v \otimes \mathcal{T}_s|\phi_i^v\rangle) = \mathcal{T}_s T(\mathcal{T}_v|\phi_i^v\rangle) = \mathcal{T}_s|\phi_i^s\rangle,$$

and

$$T(\mathcal{T}|\phi_i^s\rangle) = T((\mathcal{T}_v \otimes \mathcal{T}_s)|\psi_i^s\rangle) = T(\mathcal{T}_v(1) \otimes \mathcal{T}_s|\phi_i^s\rangle) = T(\mathcal{T}_s|\phi_i^s\rangle),$$

where I indicates a unit vector. But T applied to sounds acts as the identity, thus $T(\mathcal{T}_s|\phi_i^s\rangle) = \mathcal{T}_s|\phi_i^s\rangle$.

A possible computational goal of this approach could be the development of an app that takes as input an image, and sonifies it in real time. If the app allows a user manipulation of the image, with, for example, stretching and scaling (and any other homotopic transformation), or cutting/gluing/repeating (and any other non-homotopic transformation), the music should change in real time. This conjecture may lead to a conservation theorem. Because conservation theorems in physics are deeply linked to the concept of symmetry as a "change without change" (Wilczek 2016), and symmetry is traditionally associated with beauty in nature and in the arts (Weyl 1952), we can hope that seeking conservation and symmetry principles in music and visual arts could allow us to develop stronger tools to analyzing existing artworks and creating new ones.

Figures 3 and 4 illustrate the described procedure. We may wonder if a visual Gestalt can be translated into a sound Gestalt: formally, this means asking if the blue dotted arrows in Figure 4 are preserving the structure of Gestalt from one domain to the other. Is it possible/easier to translate a visual Gestalt into a musical Gestalt? The correspondence between visual kets and sound kets that were supposed to be joined by gestural similarity has been recently verified (Collins et al. 2018; Mannone et al. 2018). However, the correspondence of the initial Gestalt and of the final, overall Gestalt, that involve superpositions of kets with appropriate coefficients, has yet to be proven. Section 5 proposes an idea for such an experiment. If successful, this would extend the action of operator T to Gestalt(s), making the overall diagram commutative. Finally, let us focus on filtering F. We can associate to each pair of points of the original visual image the distances d_i between them in terms of connected path and composition of connected paths. Thus, each visual image can be enriched with all distances, see Figure 5. The image obtained via F can be thought of as a covering, like a triangulation, with charts (made by simple shapes) on the original image. Thus, F maps distances d_i between points of the original image onto distances between charts; this helps build a structure out of the charts, constituting a sort of atlas. In this way, we could in principle connect our formalism with hints from differential geometry.

Overall, our approach can be compared with the definition of "connecting bridges"[21] by Caramello (2016). In fact, a possible Gestalt translation reminds us of the "bridge object" discussed in Caramello's theory and recently applied within a gestural similarity framework (Mannone and Favali 2019). We can start from visuals and analyze the Gestalt extracting a core of Gestalt information as an independent object – that, in Caramello's framework, would

---

[21] According to Caramello (2016), "the importance of 'bridges' between different areas lies in the fact that they make it possible to transfer knowledge and methods between the areas, so that problems formulated in the language of one field can be tackled (and possibly solved) using techniques from a different field, and results in one area can be appropriately transferred to results in another." More specifically, Caramello refers to topoi as bridges to unify different areas of mathematics; however, the bridge metaphor can be generalized.
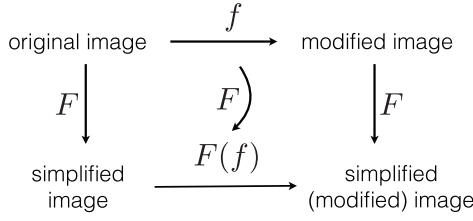
Figure 3. An image is simplified via *F*. The image can also be slightly modified via a sort of multiple transformations indicated via *f*. We can imagine simplifying the original image via filter *F*, and simplifying the modified image via *F*. The transformation between the two filtered images can be seen as the composition of *f* with *F*. We can also think of a class of transformations *f* mapped onto a class of transformations *F(f)*, as well as a class of original images and a class of modified images mapped onto a class of simplified images and a class of simplified modified images, respectively.
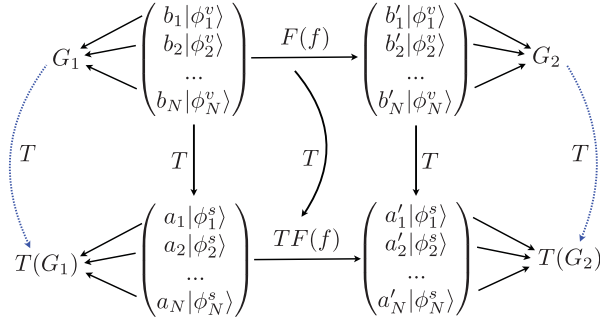
Figure 4. Each simplified image can be formally described via a superposition of visual kets. In the figure, we schematically represent the selected kets and the coefficients via an overall column vector. Transform *T* also maps basis decompositions (simple shapes with visual coefficients giving them a structure in visual space) in the domain of visuals into basis decompositions in the domain of sounds (musical sequences with coefficients giving them a structure in musical space). If we consider *f* as a class of transformations *f* as described in Figure 3, thus we can include in our description also the (class of) deformations from a simplified image into another simplified image as *F(f)*, and the comparison between a sonified simplified image with the sonified deformed shape via *TF(f)*. Finally, the symbols G_1 and G_2 are introduced here to represent the overall Gestalt of the first image and the second image, respectively.
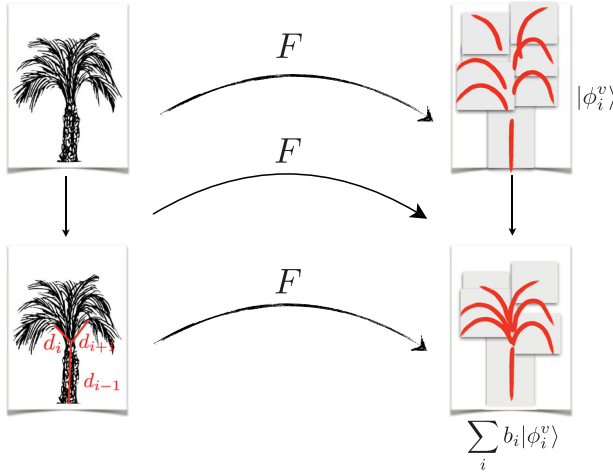

