A Multidimensional Sketching Interface for Visual Interaction with Corpus-Based Concatenative Sound Synthesis

Avgoustinos Tsiros

Submitted in partial fulfilment of the requirements of Edinburgh Napier University for the degree of Doctor of Philosophy

February 2016

Abstract

The present research sought to investigate the correspondence between auditory and visual feature dimensions and to utilise this knowledge in order to inform the design of audio-visual mappings for visual control of sound synthesis. The first stage of the research involved the design and implementation of *Morpheme*, a novel interface for interaction with corpus-based concatenative synthesis. *Morpheme* uses sketching as a model for interaction between the user and the computer. The purpose of the system is to facilitate the expression of sound design ideas by describing the qualities of the sound to be synthesised in visual terms, using a set of perceptually meaningful audio-visual feature associations. The second stage of the research involved the preparation of two multidimensional mappings for the association between auditory and visual dimensions.

The third stage of this research involved the evaluation of the Audio-Visual (A/V) mappings and of Morpheme's user interface. The evaluation comprised two controlled experiments, an online study and a user study. Our findings suggest that the strength of the perceived correspondence between the A/V associations prevails over the timbre characteristics of the sounds used to render the complementary polar features. Hence, the empirical evidence gathered by previous research is generalizable/ applicable to different contexts and the overall dimensionality of the sound used to render should not have a very significant effect on the comprehensibility and usability of an A/V mapping. However, the findings of the present research also show that there is a non-linear interaction between the harmonicity of the corpus and the perceived correspondence of the audio-visual associations. For example, strongly correlated cross-modal cues such as size-loudness or vertical position-pitch are affected less by the harmonicity of the audio corpus in comparison to weaker correlated dimensions (e.g. texture granularity-sound dissonance). No significant differences were revealed as a result of musical/audio training. The third study consisted of an evaluation of *Morpheme's* user interface were participants were asked to use the system to design a sound for a given video footage. The usability of the system was found to be satisfactory.

An interface for drawing visual queries was developed for high level control of the retrieval and signal processing algorithms of concatenative sound synthesis. This thesis elaborates on previous research findings and proposes two methods for empirically driven validation of audio-visual mappings for sound synthesis. These methods could be applied to a wide range of contexts in order to inform the design of cognitively useful multi-modal interfaces and representation and rendering of multimodal data. Moreover this research contributes to the broader understanding of multimodal perception by gathering empirical evidence about the correspondence between auditory and visual feature dimensions and by investigating which factors affect the perceived congruency between aural and visual structures.

Table of Contents

List of figures

Acknowledgements

I would like to thank my supervisor, Dr. Grégory Leplâtre for his insightful guidance, patience and support. Dr. Smyth, Michael for providing all the support needed to complete my studies. Thanks are also due to the team at Edinburgh Napier University and the Institute for Informatics and Digital Innovation (IIDI) who gave me the opportunity and the resources to do this research.

I would also like to thank all the staff at the Centre for Interaction (Edinburgh Napier University) as well as all of the participants who volunteered to take part in the experiments described in this thesis. Particular thanks go to Dr. Iain McGregor, Dr. Alistair Armitage, Dr. Susan Turner. I would also like to thank Professor David Benyon and Dr. Kia Ng for to examining my thesis.

My heartfelt thanks to Cecile for her love and patience all these years. Finally, I would like to thank my family for their unconditional support and love.

1 Introduction

1.1 Problem Definition

The research described in this thesis is concerned with visual interaction with corpus based concatenative synthesis. More specifically this thesis presents an interface that enables visual control of corpus-based concatenative sound synthesis for creative applications (e.g. sound design, electroacoustic composition). Corpus-based concatenative synthesis methods allow sounds to be synthesised from pre-recorded samples that have been sliced into small segments known as *audio-units*. The audio-units are then analysed, tagged with the analysis data, and stored in a database. Concatenative synthesis works by recalling audio-units from a database using either distance or similarity measures to find the best match between a stream of n-dimensional feature vectors used as targets and the feature vectors of pre-analysed audio-units found in the database. Sound synthesis is accomplished by combining the audio-units that are retrieved from the database, creating a new sound sequence. In the context of corpus-based concatenative synthesis sound is represented using three different levels of sound descriptors, i.e. signal, perceptual and musical level (Schwarz, 2005). Unlike other synthesis methods were the sound is represented by low-level signal processing parameters related to the given sound processing unit. This opens up new possibilities for user interaction, as users can synthesise sounds by describing the properties of sound in terms of perceptual cues rather than through direct manipulation of synthesis parameters. However to fully take advantage of these new possibilities, we first need to identify appropriate models for multimodal interaction and to devise mappings strategies that create perceptually meaningful association between characteristics of the users' sensory-motor inputs and the feature dimensions of sound. When audio feature vectors are used as the target for querying the database, the question of how to associate the feature dimensions between the target and the audio-units of the corpus is easy to answer, as the same features dimensions are available in both the target and the database side. However when visual, gestural and/or haptic data are used to query the database, it is harder to decide how the feature dimensions involved in the query should be associated.

As technological developments enable us to represent and interact with sound at a level that is closer to the perception of sound, the question of how we can approach the design and evaluation of cross-modal mappings in objective terms poses a challenge to a number of disciplines including computer music, musicology, information display, sensory substitution amongst other disciplines. In recent years, there has been a growing interest in the study of cross-modal association in the context of sound and music technology and practice. These research efforts demonstrate two things, first that it is important to gain a deeper understanding of the relationships between the senses and second that we must identify ways of incorporating prior perceptual knowledge in the design of user interfaces in order to improve the human computer dialogue. Previous studies that investigated perceptual feature correspondence between the different senses have shown that there is consistency in the feature correlates which subjects perceive as a good match (Eitan & Timmers, 2010; Evans & Treisman, 2010; Kostas Giannakis & Smith, 1993; Kohn & Eitan, 2009; Kussner & Leech-Wilkinson, 2013; Küssner, 2014; Lipscomb & Kim, 2004; Marks, 1989; Rusconi *et al.*, 2006; Walker, 1987). This consistency is rather encouraging, but more empirical work would help devise an empirical framework for the association between cross-modal sensory cues and support designers in the process of designing multimodal interfaces.

Although we are still at the early stages of understanding the mechanisms that mediate multimodal aspects of perception, we have adequate evidence to support that cross-modal binding and interactions between the different sensory modalities occur. Research findings have shown that for a successful multimodal binding to occur, the causal relationship between two modal objects and their attributes must be plausible in terms of: (i) prior experience of similar events and phenomena, (ii) time (i.e. synchrony), and (iii) space (i.e. collocation), (Kubovy & Van Valkenburg, 2001; Parise & Spence, 2013; Schutz & Kubovy, 2009; Spence, 2007, 2011). Furthermore it has been argued that when spatial and temporal alignment and a plausible causal relationship exist between an auditory and a visual object, then the two phenomena collapse into a single percept (e.g. audio-visual speech perception), (Armontrout, Schutz, & Kubovy, 2009; Kubovy & Schutz, 2010; Kubovy & Van Valkenburg, 2001; Schutz & Kubovy, 2009). Causality is a concept that deserves more consideration, as in my view it is of great importance in the context of designing multimodal interfaces, and information displays. Plausible common cause implies that the phenomenon of binding can occur as long as the association between two modal phenomena appears realistic according to prior perceptual knowledge. Conversely, by enacting prior perceptual knowledge and applying it in the design of cross-modal mappings, it should be possible to create cognitively useful cross-modal associations, and improve the human computer dialogue in multimedia environments.

Research findings show that multimodal associations can occur automatically when we are exposed to sensory stimuli, (Barsalou, 1999; Gentner & Markman, 1997; Glenberg, 1997; Medin, Goldstone, & Gentner, 1993). Even when the cause or the source of the stimuli cannot be attributed to any previously experienced cause or source (i.e. an action, an event and/or an object), similar causal relationships experienced in the past are re-appropriated, causing involuntarily associations to potential sources and/or motor actions that could have caused the stimuli, (Barsalou, 1999; Bertelson & de Gelder, 2004; Lakoff & Johnson, 1980; Schutz & Kubovy, 2009). These re-appropriation of past experience enable us to interpret and understand phenomena we have not perceived in the past, in terms of phenomena that have been previously experienced. However the underlying principles that mediate cross-modal binding and congruency effects particularly beyond spatio-temporal integration are poorly understood.

The aim of this research is to integrate and elaborate on current empirical evidence and cognitive theories related to auditory and visual binding and metaphoric congruency and utilise this knowledge to inform audio-visual associations for the control of corpus-based concatenative synthesis for creative applications.

1.2 Thesis statement

This thesis focuses on the cognitive aspects of the human computer interaction in the context of feature based sound synthesis.

My thesis statement is the following:
Cross-modal correspondences can form an adequate empirical framework for the design of multimodal mappings that successfully align with prior perceptual knowledge. This alignment can significantly improve the human computer dialogue and the analytical, compositional and pedagogical value of user interfaces for computer-based sound synthesis.

The present research engages with the problem of understanding the principles that mediate cross-modal congruency and the application of computational methods for modelling sensory-motor correspondence. The main motivation for this investigation is (i) to develop appropriate models of interaction for efficient exploration of the audio corpus, and (ii) the development of perceptually meaningful mappings to enable practitioners to create novel sounds by specifying the perceptual characteristic of the sound that they want to synthesise.

This research has led to the implementation of Morpheme, a prototype for visual control of concatenative sound synthesis. Morpheme allows the control of a concatenative synthesis parameters through the act of sketching on a digital canvas. Morpheme is (to my knowledge) the first attempt ever made to use sketching as a model of interaction for concatenative synthesis. Furthermore an important aspect of this thesis is the elaboration on previous empirical findings from the field of cognitive psychology and the application of this knowledge to inform the multimodal mapping for musical interaction. This thesis utilised two methods originally used in experimental psychology (pairwise similarity judgments/ratings, and discrimination tasks) for the evaluation of the comprehensibility and effectiveness of multidimensional audio-visual mappings for visually driven concatenative sound synthesis. These methods could be applied to a wide range of contexts in order to inform the design of cognitively useful multi-modal interfaces, representation and rendering of multimodal data. The present research supports that to further utilise the knowledge of cross-modal correspondences in the design of multimodal interfaces, it is necessary to: (i) understand better the principles that mediate cross-modal binding and congruency effects, (ii) investigate the correspondence between multidimensional features such as auditory and visual texture, (iii) investigate how perceived congruency is affected in multidimensional contexts, such as the effects of presence of non-correlated features, and (iv) develop computational methods for modelling sensorimotor correspondences.

1.3 Motivations

The discussion about what is a sensory modality and what criteria can be applied to understand the similarities as well as the differences between the senses can be traced back to antiquity. Furthermore, the link between vision and sound has a long history in Western art and is handled by artists and musicians intuitively since the beginning of the 20th century. In recent years, there has been a growing interest in the study of multimodal interaction in the context of sound and music technology and practice. Research findings show that the mapping between the user sensorimotor input to sound synthesis and musical parameters can affect a number of cognitive and experiential aspects of interaction with sound synthesis as well as the expressivity that derives from the interaction (Hunt & Kirk, 1999; Hunt, Wanderley, & Kirk, 2000).

Mappings could be understood as a mediator that links a corporeal articulation with the production of a sound. Hence, mapping is a central element in the design of a music system and a determinant factor of the creative potential of the system. Although in the digital domain we have a lots of freedom in determining how the performer’s sensorimotor inputs are associated to the sound synthesis or music parameters, it is exactly this freedom that often makes the performer feel confused about the affordances of the interface and affect the appreciation and level of control over the sound produced by the systems. In order to solve this problem we should design interfaces where the mediator becomes transparent, creating an illusion of non-mediation as suggested by Leman (2008). It could be argued that the only way to achieve the impression of non-mediation is to design interfaces that conform to prior perceptual knowledge.

Researchers in psychology have systematically studied the correspondence between cross-modal sensory cues. The findings from these studies provide evidence that different individuals exhibit common patterns of perceived congruency between specific complimentary cross-modal feature dimensions and that common congruency patterns exist across different ethnic groups. This thesis supports that cross-modal correspondences could be used to inform the design of digital cross-modal mappings. However, to date no research has been carried out to test this hypothesis. The present research aimed to close this gap by creating and testing two audio-visual mappings based on empirically validated correspondences.

Furthermore, a number of experimental methods have been developed over the years for studying cross-modal correspondence mainly in the field of experimental/cognitive psychology, as it will be discuss in Chapter 2 and 3. Only one of these methods has been adopted in the fields of musicology and computer music to study sensorimotor responses to musical and other auditory stimuli. The method that has been most commonly used in music research is *sound tracing task,* a method that was first used in mental imaging research. In free tracing experiments participants are asked to perform a sensorimotor response (e.g. a gesticulation, draw a sketch) to an audio stimulus. However, there is a plethora of other methods that have been used over the years to study the correspondence between cross-modal features as well as the perceived similarity of physical structure and the selection criteria humans use when making similarity judgements. This thesis set out to investigate whether these methods could be useful in the context of design and evaluating cross-modal mappings for musical interaction.

Previous research findings suggest that the degree of perceived crossmodal correspondence between perceptual cues can be affected in dynamic multidimensional contexts. Eitan suggests that the relationships between cross-modal correspondences can be affected by at least three factors the type of the stimuli (i.e. static-dynamic), the interactions between simultaneous variation of multiple feature dimensions. Moreover, research findings show that corresponding features of one modality can sometimes match equally well more than one feature of different sensory modalities (Eitan, 2013). For example, as discussed in Chapter 3, it has been observed that visual size corresponds equally well to both pitch and loudness, while pitch corresponds well to both vertical position and lightness.

Although previous studies have explored perceived correspondence between visual and auditory parameters (as it has been discussed in Chapter3) and the literature shows that interactions between different dimensions of auditory and visual parameters in multidimensional context can occur (Eitan, 2013), no previous study has examined how the degree of perceived correspondence is affected by the harmonicity of the audio that is used to render the A/V feature associations. Further, the audio stimuli that were used for testing A/V correspondence by previous studies comprised simple synthetic stimuli such as synthesized sine tones, while sounds in nature consist of a multiple feature dimensions. Hence, it could be argued that there are evident ecological validity questions regarding the findings of most studies of cross-modal correspondence and that follow-up experiments are required in order to assess how these findings generalise and whether the knowledge could be used in applied contexts, such as in the design of new interfaces for musical expression and information displays.

In the context of this research, it is extremely important to answer these questions as corpus-based concatenative sound synthesis (the method of sound synthesis used by *Morpheme*) uses pre-recorded audio material in order to synthesise sound that can have complex timbre characteristics. In order to examine how the characteristics of timbre of the sounds used to render a complementary polar features affect the degree of perceived similarity, a pairwise similarity test was designed where subjects were presented with a series of audio-visual stimuli and rated the perceived similarly for each pair.

1.4 Summary of findings

The present research confirms the results revealed by previous studies (Eitan & Timmers, 2010; Evans & Treisman, 2010; Kostas Giannakis & Smith, 1993; Kohn & Eitan, 2009; Kussner & Leech-Wilkinson, 2013; Küssner, 2014; Lipscomb & Kim, 2004; Marks, 1989; Rusconi et al., 2006; R. Walker, 1987) which found strong relationships between the audio-visual feature associations of size – loudness, vertical position-pitch, color brightness– spectral brightness. Weaker were the relationships between texture granularity – sound dissonance and color complexity- sound dissonance, similar to the findings of Giannakis (2006). The weak correspondence reported by the subjects between these features of the auditory and the visual stimuli suggest that further research will be required to investigate the correspondence between auditory and visual textures.

The findings from the first experiment suggest that the strength of the perceived correspondence between the A/V associations prevails over the timbre characteristics of the sounds used to sonify the visual features. Furthermore, the findings show that there is a non-linear interaction between the harmonicity of the sound and the perceived correspondence of the audio-visual associations. For example, strongly correlated cross-modal cues such as size-loudness or vertical position-pitch are affected less by the harmonicity of the audio corpus in comparison to weaker correlated dimensions (e.g. texture granularity-sound dissonance).

However, the findings from the second experiment suggest that when multiple parameters are controlled simultaneously, the harmonicity of the sound has a stronger effect. For example, when the corpus consisted of harmonic sounds, participants' detection rate was significantly increased in comparison to when using none harmonic sounds. Hence, the salience and efficacy of the cross-modal associations of a multidimensional mapping is affected by the harmonicity of source audio which is used to sonify the visual features. Moreover, the level of difficulty of the task influences the effect of the harmonicity of the corpus. With harmonicity having a stronger effect the more cognitively demanding a tasks is. Finally, results suggest that sound/musical training had no significant effect on the perceived similarity between A/V features or the discrimination ability of the subjects.