Figure 5. *F* maps original images onto their decomposition in terms of 'covering' charts made of simple shapes. If we define distances d_i between points of the original image as connected paths, we can map these distances via *F* to give a structure to the charts.
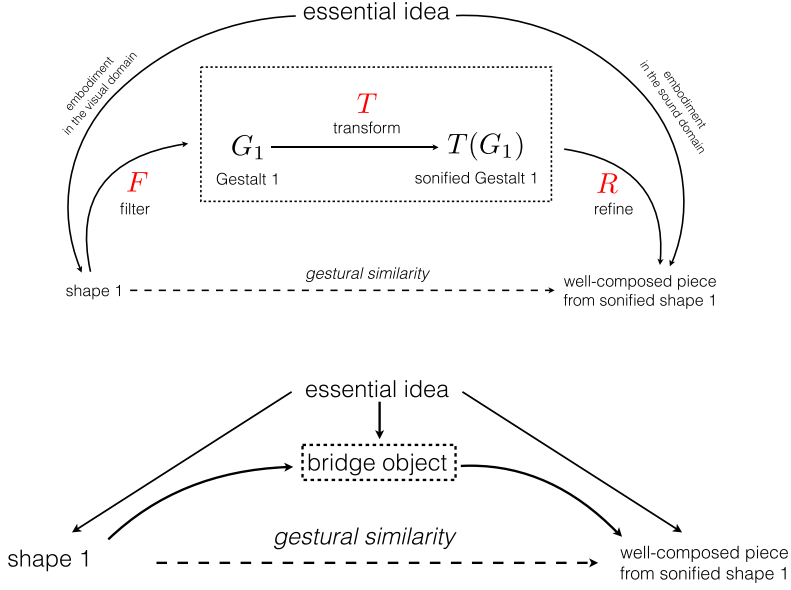
Figure 6. The text in the dotted box is freely inspired of Caramello's bridge (Caramello 2016), that has been recently enriched with gestural similarity/essential idea (Mannone and Favali 2019). The red letters highlight the discussed *RTF* procedure, that constitutes the overall bridge.

be metaphorically called a bridge object – and this core could be applied within the musical domain, constituting a basic composition whose Gestalt is, hopefully, the musical rendition of the initial one. The lower part of Figure 6 represents the relationship between the approach with an essential idea that can be embodied within the visual or the sonic domain, and a bridge object, for example a mathematical formalization of some shape properties, that joins visual and sounds. The upper part of Figure 6 shows that our approach, with a possible Gestalt translation, can be contextualized within the bridge object, where the entire bridge is constituted by the composition RTF, with filter *F*, transform *T*, and refinement *R*. We start from the domain of the continuous, with the original image; then, we take a discrete version of it with the filter, and, finally, with the refinement step, we get back to the continuous. A sonified filtered image (the result of *TF*) can be thought of as a discrete superposition of sonified visual kets, that is, specific sound kets. If a composer starts from it and he or she adds more notes and refines passages, the result of the *TF* sequence is turned into a complete and esthetically satisfactory musical piece. This is the refinement step (*R*), and it can be thought of as an interpolation.

If the work is done well, we might expect that the final result will remind the listener of some of the essential characteristics of the initial shape. Thus, the gestural similarity between the very initial shape and the very final shape will be verified, and the overall diagram will be commutative. Of course, if we reverse all the arrows, we can get the converse, that is, visual shapes from music (Mannone 2011). We can wonder if we can go from image to sound and back, and what is the amount of the information loss. This also depends on the amount of information encoded in the coefficients during the filtering step. If we sonify and image and then we want to reconstruct the original one, we do expect a partial and approximate reconstruction, because of the (two) filtering passages. We expect to preserve – and we can select parameters appropriately – the general recognizability of the original image, or, at least, of the equivalence class to which the image belongs. The time/loudness/pitch domain for music and (visual) space domain for images are independent, but they can be related via the reading-through-time of the (visual) space elements. We can consider a global time, corresponding

to the overall (visual) space, with some local times in it, flowing within each small figure or simple visual element. This can be compared with musical fragments performed by each musician during a whole symphony. Also, the visual–sound correspondence is more robust if we limit ourselves to drawing, or if we simplify complex visuals with collections of lines. This is consistent with some recent experimental approaches (Thoret et al. 2014a; Collins et al. 2018; Mannone et al. 2018).

3.1. *Gestalt and topology*

Through experiments, we can find out which, and how many visual kets are needed to recognize a given image; that is, what is the smaller index N^* necessary to recognize an image or part of it. A similar criterion can be adopted to find the minimum amount of base elements to recognize a melody or a generic musical piece. If we reduce musical pieces to their skeletal structure, a number of same-genre compositions are expected to be indistinguishable, as happens with Schenkerian analysis. An extension of our quantum-like approach can involve the topological definition of neighborhood. We can define a Gestalt neighborhood of a state $|\psi^v\rangle$ for visual (the same idea applies to sound states), where the center is the original image (that can be perfectly approximated with infinite coefficients), and where the most external layer is represented by the decomposition with the least amount of elements, N^* elements. A decomposition with a number $n \in \mathbb{N}$, with $n < N^*$ of elements, will produce ambiguities with other images; see Figure 7.

3.2. *Examples of visual and sound kets*

The proposed method can be used to obtain concrete musical sequences from a given image. The preliminary condition is the definition of a function from image to music, and the use of a dictionary of simple shapes and simple sound sequences connected via gestural similarity. An appropriate selection of suitable elementary visual kets and of corresponding sound kets should involve some experimental validation. For example, we can start from some hypothesis, based on gestural similarity, and we can check it via an experiment. We can give people a picture of
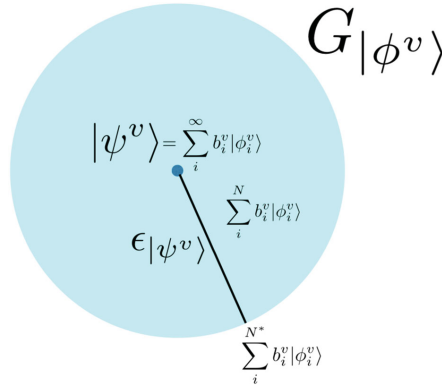

Figure 7. A topological Gestalt neighborhood G of a visual state $|\psi^v\rangle$. The closer we are to the center, the higher the number of base elements is, and the closer the decomposition is to the original state. Conversely, the farther we are from the center, and the more coarse-grained the approximation is. With a number of elements less than N^*, the original image is no longer recognizable. The same idea applies to sound states as well. Bistable states can be thought of as two apparently superposed circles, as the two faces of a cylinder with a small height. We cannot look at both faces at the same time. In the Figure, the choice of color is arbitrary.

a clear form, for instance, a palm, and a set of possible simple shapes, with the possibility to repeat a shape several times. Which shapes will be more often used for approximately sketching or tiling the given form? A similar experiment could involve an excerpt of a musical piece and a collection of short sound sequences. Which short sound sequences will be more often used to approximately sketch and represent the given musical piece? Another experiment would, of course, involve the correspondence between visual kets and sound kets. Initial validations with simple visuals and simple musical sequences has been done (Collins et al. 2018; Mannone et al. 2018); see Figure 8 for some of the visual and sound examples used in that cognition experiment. In Figure 8, the visual kets could be constituted by a dot, an arch, and a segment. Their combination (with repetition, angle selection, position in the space, formally expressed by coefficients b_i) generates a curved line, a collection of dots, and a sequence of angled segments, respectively. Each of these combinations appears as being generated by a specific gesture, and the same gesture can be thought of as being at the origin of the sound sequences on the right of the image.

Summarizing, we need both visual and sound kets with their coefficients. Coefficients for the visual kets elements characterize position and size, and their information will be translated into onset, duration, pitch range – and eventually also loudness, for tridimensional images (Mannone 2011) – values for the coefficients of sound kets. Also, we can use a (mathematical) composition of visual kets: e.g. a dotted spiral can be considered as the composition of a spiral form (this is an example of what we call here "envelope") and of dots ("pattern"). Musically, we will have an ascending–descending (along the spiral) sequence of notes, with *staccato* articulation. The

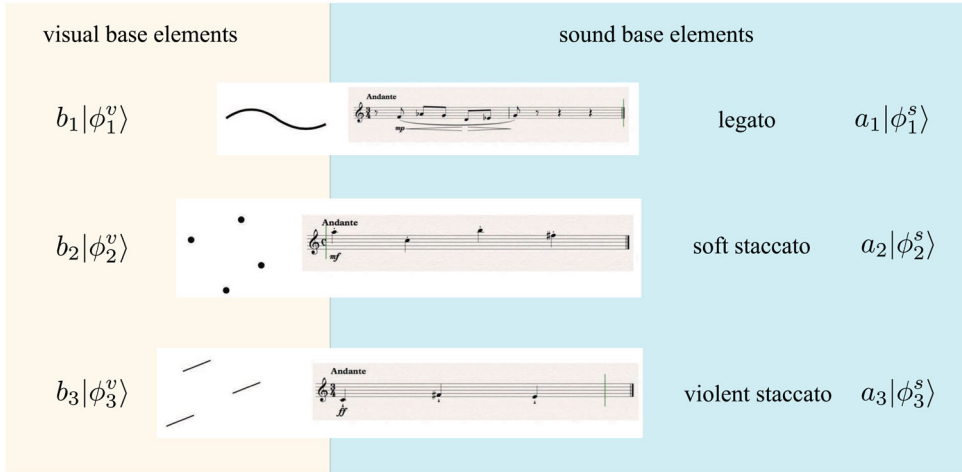| visual base elements				sound base elements	
$b_1\|\phi_1^v\rangle$				legato	$a_1\|\phi_1^s\rangle$
$b_2\|\phi_2^v\rangle$				soft staccato	$a_2\|\phi_2^s\rangle$
$b_3\|\phi_3^v\rangle$				violent staccato	$a_3\|\phi_3^s\rangle$

Figure 8. Let us consider three visual kets: let $|\phi_1^v\rangle$ be a dot, $|\phi_2^v\rangle$ a short segment, and $|\phi_3^v\rangle$ an arch. A very simple superposition of visual kets of the same kind gives, in this example, a curved line, a sequence of dots, and a sequence of angled segments, respectively. Because, in this case, the superpositions concern elements of the same kind, we can see them not as sums but just as the results of the action of coefficient b_1 for the first visual ket (i.e., "connect two arches to get a curved line," that is, "repeat the gesture of creating short curved lines in the described way"), coefficient b_2 for the second visual ket (i.e., "repeat the gesture of creating dots with this distance"), and coefficient b_3 for the third visual ket (i.e. "create the segment with this inclination as the result of a nervous gesture and repeat it with this distance"). These images (and many others), and the corresponding sounds (and many others), have been used in an experiment (Collins et al. 2018; Mannone et al. 2018) to measure the degree of the assessed gestural similarity. The results confirmed the theoretical expectations: in this example, a curved line is associated with a *legato* melody regardless of the speed or the pitch range; a sequence of dots is associated with *staccato* notes regardless of their pitch range or their speed or loudness, and, finally, a sequence of angled straight segments is associated with a *staccatissimo*, a "violent" *staccato* sequence of notes, independently by their pitch range (Collins et al. 2018; Mannone et al. 2018). The main parameters involved in gestural similarity in the analysis run so far are pitch contours – not within a specific pitch range – and articulations. There is a correspondence between a few elements of the visual base, and some elements of the sound base. In the pseudocode of Section 4, pitches of the sound sequence can be customized.
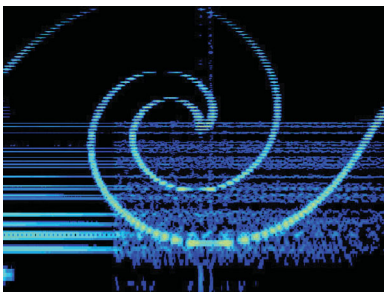
Figure 9.   A drawing within a spectrogram. Credit: Jarmo Niinisalo 2017.

staccato articulation corresponds to the isolated dots, while the spiral form can be translated into an equivalent pitch movement in the space of pitch. In fact, aside from time, which could be considered as flowing overall from left to right in an image, the small images a complex image is made of (the small images a complex image can be decomposed into) present a local time that flows on them. For example, if we see a straight line and a spiral, we can depict a melodic line following the straight line, and another melodic line following the profile of the spiral. This is intrinsically different from a simple scanning of an image from left to the right, which could be the sound rendition of the spiral within the spectrogram of Figure 9 – where, as an additional difference, time-frequency space is used rather than time-pitch space. We can also think of representations including a spectral dimension.

4.   A pseudocode and computational developments

In this section, we propose a possible computational application of the theoretical method described in Section 3. This procedure can be contextualized within the framework of algorithmic composition. Using a set of rules to compose music is not new, and the use of computer has quite a long tradition as well (Xenakis 2001). In algorithmic music, part of the compositional process is delegated to the machine, but the human creativity is preserved in the choice itself of the process and of the specific parameters that can dramatically change the final outcome. Also, in our method, images and their parameter choices, as well as the number of base elements to schematize them, play a relevant role.

Translating visual cues into auditory feedback is a well-known process among digital artists. Different techniques have been previously adopted for auditory image representations. One of the earliest approaches in digital art comes from the digital conversion of image into an audio spectrogram. This method involves the translation of the image's vertical components into the frequency of spectral components and the horizontal dimensions of the image in how the spectral properties of the sound evolve during time. An example of the application of this technique is *Windowlicker* by Aphex Twin (1999), which includes the image-sound conversion of a spiral (see Figure 9). A similar approach, that instead uses traditional instruments, has been adopted in an experiment conducted by Classical Music Reimagined: the sunset is translated into music (in the form of a score), performed and recorded, and then translated back into an image (in both cases using the spectrogram method), see Figure 10.