1.5 Thesis Overview

Chapter 2, reviews how theoretical and technological developments in the domain of music have challenged the analytical and representational value of traditional music notation and have led researcher to seek for new forms of sound representation. Furthermore this chapter reviews the state of the art in graphical interaction and discuss issues related to user interaction with sound synthesis and signal processing tools.

Chapter 3, reviews existing research related to the question of cross-modal correspondence. This chapter aims to help the readers understand the problem of cross-modal correspondence between auditory and visual sensory cues and discuss the state of the art. Further reviews on audio-visual correspondence have resulted in the design of two multidimensional mappings used in the present prototype and have contributed in the formation of my empirical investigation.

Chapter 4 presents the interface that was developed in the context of the present research, named Morpheme. This chapter explains the motivations for developing the system. Further, it provides implementation details about the user interface, the system architecture, and the mappings developed for visually querying the sound database. Additionally, this chapter describes a number of algorithm that were developed for visual feature extraction and for improving user exploration of the audio corpus, see 4.6.6 and 4.6.7.

Chapter 5 presents the experimental design and the findings from the first empirical study that was conducted in order to investigate the level of perceived correspondence between six audio-visual associations and the effects of the harmonicity of the audio corpus on the perceived congruency of the associations. The audio-visual stimuli were prepared using the Morpheme interface, which is presented in chapter 4.

Chapter 6 presents the experimental design and the findings from the second empirical study that was conducted in order to investigate the effects of the audio-visual mapping and the harmonicity of the audio corpus on the ability of the participants to discriminate between a series of audio-visual stimuli. Similarly to the previous experiment the stimuli in this experiment was prepared using the Morpheme interface.

Chapter 7, presents a user study that aimed to detect usability issues of the Morpheme interface and gather participants' opinions regarding cognitive, experiential and expressive aspects of the interaction with the interface developed in this study (i.e. Morpheme).

Chapter 8, makes concluding remarks about the scientific contributions of my thesis and presents suggestions for further work following on from the research presented in this thesis.

2 Visual Representation of Sound

2.1 Visual Representation of Sound: technological and theoretical developments

Music representation and its history can be traced back to ancient Iraq, Turkey and Greece. The oldest representations of music surviving today such as Seikilos Epitaph and the Delphic Hymns were used to annotate pitch, note duration and to some degree harmonic relationship between musical structures (Christensen, 2002). Music notation as a tool has enabled composers to conceptualise musical ideas and develop complex musical structures for multiple instruments. Sound symbolism could preserve composers' ideas in time, help instrument players to perform a musical piece, but also be used as a mechanism to assess musical performance (Karkoschka, 1973). For the majority of musicians a music score is not an accurate representation of the music. It is a rather loose representation that leaves space for interpretation by the musicians (Patton, 2007). The long history of visual representation of music suggests that music as a compositional process and as performing art, has a strong visual element associated with its inception and reproduction. However it should be mentioned that some musical cultures never used any written form of representation (i.e. free improvisation, other tribal musical cultures), but even in this cases musicians most likely than not rely on visual communication for synchronization.

The development of analogue and digital technologies for capturing, synthesizing and manipulating sound, has significantly transformed the way we compose and perform music as well as our aural culture in general. Musicians have long been experimenting with electronic instruments and pre-recorded sound to create music (e.g. music concrete, electroacoustic, electronica, and other pop), (Emmerson, 2007). Advances in technology have led to theoretical developments in musicology, as now the art of music and sound are no longer limited to the sound palette provided by the material and acoustic properties of music instruments. Soon after these developments took place question have risen about the approaches and language to be used to notate these new musical styles, which consists of such a diverse range of sounds with complex timbre qualities and relationships between them. As Thoresen & Hedman, (2007) discusses

"in order to be able to articulate and discuss these new experiences, there is need for a corresponding novel set of terms and concepts".

These developments have driven researchers to seek for new ways to represent sound for notational amongst other applications. Traditional music notation represents only a very limited set of perceptual aspects of sonic events, and it has been argued that it might not be the best approach for contemporary music analysis and comprehension (Blackburn, 2011; Emerson, 1986; Patton, 2007; Smalley, 1986; Wishart, 1986). One of the first to propose alternative forms of notation system in attempt to address these issues was Pierre Schaeffer in the 1960s, who was also the founding father of what is known today as music concrete (i.e. composing music solely by the manipulation of pre-recorder sounds captured from the natural world). Thoresen & Hedman, (2007) explains that what Schaeffer attempted to do was

"to address the question of correlation between the world of acoustics and engineering to that of the listener" (ibid: p.1).

Schaeffer's new form of sound notation known as typomorphology, sought to graphically illustrate experiential and perceptual aspects of sound (ibid: 1). Schaeffer's work on typomorphology involved a large number of symbols and explanations about the function of these symbols for conceptualization, notation and analysis of perceptual sound qualities. Figure 1 shows a minimal extract of Schaeffer's categorization for sound, were the y-axis represents sound spectrum and the x-axis energy articulation.

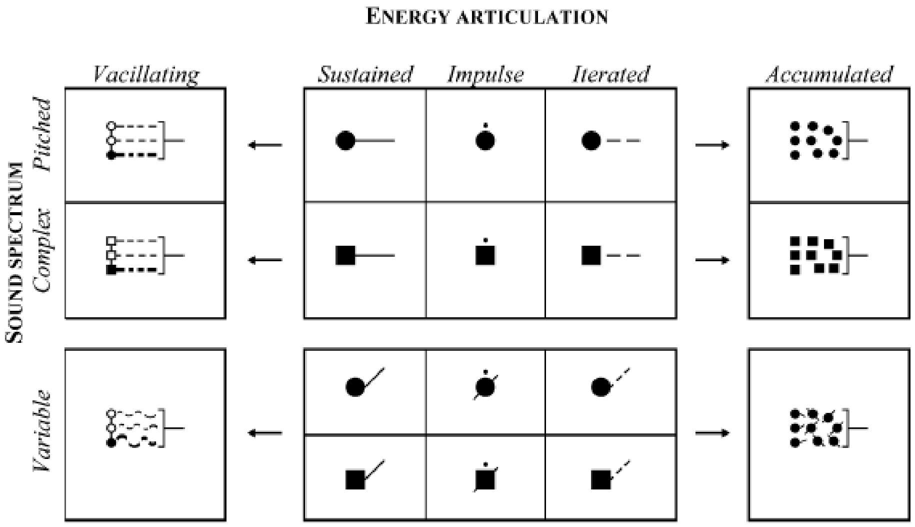


Figure 1. Example of Schaeffer's Typomorphology (Thoresen & Hedman, 2007).

Schaeffer distinguished between three major spectral categories pitched, complex and variable sounds (e.g. notes, complex unpitched, evolving sound objects). Moreover he distinguishes between five types of energy articulation: vacillating, sustained, impulse, iterated, and accumulated. All five minimal categories describe a potential energy motion articulation of sound.

Other composer have experimented with conceptual design of notation systems for representation of sound and as a tool for sound analysis, conceptualization and music performance (Barry, 1977; Climent, 2001; Smalley, 1986, 1997; Thoresen & Hedman, 2007; Thoresen, 2010). Stockhausen developed a set of graphical symbols to notate his piece Plus Minus (Stockhausen, 1963), see Figure 2. Stockhausen purposely developed this score with symbolic notations which are rather vague in nature and provided several pages of text interpretation of the semantics and functions attached to each symbol (Fox, 2000). Additionally several pages of conceptual instructions related to his rationale were also provided. However he did not provide any specific information regarding either instrumentation of the piece or the duration. He left this open to interpretation of the performers. Each symbols defined how to manipulate materials or instruments to produce sonorous effects or sonic gestures and each square represented a moment of the piece, a sonic pattern that it was repeated for as long as the performers wanted to. Stockhausen’s intention was to provide an alternative symbolic system for music improvisation and conceptualization of complex musical structures, textures and gestures rather than a descriptive type of graphical notation system to represent the perceptual qualities of sound (Barry, 1977).

Stockhausens' score is particularly interesting as it proposes a new paradigm of sound representation that is in comparison to traditional forms of music notation is less deterministic in nature. This offers more space for the performers' creative expression, which in turn could lead to the emergence of sonically interesting results and ideas. Additionally it could be argued that this paradigm suggests models of interaction that could have applications in contemporary graphical interaction design for music composition and performance.

The shift from strict forms of notation to more loose forms of musical representation can be observed since the middle of the last century. These new forms of musical scores extended the freedom as well as the responsibility to the performers, which in turn made the element of chance more central in the performance of the notated music pieces (Baveli & Georgaki, 1994). According to Logothetis:

“*What fundamentally differentiates graphic notation from traditional notation is the aforementioned polymorphism, which clearly enables all performers to retain their subjective reaction times. The composer takes into consideration the divergences of the different performers in composing and expects a certain degree of surprise through the new formalization of musical form in every performance*.” (ibid: p.1).

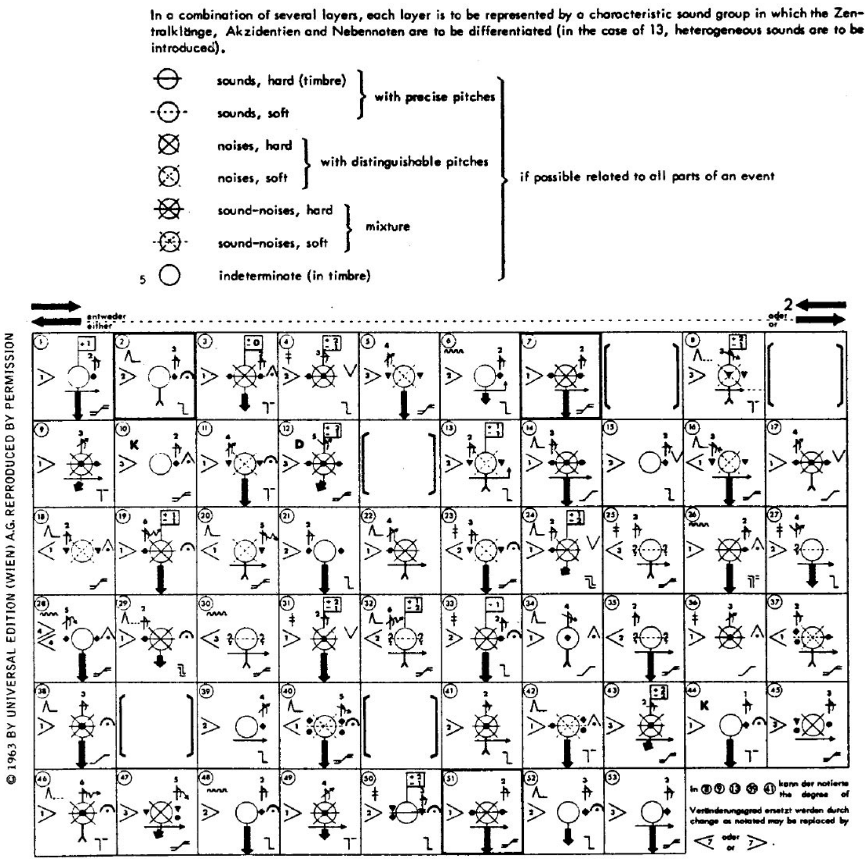

In a combination of several layers, each layer is to be represented by a characteristic sound group in which the Zentralklänge, Akzidentien and Nebennoten are to be differentiated (in the case of 13, heterogeneous sounds are to be introduced).
sounds, hard (timbre)
sounds, soft
with precise pitches
noises, hard
noises, soft
with distinguishable pitches
sound-noises, hard
sound-noises, soft
mixture
5
indeterminate (in timbre)
if possible related to all parts of an event
entweder
either
oder
or
2
K
D
in ⑧ ⑨ ⑬ ㊴ ㊶ kann der notierte Veränderungsgrad ersetzt werden durch
the degree of change as notated may be replaced by
oder
or
© 1963 BY UNIVERSAL EDITION (WIEN) A.G. REPRODUCED BY PERMISSION

Figure 2. Stauckhausen's Plus and Minus score (Levin, 2000)

Logothetis proposed his own form of graphical notation system, see Figure 3. The first type of symbols seen in the left side of Figure 3 was used to annotate pitch. The second category of symbols seen in the centre of the figure represented loudness, changes in timbre and duration. The third type of symbols shown in the right of the figure represented actions. A difference between Stockhausen's annotation system and that of Logothetis or Schaeffer's is that Stockhausen's annotation system is more symbolic, while the representations used in Schaeffer's system are more iconic. As Couprie discusses, iconic representations usually link perceptual/structural qualities of a sound to perceptual qualities of a graphical representation. Therefore iconic representations can be relatively easy to read and comprehend (see example in Figure 4). On the other hand symbolic representation methods can provide a well coded system that can support significant analytical accuracy, but the symbols are usually very inaccessible and difficult to read (e.g. music notation, mathematic, language), (ibid: p.110). However the same might be the case with iconic representation when they are very precisely tied/linked to sound qualities, (i.e. high resolution, signal level representation).

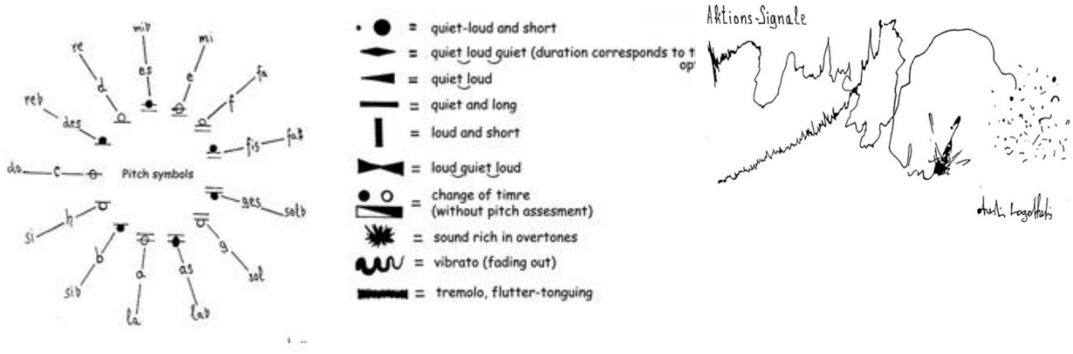


Figure 3. Graphical symbols developed by Logothetis to annotate his musical compositions (Baveli & Georgaki, 1994).

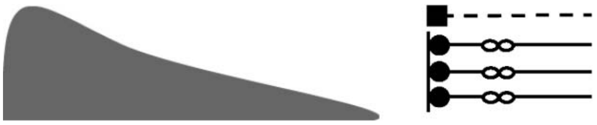


Figure 4. Iconic and Symbolic representation of sound (Couprie, 2004)

Due to the fact that iconic representations are easier to read and comprehend because they relate perceptual qualities between the sound and the representation, it could be argued that iconic representations are particularly useful for pedagogical purposes. Because the representations are perceptually meaningful, their ability to aid comprehension and analysis of the musical material is greater in comparison to symbolic representations. For instance, when a person is listening to sounds while seeing the sounds being illustrated by iconic means, perceptual links are established between sound and image that might have not been established otherwise. This in turn might help learners assimilate information quicker and better. So defining a framework for visual representation of sound is very important for pedagogical, analytical and compositional purposes (Blackburn, 2009; Couprie, 2004; Patton, 2007). Smalley discusses that descriptive and conceptual tools which can relate and organize sounds and structures could be a valuable compositional aid (Smalley, 1997). He points out that the importance lies in forms of sound notation which are more concerned with "spectral qualities than actual notes, more concerned with varieties of motion and flexible fluctuation in time than metrical time, more concerned with sounds whose source are mysterious or ambiguous rather than blatantly obvious", (ibid: 109). Furthermore Smalley argued that the focus should be in the representation of intrinsic qualities of the sound instead of the technology. Smalley has proposed a set of descriptive tools and a theoretical framework known as spectromorphology. This framework is meant to be used for analysis of sound based on aural perception. Spectromorphology as he says:

"is concerned with perceiving and thinking in terms of spectral energies and shapes in space, their behaviour, their motion growth process and their function in musical context" (ibid: 125).