Research between machine learning and meaningful artistic applications is a fertile field (Sturm et al. 2018). Rather than adopting a one-to-one relationship between images and sounds, we propose a possible computational application of the method discussed in Section 3 using machine learning approaches. Specifically, we aim to utilize image recognition and segmentation
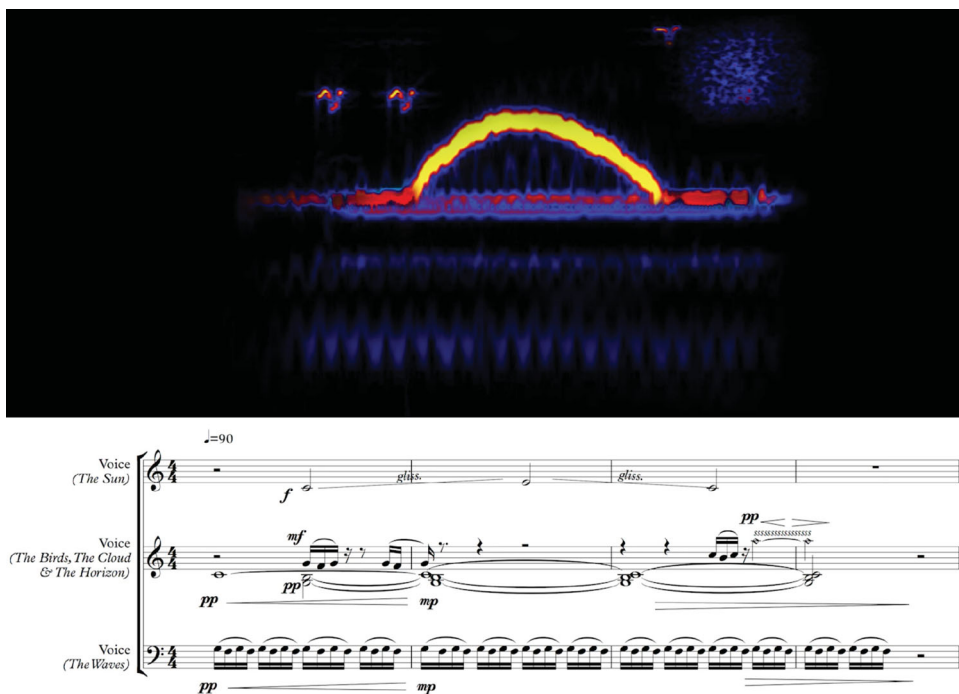
Figure 10. A drawing within a spectrogram. Image from https://youtu.be/N2DQFfID6eY. Credit: Classical Music Reimagined.

process using the TensorFlow API (Abadi et al. 2016). TensorFlow is an open source machine learning framework, originally developed by Google to empower a wide range of users with easy deployment of computation in different platforms.

By using TensorFlow's image recognition functionality, we will be able to identify the main content of the image, i.e., a tree in the case of Figure 11, and a street with cars in the case of Figure 12. TensorFlow can be used to apply machine learning to both music and images. The image recognition exploits a segmentation algorithm implemented using TensorFlow, that segments the image. At this point, the application can draw contours and mutate these into musical parts. The degree of segmentation may deal with more precise or coarse-graining segmentation of the same object, but also with higher and lower level objects: for example, a detailed view of a mountain will require trees, a detailed view of trees will require branches and leaves, and so on. Finally, leaves have veins that are similar in shape to branches, with a fractal-like self-similarity. Mathematically, we may describe this via nested categories (Mac Lane 1978; Mannone 2018c). In fact, this reminds us of a *n*-categorical depiction. Categories can be used in programming,

Figure 11. A schematic procedure: the input is an image; then, the envelope if extracted, and it is translated into a musical score. A more precise score takes into account also patterns of the image. The pictures are adapted from Krähenbühl and Koltun (2011).
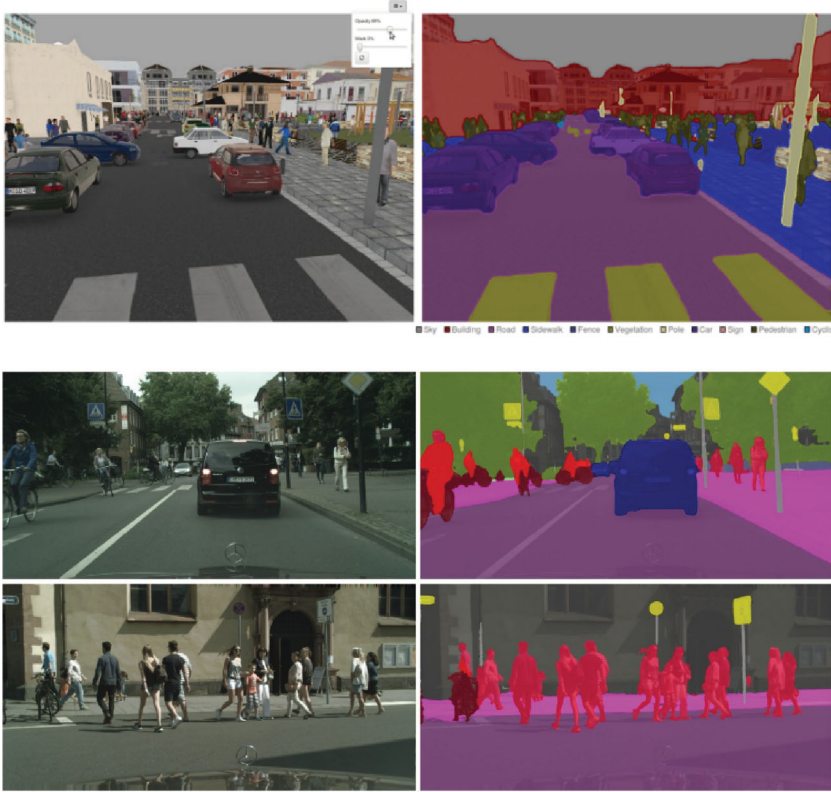
Figure 12.   An example of envelope extraction from different images as input. Top: from Heinrich 2014; bottom: from Kundu, Vineet and Koltun 2016.

suggesting new approaches. This topic may strengthen the connections between mathematical, artistic, and philosophical views of the reality around us, within an overall a esthetic perspective.

4.1.   *Details of the pseudocode*

This research informs the implementation of a software that empowers musicians in creating the skeleton of their compositions based on images. Precisely, it can generate a MIDI file out of an image, which can then be converted into a sound file or a musical score. We report the pseudocode of this future software; see Algorithm 1. As a general remark, the main differences between our pseudocode and implementations such as Xenakis' UPIC tablet (Xenakis 2001) lie in the role of time, the separation between patterns and envelopes, the potential applicability of the method to tridimensional images, and the use of the conjecture of gestural similarity.

As input, the algorithm takes (i) an image to convert, (ii) the pitch range within which the music has to be realized, (iii) the chosen instrumentation, and (iv) the discretization level. This last one corresponds to the resolution of the image analysis and the number of musical features (pitch progression, interval, dynamic, and pauses) that have to be considered in the image-music conversion process.