Two software implementations developed around the idea of spectromorphology have been proposed (see Figure 5 and Figure 6). Figure 5 demonstrates a two dimensional analytical and publication tool for developing scores based on perceptual qualities of the sound. Qualities of pitch, dynamic shape, duration, spectral thickness and texture are illustrated (Couprie, 2004).

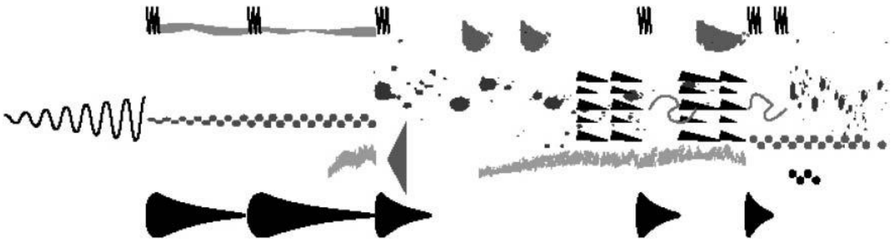


Figure 5. Ionic representation of Reflet by Ivo Malec (0'00"–1'00") (Couprie, 2004).

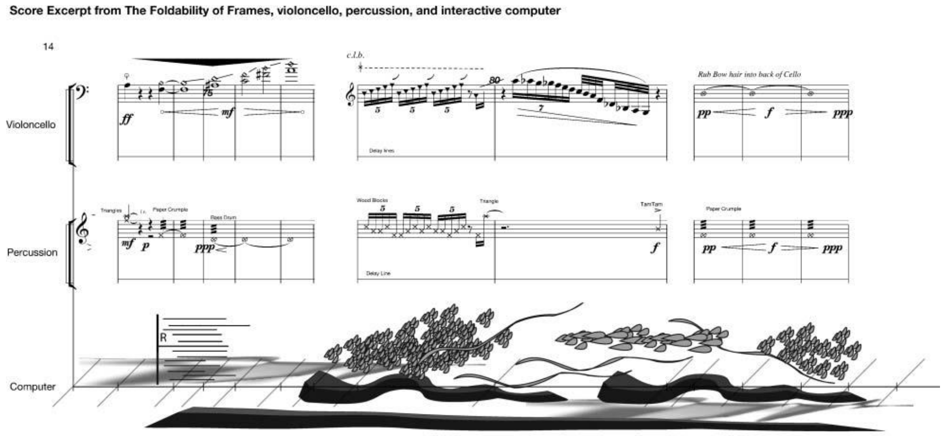

Score Excerpt from The Foldability of Frames, violoncello, percussion, and interactive computer
14
Violoncello
Percussion
Computer
c.l.b.
Rub Bow hair into back of Cello
ff
mf
pp
f
ppp
Triangles
Paper Crumple
Bass Drum
Wood Blocks
Triangle
TamTam
Delay lines
Delay Line
R

Figure 6. Score example from The Foldability of Frames (Patton, 2007).

Figure 6 shows a three dimensional visual morphological notation tool developed by (Patton, 2007). This application focuses on the representations of parabolic motion, linear motion, and granular clusters. Both examples shown above follow the well-established time/ pitch paradigm however Patton focuses more on spatial and morphological aspects of the sound while Coupries' approach is more descriptive of timbre and textural perceptual aspects of sound.

Figure 7, illustrates two examples of visual representations developed by Coupries' analysis of sound based on Smalley's spectromorphological framework. To the right of the representations the criteria used to create the representations are indicated. Each example is based on perceptual analysis of the sound which was done manually by listening to the sonic material, and through the analysis of a sonogram using an iterative segmentation process. The two examples in Figure 7 illustrate two extremes cases in terms of sound qualities, object 1 represents a note, while object 2 a noise. It should also be noted that each sound uses different set of criteria that defines it as an object. So the vertical location and the area occupied by object 2 represents spectral thickness rather than pitch and dynamics, which is the case with object 1.

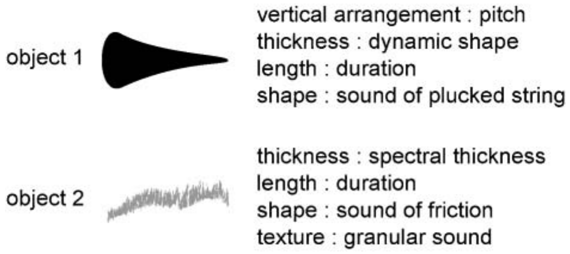

object 1
vertical arrangement : pitch
thickness : dynamic shape
length : duration
shape : sound of plucked string
object 2
thickness : spectral thickness
length : duration
shape : sound of friction
texture : granular sound

Figure 7. Two dimensional iconic representation of sound indicating relations between image and sound (Couprie, 2004).

Figure 8, illustrates a three dimensional representation of sound were temporal dimension is represented across the x-axis. This example uses both y-axis and z-axis to illustrate two kinds of frequency related information: pitch, and spectra. The fundamental frequency is represented by the vertical location of the object in the three dimensional space as in traditional music notation. However it should be noted that unlike conventional music notation Patton's representation permits fluctuation of the fundamental frequency in a non-metric time. So fluctuation along the y-axis of the 3D space illustrates pitch evolution or variations of frequency over time. Patton points out that noisy sound can challenge the idea of pitch continuum because of the discontinuities apparent in the frequencies spectrum of noisy sounds (e.g. white noise consist of sound particles which are discontinues and scattered over the entire frequency spectrum). The spread of the curve representation along the y-axis of the three dimensional space indicates the overall register occupied by the sound, which affect the perceived width of the sound and it can also vary over time. Patton explains that a tuned viola note will be perceived as being much thinner than a piano chord or an out of tune trombone section (see Figure 3.11). Moreover the spread of the object along the z-axis represents the range of spectral variance, so the bigger the range of spectral variance of the sound the wider it will appear along the z-axis. The object's fluctuation along the z-axis represents the degree of harmonicity. Sounds that consist of a high degree of harmonic frequencies will tend to move further away along the z-axis while non-harmonic sounds will appear closer.

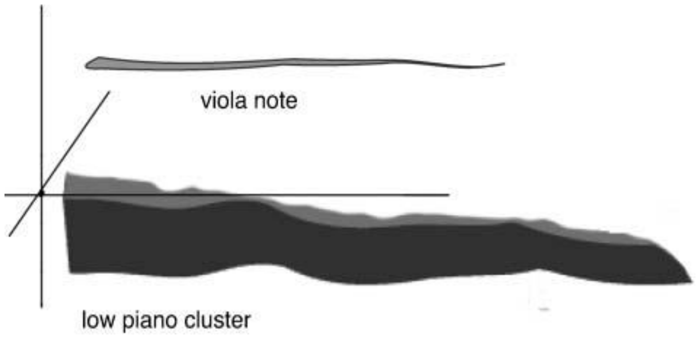


Figure 8. A three dimensional iconic representation of two sounds (Patton, 2007).

The view that defining a framework for visual representation of sound is very important for pedagogical, analytical and compositional purposes, is confirmed by recent studies in perceptual learning. There is adequate empirical evidence to claim that multisensory stimulation during learning can provide a more natural and effective learning setting, as most sensory phenomena in nature are multimodal and therefore it is likely that the human brain has evolved to develop, learn and operate optimally in multisensory settings (Shams & Seitz, 2008). It has been argued that the integration of multisensory information is essential to construct a meaningful representation of phenomena in the environment. Furthermore, the ability to construct a meaningful internal representation of the environment, actually depends on integrating and segregating between multisensory stimuli received from the environment (Shimojo, 2001; Brandwein *et al.*, 2011). When listening to a sound stimulus while seeing a visual representation that links perceptual qualities of the sound to visual parameters, the multisensory stimulation that is caused (i.e. auditory and visual) will activate larger sets of cognitive structures and therefore lead to better internal representations (Shams & Seitz, 2008), see Figure 9.

According to Goldstone, (1998) perceptual learning involves mainly four mechanisms: attention weighting, imprinting, differentiation, and unitization. By attention weighting, perception becomes adapted to tasks and environments by increasing the attention paid to important dimensions and features. By imprinting, receptors are developed that are specialized for stimuli or parts of stimuli. By differentiation, stimuli that were once indistinguishable become psychologically separated. By unitization, tasks that originally required detection of several parts are accomplished by detecting a single constructed unit representing a complex configuration.

However the relative weight of each of these variables in a skill acquisition context has not been determined yet, neither has it been attempted to map the components of learning to the specific interface features in the context of computer aided musical composition. Hence, it could be argued that future research efforts in designing music learning and creative applications should strive to identify interface design frameworks that utilise these learning modes.

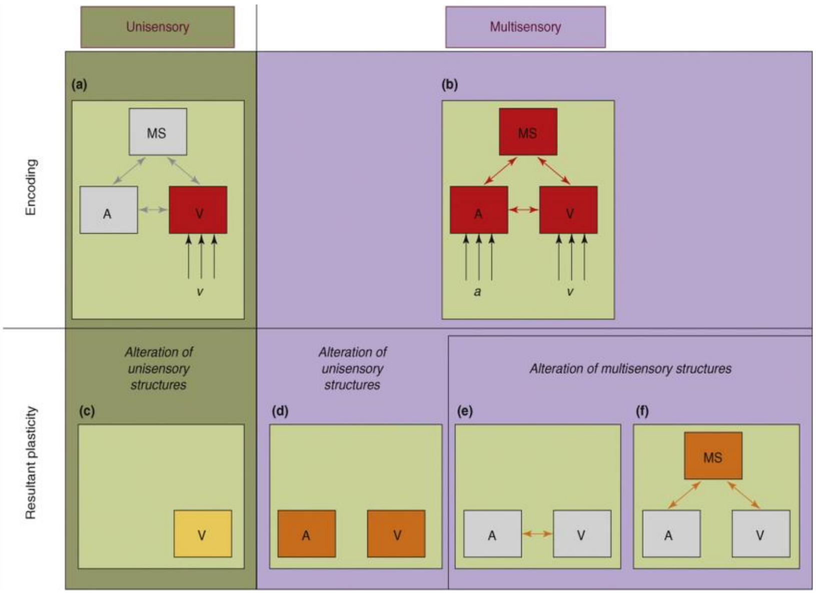

Unisensory
Multisensory
Encoding
Resultant plasticity
(a)
(b)
MS
A
V
a
v
Alteration of unisensory structures
Alteration of unisensory structures
Alteration of multisensory structures
(c)
(d)
(e)
(f)

Figure 9. Cognitive structure activation during learning in unimodal and multimodal contexts (Shams & Seitz, 2008).

Another aspect of traditional music notation that was challenged by the new graphic systems for notation of sound and musical composition was the temporal trajectory/evolution of a musical piece. Both the ideas of metrical time as well as the linear temporal trajectory suggested by traditional music notation have been challenged (Baveli & Georgaki, 1994). According to Logothetis the polymorphism of graphical notation has both to do with space and with the method by which it is read.

"*Traditional notation is divided into systems and is read from left to right, like books. But since sound does not behave in the way written word does, we could think about using pictorial notation to represent musical events. (...)because musical time doesn't follow any direction, let alone the conventional left to right writing found in literary forms" (ibid: p.2).*

Smalley (1997), distinguishes between three common types of scores or sketches and conceptual maps used in electroacoustic music that might contain information which is relevant to perception and used for conceptual and analytic purposes. The first type is used by a performer as a guide to be followed during a performance of live electronics or mixed works, and it usually contains graphic representations of acousmatic materials. The second type is a conceptual score usually produced to explain mainly technical aspect of how a piece has been created/conceived and it illustrates in some form the overall sonic structure and musical content of a work. The third type of score is diffusion maps used to notate spatial distribution of audio signal across an array of loudspeakers. Musical sketches unlike architectural and product design sketches can be very idiosyncratic and vague in nature (Thiebaut, Bello, & Schwarz, 2007). There is evidence supporting that benefits related to creativity might arise exactly from the ambiguity of these artefacts/representations, see (Goldschmidt, 1991; Neilson & Lee, 1994).

Further to the three types of scores suggested by Smalley, we could also distinguish between different typologies of scores based on the way the score is presented and read by the performer. Vickery (2014) made an extensive survey of contemporary approaches to computer screen based scores for live performance and suggested the following classification, segmented scores, scrolling scores, three dimensional scores and animated scores, see Figure 10. Segmented screen-scores resembles tradition sheet music notation, where music segments of a musical piece are presented sequentially. Rhizomatic and three dimensional and scrolling scores paradigms often employ the technique of moving the score from a fixed point on the screen that is the performers fixation point. Animated score paradigms often use an animated graphic that represent the time overlaid a more fixed structure indicating metrical note or music sequence trigger points. The performer musical sequence occur when the cursor crosses the predefined trigger points.

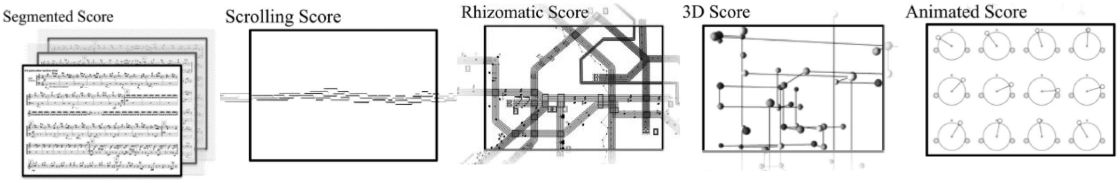


Figure 10. Five approaches for screen based music score presentation (Vickery, 2014).

Although several researchers have engaged in the development of a commonly accepted notation system or a framework for sound representation for composition and analysis of electroacoustic music, there is still no common consensus (Couprie, 2004). It could be argued that the only way to form a commonly acceptable system of sound representation is by relying on empirical founded correspondences between auditory and visual perceptual cues rather than idiosyncratic ones. As it will be discussed in chapter 3, evidence suggests that humans naturally perceive correspondences between a number of auditory and visual feature dimensions. Hence, I propose that in order to reach a common consensus about the visual representation of sound feature it is necessary to for a framework of correspondence underpinned by empirical evidence rather than based our personal choices, social conventions and technologically determined factors. Of course contextual factors such as the technology used, the application of the technology as well as the level of expertise of the user can be informative in the context of audio-visual interaction for music and sound design applications, however the focus should be primarily in the representation of perceptual qualities of the sound instead of contextual factors.

2.2 Common types of graphical representation in audio applications

According to Vinet (1999) we distinguish four major categories under which we could classify most graphical representation of sound in audio applications these are:

- Graphical User Interface (GUI) Objects, visual metaphors used to control sound parameters (i.e. knob, sliders, buttons etc.).
- Temporal representation revealing two or three variable dimensions linked to time such as spectra, amplitude (i.e. Waveform, 2D or 3D Spectrogram, Oscilloscope).
- Representation used for complex data comprehension and management such as pattern recognition, pattern similarity, classification (e.g. data visualization, multidimensional scaling, hierarchies).
- Visual programming representation enables the use of simple elementary modules that can be assembled together to represent more complex procedures, calculation and functions (e.g. Max MSP, PD, Reaktor etc.)

In this section, techniques for temporal, spatial and complex data representations of sound signals are analysed. Additionally an extensive analysis of GUI design metaphors for the control of sound parameters/processes is presented here.

2.2.1 Temporal representations of sound signals

Due to the temporal dependency of sound most graphical signal representations will illustrate time as a variable in some way. Examples of temporal representations of sound can be found in many applications with the most common one, the waveform representation that illustrates air pressure waves over time and it can provide information about dynamics/pressure, fluctuations and duration of a sound, see Figure 11. A sonogram representation of the same is presented in

Figure 12. The horizontal axis represents time, and the vertical axis frequency, while the colour indicates the amplitude of the particular frequency, the brightest the colour the higher the amplitude. Moreover many variations in the implementation of spectrograms can be found nowadays among other temporal representation of sound signals such as sonogram, oscilloscope, phasescope.

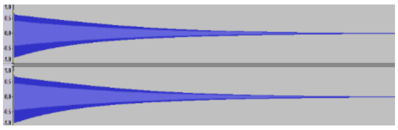


Figure 11. Waveform example.

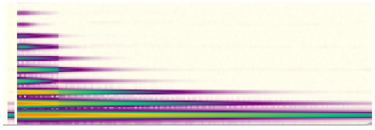


Figure 12. Spectrogram example.

The type of representations are used mainly for comprehension of the auditory information. As it will be discussed in the following section, analysis and representation of sound can be used to enable direct and intuitive interaction with sound parameters through direct manipulation of graphical representations.