After passing the check on the image file format (pseudocode line 3), the image-recognition process takes place (line 4), otherwise the user is asked to select a different image. The image-recognition process uses a deep convolutional neural network, implemented using TensorFlow, to distinguish different objects within the image. If objects are detected, macro and micro aspects

---

Algorithm 1 MIDI file generator

1: **Input:** total time, pitch range, discretization level, instrument selection
2: **Output:** musical score
3: **if** valid image is entered **then**
4:     object recognition
5:     **if** objects are recognized **then**
6:         objects contour analysis related to discretization level
7:         **for** each object **do**
8:             object segmentation → patterns generation related to discretization level
9:             **if** pattern is a curved line **then**
10:                 musical technique → legato
11:             **if** pattern is composed of dots **then**
12:                 musical technique → soft staccato
13:             **if** pattern is a straight line **then**
14:                 musical technique → violent staccato
15:             **if** pitch progression feature is selected **then**
16:                 pattern direction (up/down) → pitch progression (low/high)
17:             **if** interval feature is selected **then**
18:                 pattern angle (small/big) → interval (small/big)
19:             **if** dynamic feature is selected **then**
20:                 pattern thickness (thin/thick) → sound dynamic (piano/forte)
21:             **if** blank space between patterns **then**
22:                 space wideness (narrow/wide) → pause length (short/long)
23:             notes transposition along object contour
24:         notes transposition according to selected musical instrument
25:         create MIDI file
26:     error: no objects have been recognized. Please, upload a different image
27: error: the uploaded image cannot be processed. Please, upload a different image

---

are extracted. Macro aspects refer to the objects contour (line 1). Micro aspects are extracted by decomposing the image contour into smaller objects (see Section 3), which are then segmented into patterns made of dots, straight lines, and curved lines (line 8).

Subsequently, the pattern-music relationships are defined with a series of tests. The first three tests establish the technique with which the music is to be played and how to define pitch and time characteristics. Referring to concepts described through Figure 8, lines 9, 11, and 13 of the pseudocode test if the music is to be played with *legato*, soft *staccato*, or violent *staccato* techniques respectively. If a pattern is made of curved lines, then the notes will be indicated as *legato*, if the pattern is made of dots, as soft *staccato*, and if made of straight lines, as violent *staccato*. Precisely, the longer the lines, the more violent is the *staccato*. Afterwards, pitch progressions (line 15), interval (17), dynamics (19) and pauses between notes (21) are established. Pitch progression has a direct relationship with (visual) pattern directions (upward/downward), which is useful to establish if pitches are rising or lowering. The interval between consecutive notes is defined by looking at the pattern angle: the higher the angle, the bigger is the interval between notes. Dynamics are determined by pattern thickness – the thicker the line in the image to be sonified, the louder the notes. Finally, pauses are defined by analyzing white spaces between each element of the pattern – the wider the white spaces, the longer the pauses. Notes created by analyzing patterns are then musically organized to represent the macro aspects of the given image, that is, the whole image contour, its envelope (line 23). As the last step, the musical score is transposed to the chosen musical instrument and register, and the MIDI file is generated.

Our pseudocode is contextualized within machine learning studies. It is worth mentioning that machine-learning studies are more and more being re-read in light of categories, using compositionality and functor definitions. This is true of Bayesian machine learning (Culbertson and Sturtz 2013) and of backpropagation seen as a functor (Fong, Spivak, and Tuyéras 2019). However, to

the best of our knowledge, our approach potentially joining music, image, machine learning, and categories together is novel and can initiate a dialogue with other ongoing or already established studies.

5.   Discussion: open questions and a possible cognitive experiment

The proposed approach can be seen as a method to compose music and to analyze music and images. The aesthetic output can raise new questions about the strength of connections between music and images, and the depth of our understanding of these relationships. In particular, future research would involve specific cases of multimodal perception, choice of the bases, and experimental validation of the minimal-shape principle cited in Section 3. In particular, two main questions arise:

(1) How to find a good basis for images, general enough to be adapted in different cases, and where each of the basis elements can be meaningfully mapped into sound via gestural similarity?
(2) With this method of basis change with well-chosen visual kets and visual/sound basis elements' correspondences, is it possible/easier to translate a visual Gestalt into a musical Gestalt?

In a future experiment, the participants can be given a few different complex forms and several possibilities for simple shapes (geometric curves, points, lines) as visual kets, and then be asked to select the minimum number of simple shapes to approximate each complex form, making the superposition recognizable as a simplified version of the original complex form. For each complex form, we can evaluate how many simple shapes are needed, and which ones they are. This can help establish a principle of *minimum* and to recognize which shapes are more needed than others. Then, each simple shape should be associated with a simple sound sequence. The second step of the experiment should test the appropriateness of this association. Participants might be asked to associate sonified complete forms with each visual forms, rating their similarity. In fact, the outcomes of a future experiment can also suggest which restrictive conditions may be needed to get an effective sonification and a Gestalt translation. In designing such a new experiment, we could refer to a recent experimental study, where participants were asked to draw lines and make arm gestures according to musical sequences; examples involved vocal sequences belonging to different genres and styles (Kelkar and Jensenius 2017).

Further conceptual developments of the proposed work can deal with the dialogue between continuous and discrete, and with new applications of Deleuze's studies within a Gestalt framework. The duality between discretization and continuity is implied in Deleuze's studies about time, not only in music. Deleuze made a distinction between "striated time" and "smooth time." Striated time meaning a discrete time, in opposition with a smooth time, a continuous one. We may see the first as a quantitative, measured time, and the second as a qualitative time. A similar distinction appears in Pierre Boulez's lessons in Darmstadt, with a distinction between pulsed and un-pulsed time (Campbell 2013). We can subdivide time in small parts, but some elements allow us to perceive it as a smooth time in a Gestalt perspective. A discrete sequence of events in time, as a sequence of photographic shots, versus a continuous flow, as a continuous movement, is relevant also in musical rendition of visual shapes. In Deleuze's philosophical research, musical conjectures have been applied to other areas of human knowledge, and vice versa. This strengthens the relevance of interdisciplinary studies, that can be adopted as metaphors in several other areas. Our open questions about the translatability of Gestalt can be connected with the problem of "translation" in the generalized sense of "translatability of knowledge." This is the object of

recent philosophical research (Alunni 2013), that also involves diagrammatic thinking (Alunni 2004). Our work can contribute to the debate about possible shared, crossmodal elements not only in the arts, but in perception: according to Reybrouck (1997, 62), this also deals with "the question of the existence of universals of knowing and perception." Also, one is reminded of the universals in philosophy.