2.2.2 Temporal representation and metaphor controls

Figure 13 shows some common examples of metaphor representation for the control of sound parameters used by most commercial audio workstations. As Vinet, (1999) discusses, the concept that underpins these interfaces is enabling the parallel placement of sound and sound related temporal information. While the x-axis in all of the examples shown Figure 13 is used to represent time the y-axis is used to represents various parameter, depending on the application and the context. For example at the top left corner in Figure 13 we can see an example of a sequencer for audio and midi data. The y-axis here is used to indicate the channel to which the audio or midi data belong to. This interface metaphor is based on the principle of multi-track mixing console were each instrument has its own channel for recording or playback, (ibid: pp. 3). Below, the sequencer picture illustrates a midi editor where the y-axis is for registering and representing pitch related information. On the right of Figure 13, several examples are given of other graphical approaches using the x-y coordinate system for the representation of temporal related audio data and /or parameters. In these examples the y-axis is used to represent numerical values that control some sound parameters.

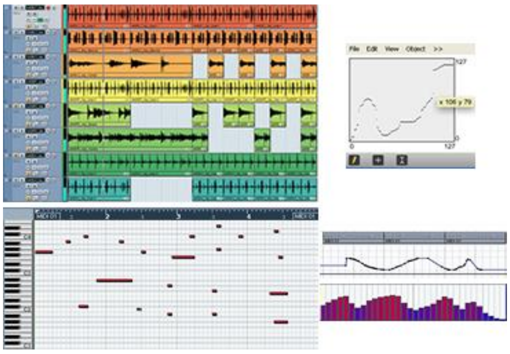


Figure 13. Examples of common graphical user interfaces applications that enable temporal placement and control of sound parameters. Top left: multi-track audio representation. Bottom left: piano roll, etc.

2.2.3 Temporal representation of sound signals & direct controls

Several audio analysis paradigms enable direct interaction with sound parameters. Figure 14 shows the main interface of the open-source software *Spear* which enables audio spectral analysis, editing and re-synthesis. Analysis is achieved using a variation of the traditional McAulay-Quatieri technique of peak interpolation and partial tracking. Linear prediction of the partial amplitudes and frequencies is used to determine the best continuations for sinusoidal tracks (Klingbeil, 2005). Figure 15 shows a screenshot from the software Audiosculpt created by IRCAM research centre. In this example Audiosculpt enables filtering of sound by drawing directly on the sonogram, this allows the elimination or attenuation a specific frequency bands with precision (Bogaards, 2005). It is an excellent tool for sound restoration or creative editing as it gives a lot more precision and flexibility compared to conventional equalization tools. Figure 16 shows a screenshot from the software Celemony Melodyne which uses FFT analysis to identify pitch and represents it on a time to pitch grid. Melodyne allows direct manipulation of the pitch by dragging the sound tone on the grid. One difference that can be observed between the Spear's (see Figure 14) and Melodyne's graphical user interfaces is related to the amount of audio signal detail represented. Melodyne represents much less detail: only the pitch, the amplitude and the duration, which makes it simpler and more effective for correcting musical pitches such as vocals and instruments notes.

2.3 Sketch-based interfaces for sound synthesis

'Audio-visual systems' follow slightly different approach than score based system in the sense that properties of the pictorial representations are used to interact with or generate sound. Many

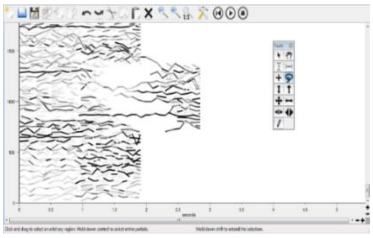


Figure 14. Spear freeware software, enables direct editing on sound spectra.

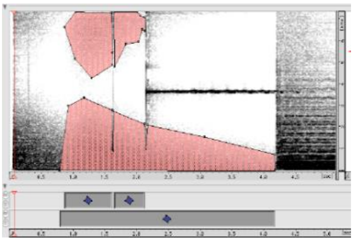


Figure 15. AudioSculpt, drawing filters on spectra.

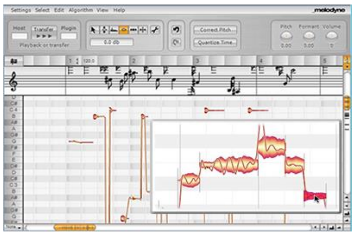


Figure 16. Melodyne pitch analysis and editing tool.

paradigms of such a system have been developed since the nineteen fifties. Visually driven systems have often be used as a compositional and sound design tool, with application in music composition, sound synthesis, signal processing and live performance. One of the first known interactive system of this type is the *UPIC* system which was the developed by Xenakis in the 1970s. Figure 17, shows the score of *Mycenae-Alpha* an electronic piece that Xenakis composed in 1978 as part of an installation of lights, movement and music. This was the first piece of electronic music to ever be created solely by the computer analysis of visual images. In these piece Xenakis was interested in the exploration of the relationship between graphic, image and sonic structure at a micro-compositional level of sound (i.e. sonic grains/particles in frequency, time and space). Xenakis developed the UPIC system that enables to scan and analyse images in order to control sound synthesis generative processes (Lohner, 1986; Nelson, 1997; Squibbs, 1996).

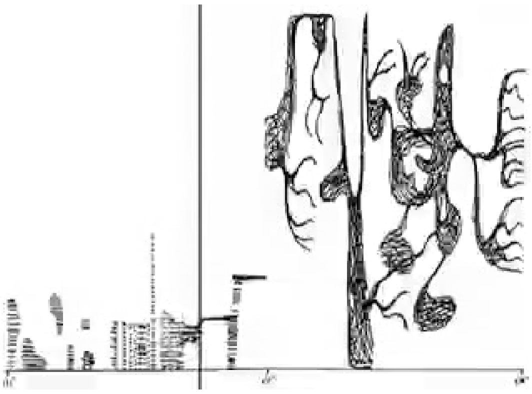


Figure 17. Iannis Xenakis Mycenae-Alpha (1978) score.

The last few decade's researches in the field of computer music have implemented several systems adopting the sketching paradigm of the UPIC system and that take advantage of these technologies to enhance the human computer dialogue and improve the means of interaction with tools for artistic expression and creativity. Figure 18, shows a screenshot from the *HighC* graphical user interface, which enables to paint sounds sequences by drawing pitch and envelopes curve tin order to control the sound processing parameters. Figure 19 shows a screenshot of the Hyperscore software system for computer-assisted composition that allows intuitive visualization and manipulation of sound (Farbood, Kaufman, Line, & Jennings, 2007). In this paradigm the sounds are mapped to the graphical elements such as shape, colour, line texture etc. The user can manipulate low level musical parameters such as pitch, dynamics but also high level such as melodic patterns, harmonic tension. Hyperscore enables the creation of musical motives and it uses the metaphor of paintbrushes to allow user to paint the musical patterns in the sequencers window. The x-axis represent time, while fluctuation along the y-axis represents/creates variations in the musical pattern based on predefined rules (Farbood et. al., 2007). The Hyperscore system was primarily designed for educational and pedagogic purposes. In overall the intention behind the design was explored graphical interfaces metaphors that follow more naturalistic approach to music composition, closer to traditional approaches (e.g. working with a music manuscript).

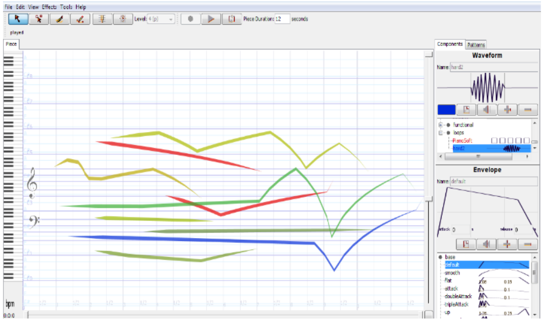


Figure 18. HighC software enables to paint sounds.

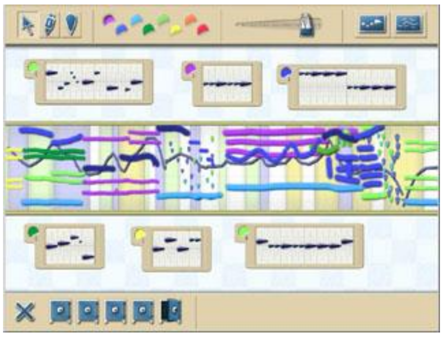


Figure 19. Hyperscore software enables to compose by drawing sounds.

Metasynth is a software solution created by U&I software that provide a range of graphical tools for sound synthesis. Metasynth[1] is using additive synthesis, spectral analysis, re-synthesis and filtering for graphical editing of the sound spectrum. Metasynth allows user control

[1] http://www.uisoftware.com/MetaSynth/index.php

of the frequency of audio oscillators using a sketching paradigm for additive synthesis. Metasynth uses line curves similar to the HighC software but images can also be used to drive the synthesis parameters. Further Metasynth and spectral manipulation. Time is represented on the x-axis and pitch on the vertical axis. The graphical user input on the horizontal axis can be quantised both in the frequency domain in order to enable graphical control of the sound frequency based on linear, logarithm and harmonic scales. Photosounder[2] is a real time application that uses images in order to control spectral processing parameters. More specifically the system works by using the pixel data of the image to control the parameters of a spectral band filter. The source audio used to synthesise the sound can derive from either a white noise generator or any other audio file. The editing tools offered by the system enable spectral editing of sound through direct manipulation of the spectrogram image representation of the sound, in many ways similar to the Spear and AudioSculpt systems which were discussed earlier.

Scrapple is augmented reality interface. Scrapple's central goal is to enable spectrographic synthesis of sound by placing various objects such as visual colored cartons cut in different shapes and morphologies on the whiteboard surface, (Levin, 2006). Scrapple uses a camera to take a snapshot of the table surface every few seconds. The image is scanned lengthwise and each object placed on the table represents a potential sound oscillator for real-time additive synthesis. The frequency of each oscillator is determined by the vertical position of the object while the darkness of the objects define the loudness of the tone. Another system developed by Levin the AVES system (Audio Visual Environment Suite) allow to synthesize in real time audio and abstract visuals through the motion of the mouse or other motion sensitive controller. Users can create, manipulate and delete the abstract visuals which are mapped and control the sound synthesis parameters. This project sought to explore the meeting point between abstract communication and interactivity as the means to derive creativity, for more information on this project see (Levin, 2005).

2.3.1 Interaction with Corpus Base Concatenative Sound Synthesis

The first attempts to synthesise sounds from pre-recorded audio were performed by Xenakis, (1971) using magnetic tape that was sliced and stitched back together to create new sound sequences. Granular synthesis was the first type of digital concatenative sound synthesis pioneered by Roads, (1978). Granular methods of sound synthesis works by temporal layering of small audio segments known as grains. By combining thousands of grains over time complex

[2] http://photosounder.com/

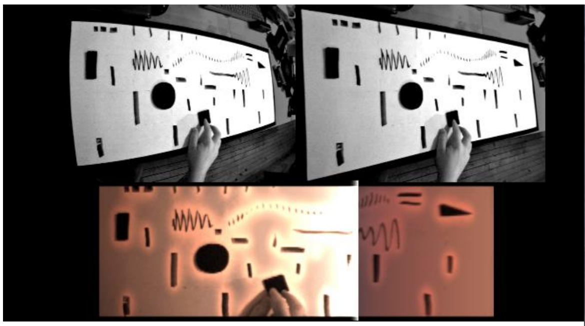


Figure 20. Scrapple in The Table is The Score: An Augmented-Reality Interface for Real-Time, Tangible, Spectrographic Performance.

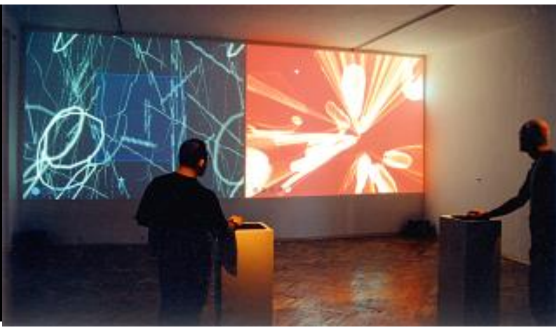


Figure 21. AVES system for real time audio and visuals synthesis.

sounds and sonic textures can be created. Grains are small pieces of audio data with duration between 10ms to 100ms. Grains are usually taken from pre-recorded audio. Amplitude envelop curves are applied to eliminate distortion artefacts caused by splicing the audio at a not at zero amplitude crossing point, (Roads, 1996, 2001). Moreover envelops are also used for shaping the grains. Granular synthesis has a long history in electroacoustic music, the most known composers that have explored granular synthesis in the context of musical composition are Roads, (1988); Truax, (1990); Xenakis, (1971). Implementations of granular synthesis provide control over a number of audio parameters such as the grains' duration, transposition, playback speed, number of grains, and the temporal location from where the grains are taken. Due to the vast number of parameters that have to be determined in order to synthesise a sound that last only for a few seconds, often implementations of granular synthesis algorithms allow the composer define the sound in global terms (i.e. setting minimum and maximum limit for each parameter) and the granular synthesis method automatically generates individual parameters for each grain using randomisation constrained by global limits defined by the composer (Roads, 2001).

A more recent development of granular synthesis known as concatenative synthesis allows the retrieval of grains from a database of pre-segmented and pre-analysed audio segments known as audio-units. According to Schwarz, (2004), we can broadly distinguish between two methods of sound synthesis: parametric synthesis and concatenative synthesis. Obviously concatenative synthesis also has signal processing parameter, so it could be thought as parametric from this point of view. However what really distinguishes concatenative synthesis from other parametric methods is that all parametric methods have an implicit way of generating signals from scratch which can be parametrised, while concatenative synthesis methods require some recorded audio source from which the sound is synthesised. Corpus-based synthesis methods were initially developed for speech synthesis and natural language processing. We could broadly distinguish between three types of corpus-based sound synthesis (i) with fixed inventory, (ii) unit-selection-based, and (iii) statistical parametric synthesis, for a review see Dutoit, (2008). Corpus-based concatenative sound synthesis methods work by recalling audio-units from a database using either distance or similarity measures to find the best match between a stream of n-dimensional feature vectors used as target and the feature vectors of pre-analysed audio-units found in the database. Feature vector in this context is a numerical vector used to represents the audio-units and the user input. Sound synthesis is accomplished by combining the audio-units that are retrieved from the database creating a new sound sequences.

Recently, corpus-based concatenative synthesis methods have been used for creative and musical applications. However as Schwarz (2005) suggests, concatenative synthesis methods have been used for in music composition long before the introduction of automatic unit-selection. The very first concatenative approaches to sound synthesis were manual splicing and stitching of magnetic tape that contained audio recording. According to Schwarz, concatenative synthesis methods could be classified based on two characteristics selection and analysis (ibid: p.3). Selection refers to the method based on which the audio units are retrieved from the audio source material. Schwarz explains that we could distinguish between two types of selection manual and automatic. The second criterion based on which we could classify musical concatenative sound synthesis is the analysis methods. We can distinguish between four types of analysis methods ranging from manual (i.e. no analysis) to low and high level descriptor analysis, see Figure 22. Figure 22, shows the taxonomy of concatenative sound synthesis methods based on these two criteria selection and analysis proposed by Schwarz, (2004).

We can further distinguish between two main applications domains of musical data-driven concatenative synthesis methods: target sound re-synthesis (Janer & Boer, 2008; Schwarz & Caramiaux, 2014; Schwarz & Schnell, 2010; Stevens, Doug, & Marschner, 2012), and free sound synthesis through the exploration of the audio corpus (Comajuncosas, 2011; Leslie *et al.*, 2010; Navab, Nort, & Wei, 2014; Schwarz & Hackbarth, 2012; Schwarz, 2012). The difference between these two categories is that in target sound re-synthesis the aim is to re-create a sound or a sounds' characteristics by providing audio examples, while free sound synthesis is less deterministic and focuses on exploration of the audio corpus in order novel sounds that do not necessarily resemble the input.