Finally, some new, brave questions may arise, in particular regarding not only the Gestalt-translatability, but also the translatability of *beauty* and of *beautiful Gestalt*: can a beautiful visual shape generate a beautiful music and vice versa? Should we carefully choose all parameters, including discretization levels, to control resulting harmonies, and so on, to translate beauty? We expect that the answers to be positive, yet we have to carefully play around with the degrees of freedom allowed by the proposed method, which primarily aims only at preserving the recognizability of the original shape. This could open a debate about hypothetic quantitative definitions of beauty independent from a specific artistic language. We wonder if the perception of *beauty* follows some criterion of equilibrium between *anticipation* and *surprise*, that can be quantitatively defined in terms of pattern variety levels – with some *aesthetic indeterminacy principle*. The relationship between beautiful Gestalt and gestural similarity should also be investigated. The connection between gestural similarity and expressivity in music has already been discussed (Mannone 2018a). The concept of beauty can be lifted above a specific artistic language; it is the case of smoothness in visual lines and in singing, a visual analogy for a good vocal technique used by Truslit (Repp 1992). Finally, the interest of further studies about Gestalt invariance can be useful not only within an artistic framework, but also for practical, and specifically diagnostic purposes: it is the case of the sonification of patients' brain, to create, in the mind of the listener, an auditory idea of such a complex 3D-physical object (Rogińska, Friedman, and Mohanraj 2013).

6. Conclusions

In this article, we started with insights from philosophy, cognition, physics, and mathematical theory of music, looking for correlations among these different areas. Mathematics and diagrammatic thinking with their abstraction power can be used to formalize music and mapping strategies from image to sound. We expected to be able to decompose complex images and musical sequences in terms of simple elements combined to be together with suitable kets and coefficients. Following the Deleuzian *rhizomatic thought*, it is possible to look for links among these elements of images and sounds.

With the idea of a quantization of images and sounds, we borrowed the formalism of Dirac notation from quantum physics, and we connected each step through diagrams. We formally described a method to derive music from visual shapes based on the *gestural similarity* criterion between simple shapes and simple musical sequences. This method can be used as a compositional technique to develop creativity and sonify different shapes and forms. However, this research also raises open-ended questions about the translatability of Gestalt from a sensory domain to another. Such a question requires cognitive experiments to help improve our theoretical model. We also described a pseudocode that could automatize (almost) each passage of the described method. Further developments of the proposed research may involve computer science applications, psychological validation of the method, and evaluation of the validity criteria for a hypothetical Gestalt translation. These criteria would give new blood to mathematical research in this framework with the definition of conservation principles.

The ultimate goal, from a STEAM (Science, Technology, Engineering, Arts, and Mathematics) perspective, would be a better connection and a dream of unification of different areas of human knowledge.

Authors' contributions

Acknowledgments

Disclosure statement

ORCID

Maria Mannone http://orcid.org/0000-0003-3606-3436
Federico Favali http://orcid.org/0000-0002-8505-6163
Balandino Di Donato http://orcid.org/0000-0001-6993-2445
Luca Turchet http://orcid.org/0000-0003-0711-8098

References

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. "TensorFlow: A System for Large-Scale Machine Learning." In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 265–283. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Allen, Paul G., and Paul A. Kolers. 1981. "Sensory Specificity of Apparent Motion." *Journal of Experimental Psychology: Human Perception and Performance* 7 (6): 1318–1326.

Alunni, Charles. 2004. "Diagrammes et catégories comme prolégomènes à la question : qu'est-ce que s'orienter diagrammatiquement dans la pensée ?" In *Penser par le diagramme – de Gilles Deleuze à Gilles Châtelet*, edited by Batt Nöelle. Paris: Presse Universitaire de Vincennes.

Alunni, Charles. 2013. "De La Traducibilité Des Savoirs." *Review Sciences/Lettres* 1, 1–14. http://rsl.revues.org/293.

Aphex Twin, 1999. "Windowlicker." CD. Warp Records – WAP105CD.

Amiot, Emmanuel. 2016. *Music Through Fourier Spaces: Discrete Fourier Transform in Music Theory*. Heidelberg: Springer.

Arias, Juan Sebastián. 2018. "Spaces of Gestures are Function Spaces." *Journal of Mathematics and Music* 12 (2): 89–105.

Baez, John, and Lauda Aaron. 2011. "A Prehistory of n-Categorical Physics." In *Deep Beauty: Understanding the Quantum World Through Mathematical Innovation*, edited by Hans Halvorson, 13–128. Cambridge: Cambridge University Press.

Campbell, Edward. 2013. *Music after Deleuze*. London: Bloomsbury.

Caramello, Olivia. 2016. "The Theory of Topos-Theoretic 'Bridges' – A Conceptual Introduction." *Glass-Bead*. Accessed 9 November 2018. https://www.glass-bead.org/article/the-theory-of-topos-theoretic-bridges-a-conceptual-introduction/?lang = enview

Caramello, Olivia. 2018. *Theories, Sites, Toposes*. New York, NY: Oxford University Press.

Coecke, Bob. 2006. "Introducing Categories for the Practising Physicist." In *What is Category Theory? Advanced Studies in Mathematics and Logic*. Vol. 30, edited by Giandomenico Sica, 45–74. Monza: Polimetrica Publishing.

Cohen-Tannoudji, Claude, Bernard Diu, and Franck Laloë. 2005. *Quantum Mechanics*. Paris: Wiley-VCH.

Collins, Tom, Maria Mannone, Delilah Hsu, and Dimitri Papageorgiou. 2018. "Psychological Validation of the Mathematical Theory of Musical Gestures." Submitted.

Culbertson, Jared, and Kirk Sturtz. 2013. "Bayesian Machine Learning via Category Theory." ArXiV. https://arxiv.org/abs/1312.1445.

Dalla Chiara, Maria L., Roberto Giuntini, Roberto Leporini, Eleonora Negri, and Giuseppe Sergioli. 2015. "Quantum Information, Cognition, and Music." *Frontiers in Psychology* 6, 1583.

D'Arcy Thompson, Wentworth. 1966. *On Growth and Form, an Abridged Edition*, edited by John Tyler Bonner. Cambridge: Cambridge University Press.

Deleuze, Gilles. 1986. *Foucault*. Paris: Les Éditions de Minuit.

Deroy, Ophelia, Anne-Sylvie Crisinel, and Charles Spence. 2013. "Crossmodal Correspondences Between Odors and Contingent Features: Odors, Musical Notes, and Geometrical Shapes." *Psychonomic Bulletin and Review* 20, 878–896.

Deroy, Ophelia, and Charles Spence. 2013. "Why We Are Not All Synesthete (Not Even Weakly So)." *Psychonomic Bulletin and Review* 20 (4): 643–664.

Dirac, Paul A. M. 1939. "A New Notation for Quantum Mechanics." *Mathematical Proceedings of the Cambridge Philosophical Society* 35, 416–418. http://journals.cambridge.org/abstract_S030500410000 21162.

Ehresmann, Andrée C., and Jaime Gomez-Ramirez. 2016. "Conciliating Neuroscience and Phenomenology Via Category Theory." *Progress in Biophysics and Molecular Biology (JPMB)* 19 (2): 347–359.

Ehresmann, Andrée C., and Jean-Paul Vanbremeersch. 2007. *Memory Evolutive Systems*. Amsterdam: Elsevier.

Fiore, Thomas, and Thomas Noll. 2011. "Commuting Groups and the Topos of Triads." *Proceedings of the 3rd International Conference on Mathematics and Computation in Music*, edited by Carlos Agon, Emmanuel Amiot, Moreno Andreatta, Gérard Assayag, Jean Bresson, John Manderau, 69–83. Heidelberg: Springer.

Fong, Brendan, David Spivak, and Rémy Tuyéras. 2019. "Backpropagators as Functor: A Compositional Perspective on Supervised Learning." ArXiV. https://arxiv.org/pdf/1711.10455.pdf.

Gabor, Dennis. 1947. "Acoustical Quanta and The Theory of Hearing." *Nature* 159 (4044): 591–594.

beim Graben, Peter, and Reinhard Blutner. 2017. "Toward a Gauge Theory of Musical Forces." In *Proceedings of the 10th International Conference on Quantum Interaction (QI 2016)*, ed. Jose Acacio de Barros, Bob Coecke, and Emmanuel Pothos, 99–111. Cham: Springer International Publishing.

beim Graben, Peter, and Reinhard Blutner. 2019. "Quantum Approaches to Music Cognition." *Journal of Mathematical Psychology* 91, 38–50.

Hallé, Francis, Roleof A. A. Olderman, and Philip Barry Tomlinson. 1978. *Tropical Trees and Forests: An Architectural Analysis*. Berlin: Springer-Verlag.

Heinrich, Greg. 2014. "Image Segmentation Using DIGITS 5", NVDIA Developer Blog, November. https://devblogs.nvidia.com/image-segmentation-using-digits-5/

Hermann, Thomas, Andy Hunt, and John Neuhoff. 2011. *The Sonification Handbook*. Berlin: Logos Verlag Berlin.

Hillier, Jean, and Gareth Abrahams. 2013. *Deleuze and Guattari: Jean Hillier in Conversation with Gareth Abrahams*. The University of Liverpool Repository. Association of European Schools of Planning, Wroclaw, Poland. https://livrepository.liverpool.ac.uk/3004625/1/aesop-ya-booklet-series-a-vol-1-deluze-and-guattari-jean-hillier-in-conversation-with-gareth-abrahams.pdf.

Jedrzejewski, Franck. 2006. *Mathematical Theory of Music*. Paris: Delatour France – IRCAM.

Jedrzejewski, Franck. 2019. *Hétérotopies musicales. Modèles mathématiques de la musique*. Paris: Hermann.

Kainen, Paul C. 2009. "On the Ehresmann-Vanbremeersch Theory and Mathematical Biology." *Axiomathes* 19, 225–244.

Kelkar, Tejaswinee, and Alexander Refsum Jensenius. 2017. "Analyzing Free-Hand Sound-Tracings of Melodic Phrases." *Applied Sciences* 8 (135): 1–21.

Khrennikov, Andrei. 2010. *Ubiquitous Quantum Structure: From Psychology to Finance*. Berlin: Springer.

Krähenbühl, Philipp, Vladlen Koltun. 2011. "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials." *Advances in Neural Information Processing Systems* 24: 109–117.

Kubota, Akihiro, Hirokazu Hori, Makoto Naruse, and Fuminori Akiba. 2017. "A New Kind of Aesthetics – The Mathematical Structure of the Aesthetic." *Philosophies* 3 (14): 1–10.

Kubovy, Michael, and Michael Schutz. 2010. "Audio-visual Objects." *Review of Philosophy and Psychology* 1 (1): 41–61.

Kubovy, Michael, and Minhong Yu. 2012. "Multistability, Cross-modal Binding and the Additivity of Conjoint Grouping Principles." *Philosophical Transactions of the Royal Society B* 367, 954–964.

Kundu, A., V. Vineet and V. Koltun. 2016. "Feature Space Optimization for Semantic Video Segmentation," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 3168–3175.

Küssner, Mats, and Baptiste Caramiaux. 2015. "Motor Invariants in Gestural Responses to Music." *Proceedings of the International Conference on the Multimodal Experience of Music*. https://www.hrionline.ac.uk/openbook/chapter/ICMEM2015-Kussner.

Lacquaniti, Francesco, Carlo Terzuolo, and Paolo Viviani. 1984. "Global Metric Properties and Preparatory Processes in Drawing Movements." In *Preparatory States and Processes*, edited by Sylvan Kornblum and Jean Requin, 357–370. Hillsdale: Psychology Press.

Lawvere, William, and Stephen Schanuel. 2009. *Conceptual Mathematics. A First Introduction to Categories*. Cambridge: Cambridge University Press.

Lerdhal, Fred, and Ray Jackendoff. 1983. *A Generative Theory of Tonal Music*. Cambridge: MIT Press.

Letelier, Juan-Carlos, Jorge Soto-Andrade, Flavio Guinez Abarzua, Athel Cornish-Bowden, and Maria Luz Cardenas. 2006. "Organizational Invariance and Metabolic Closure: Analysis in Terms of (M,R) Systems." *Journal of Theoretical Biology* 238, 949–961.

Mac Lane, Saunders. 1978. *Categories for the Working Mathematician*. Heidelberg: Springer.

Mannone, Maria. 2011. *Dalla Musica all'Immagine, dall'Immagine alla Musica*. Palermo: Compostampa.

Mannone, Maria. 2017. "Gestures between Scores and Acoustics: A Creative Application to Orchestra." PhD dissertation, University of Minnesota, 2018. https://conservancy.umn.edu/handle/11299/188931.

Mannone, Maria. 2018a. "Introduction to Gestural Similarity in Music. An Application of Category Theory to the Orchestra." *Journal of Mathematics and Music* 12 (2): 63–87.

Mannone, Maria. 2018b. "cARTegory Theory: Framing Aesthetics of Mathematics." *Journal of Humanistic Mathematics* 9 (1): 277–294. https://scholarship.claremont.edu/jhm/vol9/iss1/16/.

Mannone, Maria. 2018c. "Knots, Music and DNA." *Journal of Creative Music Systems* 2 (2): 1–20. https://www.jcms.org.uk/article/id/523/.

Mannone, Maria. 2019. *Mathematics, Nature, Art*. Palermo: Palermo University Press.

Mannone, Maria, Tom Collins, Delilah Hsu, and Dimitri Papageorgiou. 2018. "Mathematical Theory of Musical Gestures, Gestural Similarity, and Psychology: A First Validating Experiment." Poster presentation at WiMIR Workshop, MIR satellite event, Paris, September 2018.

Mannone, Maria, and Giuseppe Compagno. 2014. "Characterization of the Degree of Musical Non-Markovianity." https://arxiv.org/abs/1306.0229.

Mannone, Maria, and Federico Favali. 2019. "Categories, Musical Instruments, and Drawings: A Unification Dream." In *Proceedings of the Mathematics and Computation in Music Conference 2019*, Madrid, edited by Mariana Montiel, Francisco Gómez-Martín, and Octavio A. Agustín-Aquino, 59–72. Heidelberg: Springer.