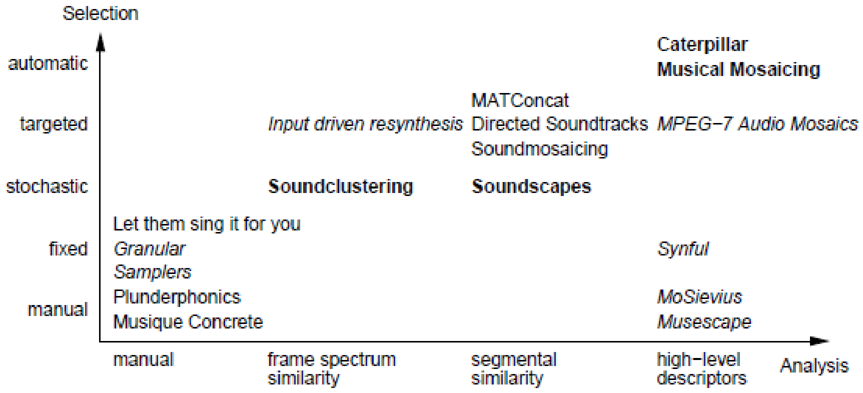


Figure 22. A taxonomy of concatenative sound synthesis methods classified based on the audio-unit selection and analysis methods (Schwarz, 2004).

In both cases a number of challenges have to be overcome in order to gain efficient control over the outcome sound. The first problem is related to the fact that it is almost impossible to define and extract a set of features that capture all of the perceptual information present in both the input query and the sound corpus (Hoffman & Cook, 2006), what I like to think of as the problem of distance between the phenomenal and the descriptor space. As Hoffman *et al.* (2006) suggest this is evident by the continually improved features for music information retrieval. Digitally extracted feature sets provide a relevant but however reductionist interpretation of what it is being perceived when we listen to music or sound at an experiential level. For example high level features such as auditory timbre, and visual texture are hard to fully describe in statistical terms. The dimensionality of real world features can be really high to compute in real-time and there is a trade-off between dimensionality reduction (which is inevitable in order to reduce computational costs) and the descriptive accuracy of the feature sets being extracted (ibid: p. 537). In the present thesis we will present the findings of an experiment which was conducted to assess how the problem of the gap between phenomenal and descriptor space affects the comprehensibility and perceived similarity of the relationships of two multidimensional audio-visual mappings.

The second problem is related to the type of input used to drive the sound synthesis. As it was mentioned earlier the input data stream used as the target for data-driven sound synthesis could derive from any source including control devices, multimedia feature extraction algorithms and sensor data. When audio feature vectors are used as the target for querying the database, the question of how to associate the feature dimensions between the target and the audio-units of the corpus it is easier to answer in comparison to when the input data derive from a feature set of none audio origin. As in the former case, the same features dimensions are available in both the target and the database side. However when visual, gestural and/or haptic data are used to query the database, it is harder to decide how the feature dimensions involved in the query should be associated. This problem is twofold. The first cluster of problems is related to understanding better the underlying principles that mediate congruency effects and sensory-motor correspondences in order to inform the cross-modal mapping and the second cluster of problems is related to the application of computational methods for modelling sensory-motor correspondences.

The following section reviews previous and current research mainly from the field of computer music and music psychology that seeks for answers to these questions.

2.4 Mapping user inputs to sound parameters consideration

Digital instruments and audio workstations have offered new ways to synthesize and control sound that were unimaginable a few decades ago. As Merrill (2007) supports, the advent of digital technologies has challenged and dismantled the relationships between instruments' physical and acoustic properties and outcome sound. As (ibid: p.3), points out, the freedom of possibilities resulting from the decoupling of these relationships comes at a cost, because these relationships were informative and an integral part of the physical interaction between performer and instrument, instrument and outcome sound. Electronic music systems allow more flexibility and freedom for the performer, because the mappings between user input and sound parameters is not constrained by either the performers’ sensorimotor skills and/or the physical and material attributes of the sound producing object. However, it is the lack of those constrains that, according to the literature, affects the musical value and expressivity of electronic music systems.

User input in this context refers to any data the performer uses to manipulate the sound parameters. The study of mapping in the context of designing interfaces for creative sound and music practice seeks to better understand the physical relationships between performer input and instruments'/systems’ sonic reactions. When designing interactive music systems, decisions have to be made regarding the relationship between input data and sound parameters. The question here is which parameters of the sound will be controlled by which aspects of the user input. To understand better the problem of mapping in digital instrument design, Hunt *et al.* (2000) propose to consider an acoustic instrument such as the violin: How do we control the volume? Which sonic parameters does the bow control? As Hunt *et al.* point out, the volume is controlled by many variables “such as bow-speed, bow pressure, choice of string and even finger position” and the bow controls more than one qualities of the sound, (ibid: 234). Figure 5.8 below shows the common multimodal interaction model that represents the relationships between performer actions and synthesis parameters. The arrows represent the input data provided by the user, while arrows that point to other arrows highlight the potential physical relationships might exist between input data.

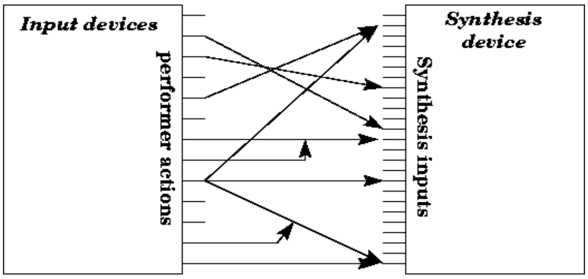


Figure 23. Mapping of performers action to sound synthesis parameters (Hunt *et al.* 2000).

In a multimodal context when the input data derive from corporeal actions of the user, it is important to consider mapping of input data to the synthesis or the musical parameters as an integral part of the interaction, as the usability or playability of the system depend on it. It is important that the relationships between user input and systems' output maintain a level of physical or symbolic correspondence: "Working with digital sound and gesture is efficient if we can keep a symbolic link between the act and the resulting sound" (Arfib, Couturier, & Kessous, 2002). Rovan & Wanderley, (1997) experimented with several mapping strategies by using same input devices and synthesis techniques but different mapping approaches and concluded that mapping is a crucial factor which affects the expressivity of the system. Through the choice of mapping, the instrument designer can create symbolic relationships between performer input and resulting sound. These relationships are important as they can improve the performer's sense of physical interaction with the interface and help him/her get more immersed in the idea that the interface which is being used is the actual object that produced the sound rather than being perceived as just a controller. Rovan & Wanderley, (1997) suggest that we could take advantage of the knowledge of physics of acoustic instruments or other objects and appropriate/formalize these to create multiparametric mapping strategies. These maintaining a perceptually meaningful relationship between user input and system output could be determinant factors concerning:

- the learnability of the interface,
- performers' perception and appreciation of the interface
- what can be achieved using the interface,
- the degree of control over the sound parameters

A series of studies have been conducted to test how mapping strategies affect a number of cognitive and experiential aspects of interaction with sound synthesis (Hunt & Kirk, 1999; Hunt, Wanderley, & Kirk, 2000). It was found that most performers enjoyed more interfaces implemented with complex mapping approaches and they believed that multiparametric interfaces were more engaging, enjoyable to use and had a longer term potential. On the contrary interfaces with simple one to one relationships between parameter and control were frustrating, confusing, odd and forced performers to mentally breakdown the different sonic properties. However, complexity on its' own is not a sufficient criterion for the success of a digital musical instrument. The quality of the associations between user input and sound parameters it could be argued to be at least as important. Hunt *et al.* (2000) proposed that the following characteristics should be carefully considered when designing multiparametric interfaces:

- *"Continuous control of many parameters in real time.*
- *More than one conscious body control (or limb) is used.*
- *Parameters are coupled together.*
- *User's energy is required as a system input."*

Research on multiparametric mapping strategies shows that multidimensional mappings are controller specific (Fels, 1994; Mulder, Fels, & Mase, 1997; Wanderley & Depalle, 2004; Wanderley, 1998). This means that implementing a successful mapping strategy for a specific controller does not necessarily means that the same mapping would be successfully used with an intermediate layer, (Hunt *et al.*, 2000). The concept here is that the controller is mapped and interacts with an intermediate layer and the intermediate layer is then mapped to the various parameters of the sound process. This is laid out in Figure 24 below. It has been argued that the issue could be addressed with the addition of an intermediate layer (Wanderley & Depalle, 2004).

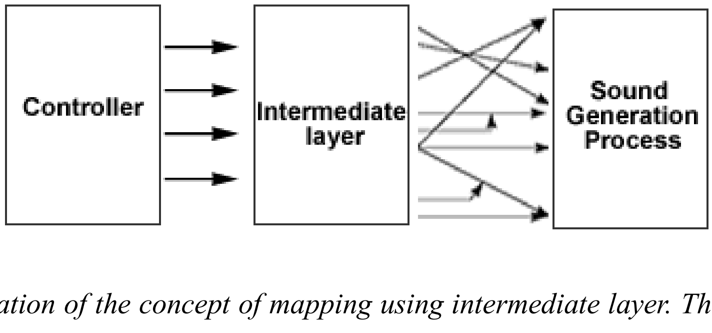


Figure 24. Illustration of the concept of mapping using intermediate layer. The arrow in the left side represent performers gestural input, the arrows in the right represent the response of the intermediate layer to the performer's gestural input (Hunt *et al.*, 2000).

Although, using intermediate layer can be useful and has applications in the implementation of multiparametric interfaces, on it is own it does not sufficient to solve the problems associated with mapping users' sensorimotor input to the sound parameters. As I mentioned in the section 2.3.1 the problem associated to this question is twofold. First, it is necessary to understand the principles that mediate cross-modal correspondence and then apply computational methods for modelling the association. Hence, it could be argued that the problem of multiparametric mapping between the user's sensorimotor input and the synthesis parameters cannot be simply solved with a technological solution. Empirical work is required in order to enact perceptual knowledge which can then be used inform and model the mapping between user's sensorimotor input and sound parameters. The present project set out to explore how and whether empirical findings from the study of cross-modal correspondence and the methods used to study multimodal aspects of perception can be used in tandem with computational modelling in order to address the problem of digital cross-modal mappings.

Mappings could be understood as a mediator that links a corporeal articulation with the production of a sound. Hence, mapping is a central element in the design of a music system and a determinant of the creative potential of the system. Although in the digital domain we have a lots of freedom in determining how the performer's sensorimotor inputs are associated to the sound synthesis or music parameters, it is exactly this freedom that often makes the performer feel confused about the affordances of the interface and affect the appreciation and level of control over the sound produced by the systems. In order to solve this problem we should design interfaces where the mediator becomes transparent, creating an illusion of non-mediation as suggested by Leman (2008). It could be argued that the only way to achieve the illusion of non-mediation is to design interfaces that conform to prior perceptual knowledge. I believe that three concepts are of great significance here affordances, causation and intentionality, similar views have been proposed by Leman (2008).

I consider the concept of affordance is more closely linked to the physical aspects of a mediating technology (i.e. the design of interface elements that enable the user input), the concept of causation to be more closely linked to the mediation of the sensorimotor inputs of the performer to sound parameters (i.e. the mapping between inputs and outputs) and intentionality to the performer goals and plans. Of course, for non-mediation illusion to be achieved in a digital interface affordances and causation should indeed both abide by the same set of physical constrains, while intentionality is less relevant in attaining the non-mediation illusion. Don Norman explains that "*an affordance is a relationship between the properties of an object and the* *capabilities of the agent that determine just how the object could possibly be used"* (Norman, 1999). By causation, we are referring to the regularities and co-variation of the performer sensorimotor input the sensorimotor feedback provided by the system. So, it could be argued that designing an interaction that is plausible in relation to prior perceptual knowledge, required considerate design of both affordances and causal relationships. Further, it could be argued that a perceptually meaningful interaction is one that aligns with prior perceptual knowledge. Hence it is necessary to enact and quantify structural and semantic correspondences, statistical and structural regularities and co-variation of events using empirical methods and carefully apply this knowledge to model and design perceptually meaningful interactions.

In recent years, there has been a growing interest in the study of multimodal interaction in the context of sound and music technology and practice. These research efforts demonstrate two things, first that it is important to gain a deeper understanding of the relationships between the senses and second that we must identify ways of incorporating prior perceptual knowledge in the design of user interfaces in order to improve the human computer dialogue and the expressivity of music systems. A number of experimental methods have been developed over the years for studying cross-modal correspondence mainly in the field of experimental/cognitive psychology, as we will discuss in Chapter 3. Some of these methodologies have been adopted in the fields of musicology and computer music to study sensorimotor responses to musical and other auditory stimuli. The method that has been most commonly used in music research is what we could name sound tracing task that was first used in mental imaging research. In free tracing experiments participants are asked to perform a sensorimotor response (e.g. a gesticulation, draw a sketch) to an audio stimulus. Godoy was one of the first researchers to conduct this type of research in musicology (Godøy, Haga, & Jensenius, 2006a, 2006b; Godøy, 2006). Godoy and his colleagues asked participants to perform gesture using a pen and a tablet in response of auditory stimuli. He used instrumental, electronic and environmental sounds as stimuli and tested three types of sounds: impulsive, continuous and iterative (Godøy, 2006). The findings from this study showed that there is consistency between the subjects' gestural responses and the audio stimuli, unfortunately the author does not provide an analysis of the nature of these coherences. This type experimental methodology known as free tracing was adopted by a number of other researchers (Caramiaux, Francoise, Bevilacqua, & Schnell., 2014; Baptiste Caramiaux, Bevilacqua, & Schnell, 2010a, 2010b; Kussner & Leech-Wilkinson, 2013; Küssner, Tidhar, Prior, & Leech-Wilkinson, 2014; Küssner, 2014).

In another study conducted by Caramiaux *et al.*, (2014), participants were asked to a free tracing gestural responses to sound stimuli. Some of the stimuli were prepared so that participants could recognize the gestures used to produce the sound while for the rest of the stimuli there was no obvious sound producing gestures associated to the sounds. The results of this study suggest that when the sound producing gesture was recognisable the participants would imitate the gesture, while when there was no obvious gesture associated to the sound stimuli, participants would trace the shape of the acoustic features of the audio (Caramiaux *et al.*, 2014). According to the authors, this suggests that identifying the cause or the source of the audio changes the strategy we follow in making cross-modal associations. For instance in this study it was shown that we can shift between gestural tracing of causal or acoustic properties of sound. For example, when the sound producing gesture of a given sound (e.g. scrapping, stroking) is recognised then we tend to mimic the gesture that we think has produced the sound and ignore its' acoustic properties. While when the cause of the sound cannot be identified, we tend to gesturally follow the acoustic properties of the sounds. Finally it was shown that between subject consistency was higher in the case of acoustic tracing (i.e. gesturally following the feature of the sound) than a causal rendering (i.e. mimicking the sound producing).

In another study, free gestural tracing was used to investigate gestural rendering of sound by 5 and 8 years old children (Kohn & Eitan, 2009). The children were presented with auditory stimuli that varied in terms of loudness, pitch and tempo. The findings showed that in most cases children associated upward and downward movements in pitch and loudness to upward and downward movements in vertical axis, while changes in tempo were associated to movement speed and muscular intensity. Another free gestural rendering study showed that musically trained subjects made more consistent associations between auditory and gestural features than non-musically trained subjects (Küssner *et al.*, 2014). This study also indicates that simultaneous changes of the sound properties affect the way in which people render the sound gesturally. Significant differences were observed between the adjustment of the speed of hand movement to varying tempo when pitch and loudness varied concurrently.

A number of studies have examined how changes in parameters of auditory stimuli are represented in the visual domain through free visual tracing of sound by people. Somers tested the effects of auditory and visual stimulation in cross-modal rendering. Visual design students were asked to create visual representations of sonic structures, while composition students were asked to create sonic representations of abstract paintings (Somers, 1998, 2000). The aim was to make the students perform structural analyses of the object in question and then render the object's individual components and their structure in sonic or visual terms. Somers found that the mental transformations from sound to vision and vice versa resulted in creative works that exhibited structural coherence between auditory and visual properties, unfortunately the author does not provide an analysis of the nature of these coherences. In another study Kussner & Leech-Wilkinson, (2013) using a free tracing paradigm investigated how musically trained and untrained participant represented the sound of simple stimuli that varied in terms of pitch, loudness and tempo, as well as short musical excerpts. Using a graphics tablet, participants represented the sound stimuli visual by drawing. The results revealed that the majority of the participants represented pitch with to height (i.e. vertical position on the tablet to represent high pitches), and loudness with the thickness of the line, and duration was usually represented by the length of the line in the horizontal axis. Furthermore, the study concluded that musically trained participants' representations were more accurate in comparison to the untrained participants.

Free tracing was used to investigate how musicians from different cultural background (i.e. British, Japanese, New Guinean) made associations between musical sound and shape (Athanasopoulos & Moran, 2013). Participants were asked to represent visually by drawing on paper short musical excerpts varying in pitch contour. The only instruction participants were given was that their visual representation should enable another community member who will see the drawing to understand the properties of the sounds being represented. The results revealed differences in the way musicians from different cultures represented pitch and time. British and Japanese participants represented time on the horizontal axis (i.e. from left to right) while pitch was represented on the vertical axis (i.e. height of pitch corresponding to of vertical position on the paper). Only a minority of Japanese participants represented time vertically. However, the majority of New Guineans did not follow the time-pitch parameters that were intentionally varied in the musical excerpts, instead they focused on the representation of loudness which was represented as color hue. This findings suggest that cross-modal correspondences may be linked though culture/nurture as oppose to be hardwired structural correspondences which form at early stages of sensory information processing. It should be noted that whether cross-modal correspondence is due to culture or due to the way the brain process sensory information is currently a topic of debate. As it will be discussed in the next chapter the evidence we have is not conclusive regarding this matter, and different researchers will have different views about whether it is nature or nurture that plays the most important role in mediating cross-modal correspondences.

2.5 Conclusions

This chapter reviews previous and related work and identifies a number of issues of concern. Firstly, previous research has emphasised that defining approaches to visual representation of sound is extremely important for analytical, compositional and pedagogical purposes. Secondly, there is need for conceptualisation tools that are capable of bridging the gap in the creative process between the phases of conceptualisation and production. Thirdly, graphical representations used for interaction with sound synthesis for sound design and musical applications should focus more on perceptually relevant sound attributes and less on low level attributes of the given synthesis methods. As digital technologies allow us to make arbitrary associations between any type of data (i.e. modal or amodal). The question of how we can approach the design of such associations in objective terms poses a challenge to a number of disciplines including computer music, information display, and sensory substitution amongst others. Finally, there is need for empirically supported design frameworks for cross-modal representation and interaction that respect users' intuitions and expectations by maintaining cross-modal correspondences which are based on that prior perceptual knowledge, when possible.

3 Cross-Modal Correspondence

3.1 Introduction

Humans and other animals inhabit a multisensory environment and are equipped with multiple senses in order to perceive as accurately as possible the information available in the environment. Our senses receive a constant flow of a variable number of unimodal stimuli at any given time. These signals derive from a number of environmental sources/properties. In order to sense the information provided by the unimodal signals the brain has to “decide” which stimuli/properties to combine and which to keep separate. How the brain achieves to combine sensory stimuli is the main topic discussed in this chapter. More specifically this chapter reviews literature mainly from the field of experimental psychology to present the state of the art in audio-visual binding and correspondence research and discuss how perception and other high level cognitive processes mediate cross-modal interactions and integrations between unimodal stimuli.

The aim of this chapter is threefold. The first aim is to understand better the mechanisms that underpin audio-visual binding and congruency effects. The second aim is to summarise experimental findings that are relevant to the understanding of audio-visual correspondences and identify which audio-visual features dimensions are the best correlates. The third and final aim is to summarize briefly the main methodologies used to study cross-modal correspondence.

3.2 Differences between Auditory and Visual Perception

The discussion about what is a sensory modality and what criteria can be applied to distinguish between the different senses can be traced back to antiquity. For example Aristotle, suggested that each of the sense organ has *proper objects*. Aristotle coined the term *proper objects* to refer to the perceptible features of the environment that can only be sensed by one sense and not by the other senses (e.g. color can only be sensed through sight (Sorabji, 2011)). He also coined the term *common sense* to refer to the perceptible features of the environment that can be sense through multiple sensory modalities. For example, the shape or a texture of an object can be sensed using either vision or touch. Since the Aristotelian times philosophers and scientists have debated what criteria can applied for distinguishing the senses. Table 1, shows what criteria philosophers have traditionally used to individuate the different senses (MacPherson 2011). The table suggests that traditionally, the senses are individuated and compared according to four dimensions (i.e. representation, phenomenal character, proximal stimulus and sense organs). Representation refers to the perceived properties of the stimulus, phenomenal character to the qualitative experience of the perception, proximal stimulus refers to the physical properties of the stimulus and the sense organs to the physical make-up of the organs used for sensing the physical energy and chemical composition of the environment.

Table 1. Criteria different theories use to individuate the senses (Macpherson, 2011).

	Vision	**Touch**	**Hearing**	**Taste**	**Smell**
Representation	Color, shape and movement, at a distance from our body and in front of our eyes.	Temperature, pressure, shape and movement at the surface of our body.	Sounds, volume, pitch, objects being struck or vibrated at locations in and at a distance front and all around our body.	Flavours (sweet, salty, bitter, sour, umami) in the mouth or on the tongue or in the food touching the tongue.	Odours located either in the nose or in the air around the nose, perhaps coming from a certain direction.
Phenomenal Character	Visual experience	Tactile experience	Auditory experience	Taste experience	Olfactory experience
Proximal stimulus	Electromagnetic waves	Mechanical pressure and temperature.	Pressure waves in a medium such as air or water.	Chemicals that affect receptors on the tongue.	Volatile molecules that affect the epithelium.
Sense organs	Eyes, particularly the retina	Skin or receptors in the skin.	Ears particularly the cochlea.	Tongue, particularly the taste-buds on the tongue.	Nose, particularly the nasal epithelium.

At first glance, sensory modalities might appear to be of such disparate nature. For example we know that each sensory organ has a unique physiology, it can detect a unique type of signals, most attributes of the signals detected by each sensory organ are processed in different areas of the brain and each sense produces or involves a distinct phenomenological experience. For example, visual signals consist of electromagnetic waves, which travel extremely fast and have extremely small wavelength. In contrast, sound energy is mechanical, travels relatively slowly, and the size of the wavelength can be rather large (Handel, 2006). In the case of hearing, mechanical waves are broken down into frequency components by the hair cells in the inner ear which bend depending on the variation of the intensity of specific frequencies. While in the case of seeing, cells fire to the intensity variations in small regions of the retina and moreover fire maximally to intensity variations that occur alongside specific directions (ibid: p.7).

Further, audition is viewed predominately as a temporal sense closely associated with the recognition of events, while vision is viewed predominately as a spatial sense closely associated with recognition of objects (Bregman, 1994; Handel, 2006). Auditory perception is mainly concerned with sources that produce sound, as opposed to the material and structural properties of the surfaces that reflect sound, which is the case with perceiving light. Surfaces that reflect sound might provide spatial cues and alter perceptible qualities of an auditory object, but they are not what we perceive as the object. A bat for example uses sound in the way we use light to perceive what we consider as visual information. Audition in humans has evolved to perceive spectral flux rather than spectral reflectance which is the case in bat vision (Mollon, 1995). Moreover visual objects tend to occlude other objects behind them while auditory objects sum common frequencies components (Handel, 2006). According to Keeley, (2011), the criteria that best account for modality differentiation:

- Physics: Sensory systems detect and respond to different forms of energy independently of any psychological and biological concerns. Physics provide a useful ontological differentiation between the senses as electromagnetic phenomena differ from mechanical energy and both notably differ from chemical composites.
- Neurobiology: Sense organs physiology and connections to the brain differ.
- Behaviour: The ability to discriminate behaviourally between stimuli of different physical type.
- Dedication: The developmental importance of a sense to an organism.

While the first two elements (i.e. physics and neurobiology) in the list above are relatively easy to interpret, the last two require further explanations. What Keely means by behaviour is the ability of the organism to act based on the stimuli. For example tanning is a reaction of the skin due to the stimulation of the skin by the light radiation, but is not something the organism consciously does as a response to light exposure, therefore tanning cannot be classified as a behaviour of the organism as it is an automatic reaction of the skin that happens beneath the conscious awareness of the organism. It could however be considered as a behaviour of the organism’s body, if a body were to be considered as conscious. In the contrary dancing to music or verbally responding to a question are types of behaviour, as these types of processes demand the organisms’ ability to discriminate behaviourally the sensory stimuli and cause a sensorimotor response. However as Keely discusses the behaviour criterion on its own is insufficient to individuate the senses. He argues that taking behaviour alone as a criterion, it could be argued that humans have an electrical modality as they can discriminate between electrical currents of different strength. For example when a person touches a nine volt battery in their tongue they are able to tell if the battery is charged or it is empty. The dedication criterion can help resolve the human electrical modality issue. Dedication refers to the evolutionary importance of the sense to an organism. So it does not necessarily means that because human have the ability to discriminate electrical currents we ought to attribute an electrical modality to humans, because humans do not use the electrical capacity to understand, navigate and interact with their environment.

Keeley essentially rejects the first two criteria from table 1 (i.e. representation and phenomenal character that derives from the Grice's original proposal (Grice, 1962)) as good criteria for differentiation between the senses. Keeley's criteria follow a more physicalist approach to the problem of sense discrimination, attempting to remove psychological dimensions as a sensible criteria for modality differentiation. These is useful as it removes the higher level sensory products such as representation and phenomenal character which could be seen as by-products of the neurobiological dimension of the senses and successors of the more primordial physical attributes. Essentially Keeley distances himself from how things appear and strives to achieve a more ontological solution to the problem of differentiation of the senses.

Attempting to differentiate between the senses is useful as it raises a number of questions that help to understand better the nature of perception as well as the sensory apparatus. However it is questionable whether answering the question of what criteria are the best for individuating the senses is helpful in understanding the way humans perceive the external environment. It could be argued that although there are different types of senses, at the same time these are all tokens of perception and therefore it is impossible to completely separate them. Consequently, the discussion of individuating the senses is constructive in understanding the individual components of perception but not for understanding the higher levels of perceptual and cognitive functions. In recent years researchers and philosophers are becoming increasingly interested in the similarities between the two modalities rather that the differences. It could be argued that focusing on the aspects of perception that are shared between the senses rather than their differences is necessary to fully understand perception, because the latter obscures rather than it reveals the purpose and function of perception. The following sections draws on literature from psychology, neuroscience, linguistics, and philosophy to portray a cross-modal conception of auditory and visual modalities focusing on the similarities in form, structure, and function and exploring the links to perception, conception, and language.

3.3 Similarities between Auditory and Visual Perception

Humans and other animals inhabit a multisensory environment and are equipped with multiple senses in order to sense the information available in the environment. The senses have evolved to receive and recognise information, with the ultimate goal to aid the organisms that possess the apparatus to survive (Kubovy & Schutz, 2010). The stimuli we receive from the environment are only rarely uni-modal. Senses should be flexible and adaptable so that they can respond to a rapidly evolving dynamic environment to sense the information that surrounds us as accurate as possible. Different senses provide information of differed level of precision. For example, on the one hand hearing is less precise than vision in spatial localisation, on the other hand hearing allows as to perceive 360 degrees of the auditory scene while vision allow us to attend only a limited area of the visual scene. Hence, multiple forms of sensing enable us to construct a more precise and robust representation of the external world. The information we perceive through the different senses can be redundant or complementary (Parise & Spence, 2013).

Redundant cues, refer to sensory information perceived through different sensory modalities which describe the same feature of the physical world. While complementary cues, refer to the features of the physical world which can be experienced only by one sensory modality (ibid: 790). In order to explain the difference between redundant and complimentary cues, Parise *et al.* (2013) bring an example of a dog barking behind a picket fence. The information received by both visual and auditory channels provide cues regarding the spatial position of the dog, assuming that we are close enough to both hear and see the dog. The spatial cues received by the two modalities could be considered as redundant, although the accuracy of the information might vary. While the colour of the of the dog's hair can only be perceived via visual sensory channels. Likewise the timbre of the dogs barking sounds can only be perceived if we can hear the sound the dog produces. Therefore the colour and the timbre could be considered as complimentary cues that are unique attributes to each modality (ibid: 791). Redundant cues are combined to create more precise and robust representations of the environment and in order provide better sensory estimates. While complimentary cues provide an additional layer of information that further enhances the sensory estimates and can account for the richness of our sensory experience.

When the sensory cues received by multiple senses are redundant, it is sufficient if the observer attends to the information from one the sensory channel to know what information to expect from the other sensory channels. Therefore when the observer assesses the relationship between two or more redundant feature dimensions from different sensory modalities, the corresponding feature dimensions should provide an equally high correlation probability in relation to the observer's expectations. However when multiple senses provide complimentary information then the ability of the observer to infer the corresponding dimension by attending only to the information provided by one modality is decreased, as the sensory estimates became highly uncertain and the observer's assessment unpredictable. For example, listening to the sound of a barking dog provides no information about their color, and similarly seeing the color of a dog provides no information about that what the dog might sound like. However empirical evidence shows that humans report perceived congruency between some complimentary features varies and that some cross-modal feature associations are preferred over others (Evans & Treisman, 2010; Parise & Spence, 2013; Spence, 2011). Moreover there is evidence that different individuals exhibit common patterns of perceived congruency between specific complimentary cross-modal feature dimensions and that common congruency patterns exist across different ethnic groups. Parise *et al.* (2013) coined the term *Complimentary polar features* to describe complimentary feature (i.e. visual size and loudness) dimensions that produce high levels of perceived congruency. Furthermore, they suggest that we can distinguish between three types of cross-modal correspondence between complementary polar features: structural, statistical and semantic correspondences. Below we explain what each type of correspondence refers to.

3.3.1 Structural Correspondence

The first type of cross-modal correspondence is structural correspondences. They occur due to the similarities of the transformation of sensory information into perceptual information. Handel (2006) suggests that there are striking similarities in the way auditory and visual perception encodes information received by the sensory organs (ibid: 2006, p.95). Handel made an extensive survey of research on the receptive fields, and the electrical firing spikes caused when encoding sensory into perceptual information. The results suggest that audition and vision are perceptually identical in this respect. It has also been suggested that the intensity of spatial, temporal and quantitative features of all sensory inputs are encoded into perceptual information using a single coding systems found in the inferior parietal cortex (Walsh, 2003). Extrapolating Walsh's proposal we might be able to argue that structural correspondence occur because of the use of common neurophysiological resources that process different quantitative features (see Figure 25).

Handel supports that there are a number of good reasons to suspect that there is a concrete set of principles that unify perceptual processing and experience between all sensory modalities (Handel, 2006). Similarities can be found in the tuning of sensory receptors to sensory energy, in the hierarchical organization of cognitive function, and in the interactions and integration of

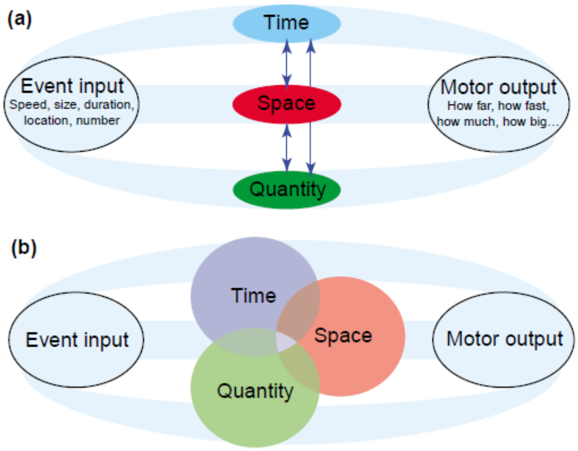


Figure 25. Walsh's two schemas for processing time, space and quantity. Panel (a) schema were magnitudes are analysed separately based on their own metric. Panel (b) shared magnitude system using common metrics for the different sensory inputs (Walsh, 2003).

sensory specific information (ibid: p.150).

Handel points out that in both auditory and visual perception of motion, motion is a direct result of the ability to detect changes in the configuration of texture over time, hence motion is tightly linked to texture segregation in both modalities. Another analogy that could be made between visual colour and auditory timbre is constancy. Color constancy refers to the capacity of the human visual system to ensure that the color of objects remains relatively constant under various illumination conditions. Likewise, the human auditory perception has the ability to recognise sound quality of an audible sources when pitch and loudness is varied. Further, Handel explains to achieve color constancy, the visual system has to remove the effects of the change in illumination in order to identify the underlying reflectance of surfaces. To achieve sound source constancy the auditory system of the listener has to remove the effect of changes in pitch and loudness in order to identify the underlying harmonic relationship of the spectrum of the sounds produced by a source, (ibid: p 372).

Further, both colour and timbre are speculative in nature in the sense that they are estimates of likeness that could be categorised under one of the major classes of the colour scale (e.g. red, blue, green), which also is the case with timbre. Further, Handel argues that the organisation principles, as well as functional similarities in object identification and the receptive fields of auditory and visual objects, suggest that both modalities are underpinned by principles which are identical, and could be though as generalised Gestalts. For example, according to Handel, the eyes and the ears each receive two signals that are slightly displaced and different (i.e. the signals received by the left and right ear and eye). The problem which cognition has to solve is establishing correspondences between displaced stimuli both at unimodal and multimodal level. In vision, the difference is mainly spatial while in audition it is mainly temporal. The task in both cases is to match the signals received by each eye and ear in order to construct a coherent mental representation. Perceiving is not merely about attending to the parts of the sensory stimuli. Perceiving is an active process, which is not modular, but involves interactions and integrations amongst sensory stimuli. A coherent mental representation of the environment is not constructed and experienced independently for each modality. Instead, we perceive the external environment as a unified phenomenon.