Mannone, Maria, and Luca Turchet. 2019. "Shall We (Math and) Dance?" In *Proceedings of the Mathematics and Computation in Music Conference 2019*, Madrid, ed. Mariana Montiel, Francisco Gómez-Martín, and Octavio A. Agustín-Aquino, 84–97. Heidelberg: Springer.

Mazzola, Guerino, and Moreno Andreatta. 2010. "Diagrams, Gestures and Formulae in Music." *Journal of Mathematics and Music* 1 (1): 23–46.

Mazzola, Guerino. 2018. *The Topos of Music*, second edition. Heidelberg: Springer.

Niinisalo, Jarmo. 2017. "The Aphex Face" (blogpost). http://www.bastwood.com/?page_id = 10.

Nobile, Luca. 2013. "Words in the Mirror: Analysing the Sensorimotor Interface between Phonetics and Semantics in Italian." In *Iconicity in Language and Literature 10: Semblance and Signification*, edited by Pascal Michelucci, Olga Fischer, and Christina Ljungberg, 101–131. Amsterdam/Philadelphia: John Benjamins.

Northoff, Georg, Naotsugu Tsuchiya, and Hayato Saigo. 2019. "Mathematics and the Brain – A Category Theoretic Approach to go Beyond the Neural Correlates of Consciousness." ArXiV. https://www.biorxiv.org/content/10.1101/674242v1.

Plamondon, Réjean, and Wacef Guerfali. 1998. "The 2/3 Power Law: When and Why?" *Acta Psychologica* 100, 85–96.

Phillips, Steven, and William Wilson. 2010. "Categorical Compositionality: A Category Theory Explanation for the Systematicity of Human Cognition." *PLoS Computational Biology* 6 (7): 1–14. https://doi.org/10.1371/journal.pcbi.1000858.

Popoff, Alexandre, Moreno Andreatta, and Andrée Ehresmann. 2015. "A Categorical Generalization of Klumpenhouwer Networks." In *Mathematics and Computation in Music. MCM 2015*, edited by Tom Collins and Anja Volk. Heidelberg: Springer.

Prusinkiewicz, Przemyslaw, and James Hanan. 1980. *Lindenmayer Systems, Fractals, and Plants*. Lecture Notes in Biomathematics. New York: Springer-Verlag.

Prusinkiewicz, Przemyslaw, and Aristid Lindenmayer. 2004. *The Algorithmic Beauty of Plants*. New York: Springer.

Repp, Bruno. 1992. *Music as Motion: A Synopsis of Alexander Truslit's (1938) "Gestaltung und Bewegung in der Musik."* Haskins Laboratories Status Report on Speech Research, SR-111/112: 265–278.

Reybrouck, Mark. 1997. "Gestalt Concepts and Music: Limitations and Possibilities." In *Proceeding Music, Gestalt, and Computing – Studies in Cognitive and Systematic Musicology*, edited by Marc Leman, 57–69. Heidelberg: Springer.

Rocchesso, Davide, and Maria Mannone. 2018. "Embryo of a Quantum Vocal Theory of Sound." In *Conference of Musical Informatics (CIM) Proceedings*, Udine. http://cim.lim.di.unimi.it/2018_CIM_XXII_Atti.pdf.

Rogińska, Agnieszka, Kent Friedman, and Hariharan Mohanraj. 2013. "Exploring Sonification for Augmenting Brain Scan Data." In *Proceedings of the International Conference on Auditory Display*. http://hdl.handle.net/1853/51653.

Rosen, Robert. 1958. "The Representation of Biological Systems From the Standpoint of the Theory of Categories." *Bulletin of Mathematical Biophysics* 20 (4): 317–341.

Rosenblum, Lawrence D., James W. Dias, and Josh Dorsi. 2016. "The Supramodal Brain: Implications for Auditory Perception." *Journal of Cognitive Psychology* 29 (1): 65–87.

Schutz, Michael, and Scott Lipscomb. 2007. "Hearing Gestures, Seeing Music: Vision Influences Perceived Tone Duration." *Perception* 36 (6): 888–897.

Sparavigna, Amelia Carolina, and Roberto Marazzato. 2016. "Ket Vectors in Three-Dimensional Space and their Cross Product." *HAL Archives-Ouvertes*. https://hal.archives-ouvertes.fr/hal-01288341/document.

Spence, Charles. 2015. "Crossmodal Perceptual Organization." In *Oxford Handbook of Perceptual Organization*, edited by Johan Wagemans. Oxford: Oxford University Press.

Spence, Charles and Cesare V. Parise. 2012. "The Cognitive Neuroscience of Crossmodal Correspondences" *i-Perception* 3(7): 410–412.

Sturm, Bob, Oded Ben-Tal, Úna Monaghan, Nick Collins, Dorien Herremans, Elaine Chew, Gaëtan Hadjeres, Emmanuel Deruty, and Francois Pachet. 2018. "Machine Learning Research that Matters for Music Creation: A Case Study." *Journal of New Music Research* 48 (1): 1–21. https://www.tandfonline.com/doi/abs/10.1080/09298215.2018.1515233?journalCode=nnmr20.

Thoret, Etienne, Mitsuko Aramaki, Christophe Bourdin, Lionel Bringoux, Richard Kronland-Martinet, and Sølvi Ystad. 2014. "Audio-Motor Synchronization: The Effect of Mapping Between Kinematics and Acoustic Cues on Geometric Motor Features." In *Sound, Music, and Motion. CMMR 2013*, edited by Mitsuko Aramaki, Olivier Derrien, Richard Kronland-Martinet, and Sølvi Ystad. Cham: Springer.

Thoret, Etienne, Mitsuko Aramaki, Lionel Bringoux, Sølvi Ystad, and Richard Kronland-Martinet. 2016. "Seeing Circles and Drawing Ellipses: When Sound Biases Reproduction of Visual Motion." *PLoS ONE* 11 (4): e0154475.

Thoret, Etienne, Mitsuko Aramaki, Richard Kronland-Martinet, Jean-Luc Velay, and Sølvi Ystad. 2014. "From Sound to Shape: Auditory Perception of Drawing Movements." *Journal of Experimental Psychology: Human Perception and Performance* 40 (3): 983–994.

Weyl, Hermann. 1952. *Symmetry* (reprint 2016). Princeton, NJ: Princeton Science Library.

Wilczek, Franck. 2016. *A Beautiful Question: Finding Nature's Deep Design*. New York: Penguin Books.

Xenakis, Iannis. 2001. *Formalized Music: Thought and Mathematics in Composition* (revised edition). New York: Pendragon Press.

Youssry, Akram, Ahmed El-Rafei, and Salwa Elramly. 2015. "A Quantum Mechanics-Based Framework for Image Processing and Its Application to Image Segmentation." *Quantum Information Processing* 14 (10): 3613–3638.

Zbikowski, Lawrence. 2017. "Music, Analogy, and Metaphor." In *The Routledge Companion to Music Cognition*, edited by Richard Ashley and Renee Timmers, 501–512. New York: Routledge.