3.3.2 Statistical Correspondence

The second type of cross-modal feature correspondence originates from statistical or functional regularities that can be commonly observed in the physical environment. For example a common statistical regularity is the size of physical body and its resonating frequencies, the changing distance to a sound source and its relative perceived amplitude level. The brain uses these learned statistical regularities to assess which cross-modal cues to combine and which to segregate. These statistical regularities are likely to be independent of culture. It might be argued that the degree of perceived correspondence between two cross-modal cues by an observer depends on the probability of these associations to occur in the environment. As Parise *et al.* (2013) suggest, observers can be trained through exposure to artificially compose cross-modal correspondences to learn new correspondences, however these newly learned feature associations will not be universal (i.e. shared by all humans like natural correspondences).

The statistical regularities based on prior perceptual knowledge hypothesis might have the strength to explain some confusing observations related to the correspondence between auditory and visual stimuli/properties, (Eitan & Granot, 2011; Eitan, Schupak, Gotler, & Marks, 2011; Eitan, Schupak, & Marks, 2008; Eitan, 2013; Marks & Eitan, 2012). For example research findings suggest that one visual feature can correlate equally well to two auditory features (e.g. lightness to pitch and loudness) and vice versa auditory features can correlate equally well to multiple visual features (e.g. pitch to lightness and vertical position), (Evans & Treisman, 2010; Marks, 1974, 1989). The couplings between different audio-visual feature sets based on prior perceptual knowledge could explain the first phenomenon that of one-to-many, many-to-one correspondences between auditory and visual features. For instance, as an auditory source is getting closer to the position of the observer, the relative amplitude of the sound increases, while the opposite happens when the source is moving away from the observer's position in space. This could lead to a visual distance to auditory loudness correspondence. On the other hand as the visual object approaches the observer its relative size also appears to be getting bigger forming a second correspondence between size and loudness

Further, it has been argued that correspondences can be affected depending on whether the stimuli used for testing is static or dynamic, (Eitan, 2013; Granot & Eitan, 2011). Static and dynamic stimuli here refer to whether the variations of the stimuli are discreet (static) or continuous (dynamic). For example, using static stimuli testing visual brightness to pitch correspondence, means using visual stimuli that differ in terms of brightness in a discreet manner (i.e. one stimulus is dark while the other is bright) and two auditory stimuli that differ in terms of pitch in a discreet manner (i.e. one stimulus has low pitch the other has high pitch). While when using dynamic stimuli, the pitch and brightness is varied over time. For example, a single visual object displayed on a computer screen is resized dynamically from large to small and the pitch of the sound decrease in size. The differences between static and dynamic stimuli could also be explained, via conflicting learned cross-modal regularities Eitan *et al.* (2011) showed that while large visual objects are usually associated with low pitched sounds and small objects with high pitched sound, when the stimuli's pitch is dynamically manipulated this effect can be reversed. For example they found that a rising pitch contour is associated with an increase in size of a visual object and falling pitch contour with decreasing size object. In this case it could be argued that when perceiving the correspondence between size and pitch using static stimuli the learned regularity which is used to assess the plausibility of the cross-modal association is that of larger sounding bodies produce resonate at lower frequencies therefore produce lower pitched sounds. In contrast when using dynamic stimuli to test the correspondence between pitch and size it could be that a different learned cross-modal regularity is used to assess the plausibility of the association between A/V features. For example, according to the Doppler effect when an auditory source is moving away from the position of an observer the relative frequency of the sound is decreasing and when the object is moving closer to the observer it's frequency appear to increase. This example illustrates how cross-modal associations can be conflicting.

Furthermore, research finding suggest that there are interactions between the perceived correspondences of complementary polar features when other sound and/or visual parameters are varied simultaneously. For instance, an increase in tempo rate is often correlated to an increase in visual motion speed, however when increase in tempo is accompanied by a decreasing loudness the correspondence between tempo and speed is weakened, (Eitan & Granot, 2011; Eitan *et al.*, 2008; Eitan, 2013; Granot & Eitan, 2011; Marks & Eitan, 2012). In this case it could be argued that loudness is also somewhat associated to speed motion. For instance imagine that you hold the one end of a rope in your hands and you rotate in the air, the faster you rotate the rope, the louder the sound produced by the interaction between rope and the air. So an increase in speed will in many cases also cause an increase in loudness. When the opposite relationship is observed for example increase in tempo and speed with a decrease of loudness clashes with observes' prior perceptual knowledge, hence the correspondence between tempo rate and visual speed is weakened.

3.3.3 Semantic Correspondence

The third type of correspondences are semantically mediated correspondences which occur due to the usage of common linguistic labels to describe perceived cues from different sensory inputs. For example, human perceive colors as being cold and warm, pitch as being high and low, and sound timbre as being dark and bright. While in the case of structural and statistical correspondence the perceived similarity between two objects occurs because the two objects share structural or functional features. In semantically mediated correspondence similarity occurs when two objects are semantically linked. Let's take for example two word pairs (such as *dog* and *god*, *fridge* and *bridge*) this two words do not have much in common in the semantic domain but do share physical features such as letters. While for example the words (*aim* and *target, car* and *auto-mobile)* do not share any structural properties but they do map to the same concept or object so they share a common semantic domain. Gentner & Markman (1997) have proposed an even more refined taxonomy of the different forms that similarity can take, and the role similarity has in forming relationships between concepts and objects, see Figure 26. This taxonomy distinguishes between literal similarity, mere-appearance, Metaphor and Analogy.

According to Gentner's and Markman, analogy is a mediation device to express similarity of relational structure between two objects or domains despite that these objects differ greatly structurally for example the words *car* and *automobile*. Analogy is often viewed as a superior cognitive process in comparison to similarity which is considered somewhat primitive in the sense that it is a process that we share with all animal species, (Gentner & Markman, 1997; Larkey & Markman, 2005; Medin *et al.*, 1993; L. Smith & Heise, 1992). Gentner and Markman explain, that in analogy the relational basis that associates conceptual structures can be viewed as being somewhat detached from sensory-motor experience. While mere-appearance and literal similarity is more closely linked to sensory-motor experience. Many researchers have argued that similarity

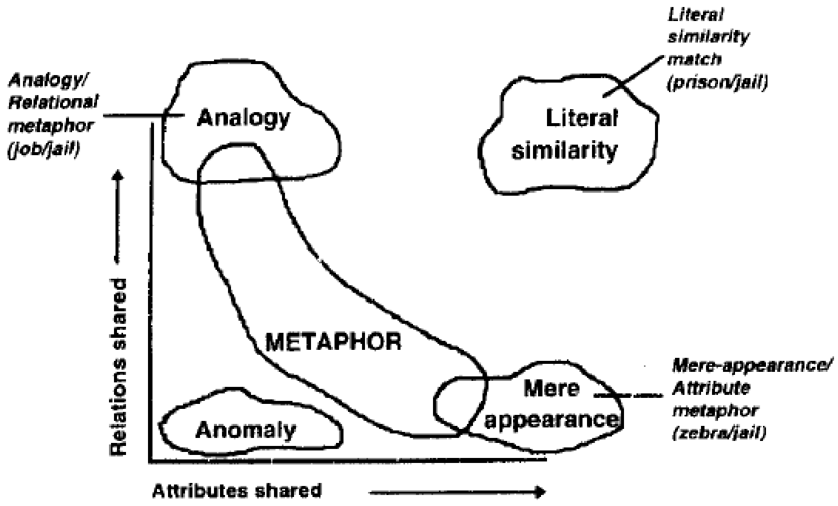


Figure 26. Similarity space showing the different kind of matches in terms of the degree of relational (i.e. semantic) versus attribute (i.e. structural) similarity (Gentner & Markman, 1997).

is more reliant on perception while analogy relies more on language and semantics, (Fauconnier, 1997; Lakoff & Johnson, 1987; Lakoff & Johnson, 1980; Medin *et al.*, 1993; Smith & Heise, 1992). Conversely, it has also been argued by many researchers that a perceptual grounding of analogy and metaphor can be identified, (Barsalou, Simmons, Barbey, & Wilson, 2003; Barsalou, 1993; Hampe, 2005; Parise & Spence, 2013; Talmy, 2000). For example research findings show that across cultures the same sensory related linguistic metaphors are used (e.g. height-pitch) to describe the same sensory phenomena that are experienced through different sensory inputs. This suggest that semantically mediated correspondence might be underpinned by statistical regularities, (Parise & Spence, 2013; Rusconi *et al.*, 2006).

Talmy explains, that there are many disagreements in psychology with regards to where perception ends and conception begins, (Talmy, 2000). Drawing a line between phenomena that are purely perceptual and phenomena that are purely conceptual has been proven a difficult task, (ibid: p.139). For example when one sees a visual object such as a bicycle, does the recognition of the object reside in perception or in cognition? As Kant argued, conception without percept would be empty while perception without concepts would be blind, (Masih, 1993). The interrelationships between perception and conception are numerous so it is impossible to have a pure perception or pure conception. Perceptual information is shaped, formed and divided by the concepts that have formed in the past, and concepts have to be filled in with perceptual information to have any coherence of contain or substance. Concepts and properties that we have learned through the sensory-motor systems and often are used as metaphors to express abstract thought, emotions, and intentions, (Barsalou *et al.*, 2003; Lakoff & Johnson, 1980; Zaltman, 2002). In order to explain how the processes that enable conceptual blending operate (Lakoff & Johnson, 1987) introduced the theory of image schemas, which has been a cornerstone in cognitive linguistic thought. The theory suggests that image schemas are pre-conceptual and central to the formation of conceptual knowledge and linguistic abilities. According to Hampe (2005), image schemas are highly schematic gestalts which capture the structural coherence of sensory-motor experience integrating information from multiple modalities, beneath conscious awareness. As gestalts, image schemas are both active and flexible, enabling to map the phenomenal structures of modal experience to abstract and amodal structures, similar to the way the anger is semantically mapped to perceptual grounded experiential phenomena such as boiling or exploding or the perceptual grounding of our mood by saying I feel high or low.

Lakoff & Johnson, (1980) proposed to think of linguistic metaphors as consisting of two domains: the concrete domain and the abstract domain. The concrete domain refers to the literal meaning of the metaphor. The abstract domain refers to the concept that a metaphor signifies, the semantic content. For example, in the metaphor *I grasping an idea,* if it is interpreted literally the meaning that it conveys is misleading and confusing as the metaphor is a symbol and consequently is defined by conventions. Therefore, it could be said that although in metaphor the concrete domain is important for the emergence and the manipulation of abstract concepts, however the concrete domain in itself is inadequate for definition and the representation of an abstract concept. Lakoff *et al.* (1980), gave the following example in order to explain this, they said that if all we know about anger was based on an analogy such as that *anger* is like *liquid exploding from a container* and all we knew about anger is based on the knowledge that derives from this analogy, these would be far from sufficient to understand what anger is. However anger is a rather abstract concept that does not have an inherently specific representational content, so it could not be mapped systematically to the concrete domain. This leads us to the conclusion that the mapping between abstract and concrete domain has to be learned.

Cross-modal correspondences according to many researchers can originate at different three hierarchical level of sensory processing including psychophysical, perceptual and post-perceptual/cognitive (Landy, *et al.*, 2012; Marks, 1974, 1989; Martino & Marks, 1999; Shams & Beierholm, 2012). However, the relative weight and of each hierarchical level in the perceived congruency is hard to determine. The next section reviews literature that explore the state of the art in sensory cue integration. Sensory cue integration refers to the process were two sensory cues received by two sensory modalities are integrated into a higher level multimodal object. For example, in ventriloquist effect, were two otherwise unrelated cues ventriloquist voice and puppets mouth movement are combined to give the impression of the talking puppet.

3.4 Cross-modal sensory cue integration

It has been suggested that there are two subsystems in perception named the dorsal stream which is concerned with identification of objects and the ventral stream which is concerned with the spatial location of object in relation to an individual, (Kubovy & Schutz, 2010). The dorsal stream is also known as the *what* subsystems and the ventral stream also known as the *where* subsystems. It has been argued that audition and vision share the two stream hypothesis. Initially the theory was concerned only with vision, but not long after the hypothesis was applied to auditory perception (Rauschecker & Tian, 2000), and later was further adapted to account for multimodal aspects of perception such as audio-vision, (Kubovy & Schutz, 2010; Kubovy & Van Valkenburg, 2001). As Kubovy *et al.* explain, the senses have evolved to receive and recognise information, with the ultimate goal to aid the organisms that possess the apparatus to survive. The senses should be flexible and adaptable so that they can respond to a rapidly evolving dynamic environment. The stimulus we receive from the environment is only rarely uni-modal. Many sensory phenomena in the nature can be experienced through multiple senses. Therefore it has been argued that studying sensory modalities in an isolated manner could only be justifiable if sensory stimulation of one sensory modality was interpreted independently to the signals received by other sensory modalities, (Shimojo & Shams, 2001).The interaction and integration between sensory stimuli raises questions regarding the mechanisms and the rules that underpin cross-modal perception.

Kubovy *et al.* (2010) provided a plausible explanation to the question of how sensory information might interact and integrate. Their theory is based on two concepts: the first is the idea of indispensable attributes and the second sensory integration based on the *what/where* subsystems. They support that auditory and visual *what* and *where* subsystems have complex relationship and interact at multimodal level. In order to explain how this might happen, Kubovy devised two thought experiments, one for vision and one for audition. His thought experiments had as objective to explain how the perception of numerosity might be affected by coinciding synchronous stimuli, and investigate which attributes are indispensable for the discrimination of numerosity between objects (ibid: p.56). In the first thought experiment, they considered two visual features light wavelengths and spatial location. Through their thought experiment, they demonstrated that when two colored light sources collapse in space and time, our ability to discriminate between the two is compromised. However, in audition, our ability to discriminate between two sound sources collapses when the frequency and the time are identical. They conclude that the indispensable attributes in the case of sound are frequency and time while in the case of visual information indispensable attributes are space and time. A particularly interesting point they make which goes beyond the individual attributes and their ability to aid in discrimination of numerosity (i.e. indispensable attribute), is the idea of collapse of numerosity. One particularly fascinating aspect of Kubovy's and Valkenburg's theory is the suggestion that, when spatial and temporal alignment and a plausible causal relationship exist between an auditory and a visual object, then the two phenomena collapse into a single percept. They argue that this happens because the auditory and visual *what* and *where* subsystems coincide. This explanation has the strength to demonstrate how shared attributes between the two sources have the ability to affect our perception of numerosity, which has the potential to explain some facets of multisensory binding.

For example in audio-visual speech perception, the visual stimulus is not perceived separately from the auditory stimuli, instead the two are perceived as one phenomenon. Moreover they explain that for a successful binding to occur, the causal relationship between the two objects and their attributes must be plausible in terms of: (i) prior experience of similar events and phenomena, (ii) in terms of time (i.e. synchrony), and (iii) in terms of space (i.e. collocation). Plausible common cause is a concept that deserves more consideration. It could be argued that it is of great importance in the context of designing multimodal systems, and information displays. Plausible common cause implies that the phenomenon of binding can occur as long as the association between two modal phenomena appears realistic according to prior knowledge. Conversely, by enacting this knowledge and applying it to digital multimodal mappings, it should be possible to create intuitive associations, associations that give the impression of collapse of numerosity between the modal elements involved. Therefore for a multimodal association to be considered as intuitive, it does not necessarily have to accomplish an absolute structural isomorphism, but rather be persuasive (i.e. create the illusion of realism by conforming to prior perceptual knowledge). Hence it will be necessary to explore these concepts further to create sensory representation that derive their significance from sensory experience and require minimal learning.

Philosopher and scientists alike have long recognised the importance of causality as an organising principle that allows humans to make sense of the physical world. Despite the fact that there have been debates about whether our ability to perceive causal relationships is deductive or inductive, scientist generally agree that humans perceive the world in terms of causal relationships. As the environment humans inhabit is rich in sensory cues, combining these cues into percepts is essential in order to make sense of the phenomena that surround us, (Wei & Körding, 2012). It has been argued that sensory cues only make sense when the brain understands their causes. When exposed to sensory stimulation the nervous system has to make a decision whether sensory cues should be combined or segregated. To make this judgment the brain has to estimate whether the sensory cues derive from an independent or a common cause. The potential role of causation in the process of multi-sensory feature integration has been discussed by a number of studies. Research findings suggest that our ability to make causal inferences (i.e. decisions about whether two sensory cues/ phenomena have a causal relationships and therefore should be integrated) are fast, automatic and distinct from the causal inferences on the cognitive level, which we will talk about later in the section 3.6. Research findings suggest that the perception of causality exist in early infancy. The perception of causality is affected by many factors such as the qualities of the sensory cues, perceptual grouping, attention, context amongst other factors. Studies have shown that the degree of integration of sensory cues is affected by the perceivers' beliefs and intentions. For example, when subject are told that multisensory stimuli with disparities are from different sources sensory integration is significantly reduced. However when subjects are told that sensory stimuli with disparities have common cause, subjects tend to integrate stimuli more, despite the lack of common cause.

Researcher have applied Bayesian theory in order to study and model sensory integration based on causal inference. Most studies on modelling sensory integration agree that sensory cues are combined linearly and weighted depending on the reliability of the individual cues, (Wei & Körding, 2012). Linear cue combination models assume that individual cues are all estimating the same feature of the world (e.g. the depth or location of the same object), (Landy *et al.*, 2012). Gaussian distributions defined by a mean and standard deviations can be used to represent the likelihood that a visual and an auditory cues are drawn from a common cause. When there is overlap on the auditory and the visual distributions then the likelihood that the sensory cues derive from the same source increases. The figure below shows an example based on ventriloquist effect, were two otherwise unrelated cues ventriloquist voice and puppets mouth movement are combined to give the impression of the talking puppet. The example in the Figure 27, show how the probability of cue combination p(c) is affected as a result of increasing spatial disparity between the auditory and visual sources. The model below takes three variables into account, the probable spatial position of a visual cue (red distribution), the probable spatial location of the auditory cue (green distribution) and a subject's estimate of the source of spatial location of the auditory source after computing/processing the distance between visual cue and auditory cue (blue distribution). When the auditory and the visual cues are coming from the same location then the two cues are combined, then the more spatial disparity there is between the two cues the likelihood that the same source caused the stimuli decreases.

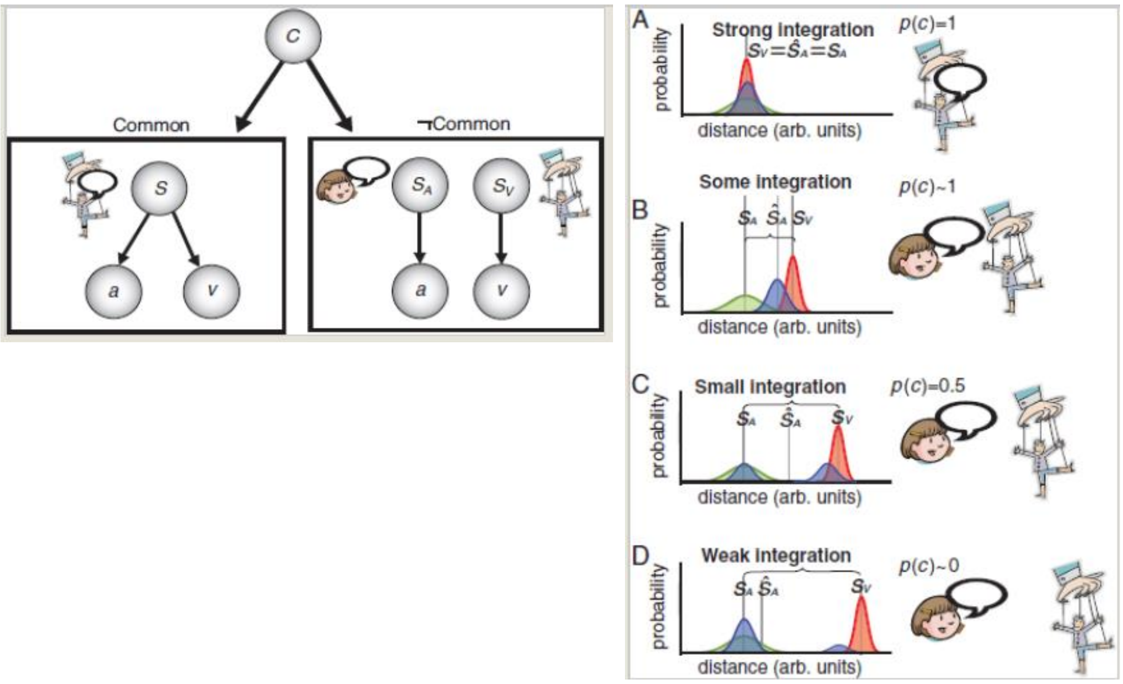


Figure 27. Example of the estimation of the probability of cue combination between auditory and visual stimuli (Wei & Körding, 2012).

The model above suggests that there are three types of interactions between sensory cues (i.e. full integration, partial integration and weak integration). In fact the model cannot account for the case when two sources do not derive from a single source. However different models has been proposed that does not assume integration and that can be extended to any number of signal and modalities, (Shams & Beierholm, 2012). Figure 28, shows three type (A) is the traditional model that assumes integration, (B) represents two cues that may or may not be caused by the same source, and (C) represents three sensory cues that one two or three may be caused by the same or distinct sources, the double arrows in (B) and (C) represent interdependency between the cues. The Bayesian observer model in (B) considers statistical independence between auditory and visual signals as signals of each modality can be more or less accurate as different type of noise can corrupt them independently. In this model XA and XV represent the actual cues (potentially noisy), while SA and SV represents the optimal estimate relationship between the cues, which results from prior perceptual knowledge. So the likelihood of XA occurring given an auditory cause SA is represented by the probability distribution P(XA|SA). Similarly the likelihood of the visual cue XV occurring given SV is represented by the probability distribution (XV|SV). So the priors P(SA/SV) indicate the perceptual knowledge of the observer about the auditory-visual events in the environment. *"This prior probability encodes the interaction between the two modalities, and thus it can be referred to as an interaction prior",* (Shams & Beierholm, 2012).

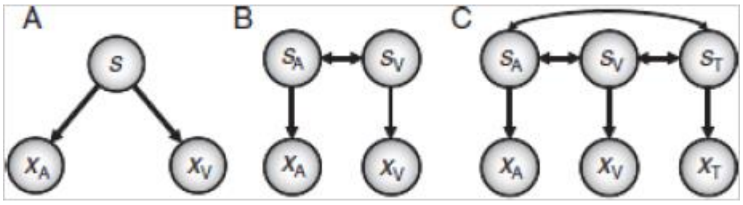


Figure 28. Generative models of different cue-combination models. The graph nodes represent random variables, and arrows denote potential conditionality. The absence of an arrow represents statistical independence between the two variables (Shams & Beierholm, 2012).

Studies on perceptual causality have shown that inferring the causality between cues is an automatic process, which takes place at early stages of sensory information processing. As it was discussed earlier cue integration is largely but not completely independent from higher level cognitive processes, (Wei & Körding, 2012). For example, when subject are told that multisensory stimuli with disparities are from different sources sensory integration is significantly reduced. Hence, it has been postulated that causal inference in perception resides at the intersection of cognitive and perceptual processing, (ibid: p18). However, as Wei and Körding discuss, the fact that causal inference seems to play an important role in cue combination and sensorimotor control raises a more primordial question which is how various cognitive and sensorimotor processes interact. Furthermore, the links between causal inference and cue combination is poorly understood, and the role of causal inference in cue combination has only been examined in very simple problems (i.e. small number of cues).

3.5 Cross-modal correspondence: Methodological Consideration

Some of the first empirical studies known that provided evidence for the existence of cross-modal correspondence can be traced back to *sound symbolism* studies. Sound symbolism is the term that was used to describe the phenomenon of association between speech sounds and shapes. The most well-known such experiment was conducted by Köhler, (1929). Köhler, observed that most people tended to associate phonetic acoustic morphology with visual shapes. The experiment involved showing two images to participants a rounded and a spiky and asking participants which image they thought was the meaningless word *"Baluma"* and which was *"Takete"*. He observed that the majority of the participants would associated *Baluma* with the rounded shape and the word *Takete* with the angular shape. Similar findings were observed by another study conducted the same year, were it was founsd that the words "mil" would be systematically associated with small and the word "mal" with large visual objects, (Sapir, 1929). Many studies followed on the topic of phonetic to visual symbolism, for a review see (Nuckolls, 2010; Parise & Pavani, 2011). The growing empirical evidence suggesting that there is consistency between individuals in making cross-modal associations between acoustic and visual features pointed that these phenomenon must be underpined by natural constrains that determine the mapping, (Parise & Pavani, 2011). In recent years, we have adequate empirical evidence to support this view.

A few years latter, psychophysisist also started to investigate these effects. For example, Stevens & Marks, (1965) conducted a study where participants were asked to adjust sound amplitude using a potentiometer to levels to different levels of light intensity and vice versa adjust light intensity in response to audio stimuli of different levels of amplitude levels. The results of the study showed that participants reliably matched loudness and brighness. The same experiment was contacted with childrens five years old who also reliably matched light and sound intensity, (Bond & Stevens, 1969). More recent studies have investigated the ability of children to match cross-modal feature diamensions, (Mondloch & Maurer, 2004; Smith & Sera, 1992; Walker *et al.*, 2010). For example, it has been shown that childrens as young as 2 years old can match reliably visual size to sound loudness (Walker *et al.*, 2010) visual size to pitch (Mondloch & Maurer, 2004). Empirical evidence suggestes that size to loudness is one of the first consistent cross-modal mappings exhibited by childrens at the age of 2, while other cross-modal features dimensions are reliably matched somewhat at a older age, (Spence, 2011). Although other studies suggest that cross-modal associations such as that of luminance (visual brightness) to pitch are common between humans and chimpatzees, which suggests that these cognitive associations evolved before the human spieces and supports the view that the phenomenon of cross-modal correspondence is not culturally learned or a linguistic phenomenon, instead is a fundamental cognitive feature shared perhaps by all primates, (Ludwig, Adachi, & Matsuzawa, 2011). Other studies have shown that consistent associations is exhibited by children between sound and visual feature dimension also exist between other modalities, such as vision and touch, audition and body movement, audition and taste/flavour, (for full list of references see Spence (2011)). However, due to the fact that this study focuses on correspondences between audition and vision our discussion is limited studies related to these to modalities.

We could distinguish between two main experimental paradigms for the study of cross-modal correspondence, speeded and non-speeded, (Marks, 2004). Marks, explain that the since researchers in cognitive science began to view the human mind as a computer that processes sensory input from different modalities at various levels of perceptual processing ranging from psychophysical to psycholinguistic, it was realised that the speed with which the brain processes information can be used to study at which levels of processing cross-modal correspondences occur and how people attend to particular features of sensory stimuli. The speeded classification paradigm which is currently the most commonly used method for the study of selective attention to multidimensional varying stimuli was first proposed by Gerner in the 1960's, (ibid :p 85). The speeded classification paradigm was initially devised by Gerner in order to study selective attention in multidimensional stimuli of a single modality. Since the paradigm developed by Gerner has been adopted in the study of selective attention in multimodal contexts.

The speeded classification/identification paradigm subjects identify or classify particular characteristics of a stimulus as quickly as possible. Most commonly speeded classification experiments involve two tasks. In the first instance subjects are asked to indicate as quickly as possible the state/magnitude of a given sensory feature (e.g. pitch is low/high or loudness is low/loud) while all other feature dimensions of the stimuli are kept constant. The reaction times, gathered by the first task which is known as *baseline condition*, while a second task is performed were two or more feature dimensions vary simultaneously, which is known as selective attention task. According to Marks, the baseline task is considered as an identification task because we only have two states of the stimuli and possible responses, while the selective attention task is considered as a classification task because there are four possible combination of the stimuli but only two responses, (ibid: p. 86). For example, the stimuli in the selective attention task might be auditory varying along two dimensions (loudness and pitch) or it might be visual varying in terms of size and brightness. Varying two dimensions of the stimuli simultaneously creates four possible stimuli combination, low pitch and loudness, high pitch/low loudness, low pitch/high loudness, high pitch and loudness. In speeded experimental conditions the subjects reaction times in the baseline task is compared to the performance of the selective attention task. If subjects' reaction times of the selective attention task is greater than in the baseline task due to the variation of irrelevant feature dimension then it is assumed that there is an interference between these feature dimensions known as *Gerner's interference*. Gerner's interference has often be attributed to the hypothesised characteristics of early sensory processing.

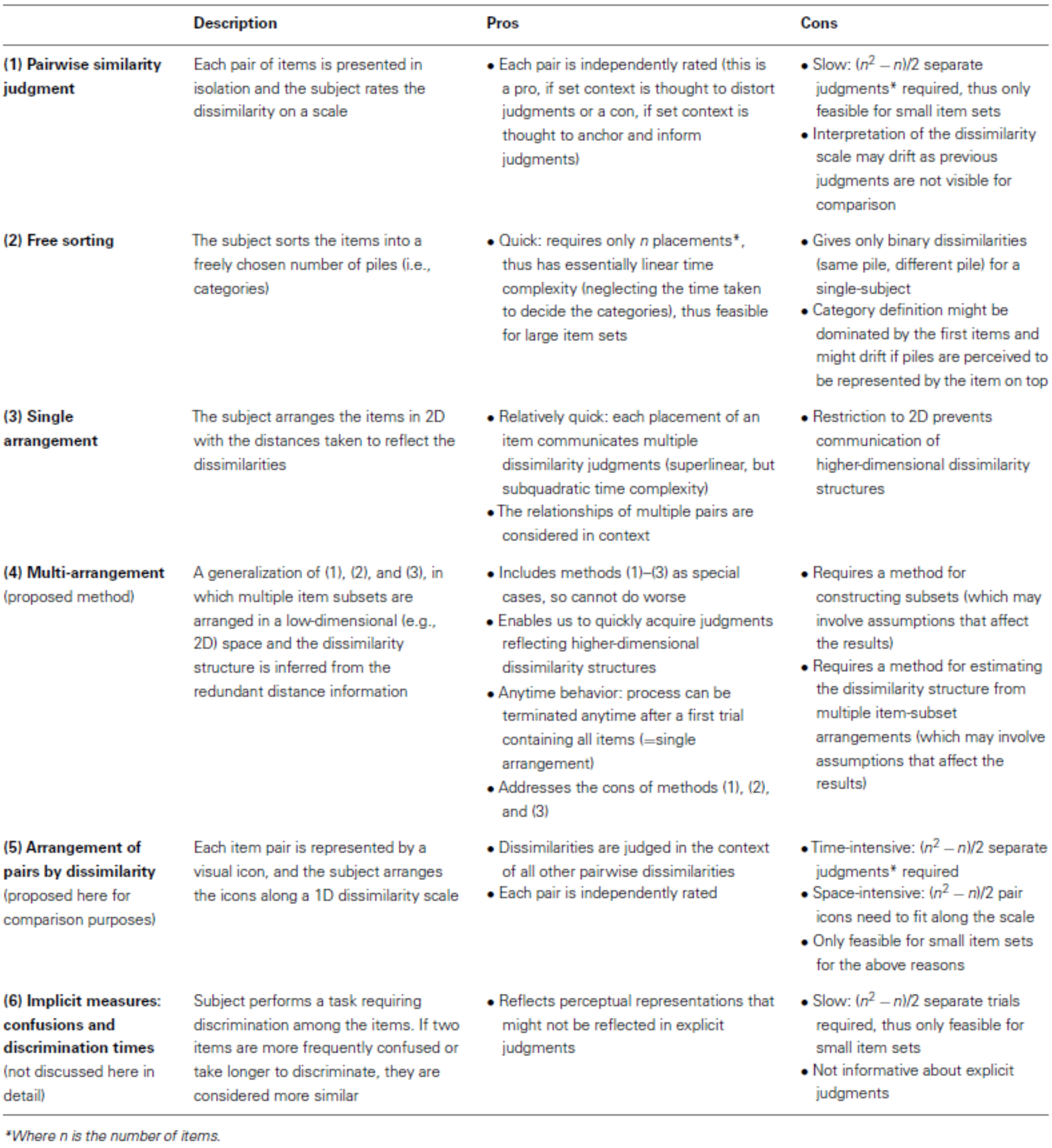
The Gerner's interference paradigm, suggests that if two irrelevant feature dimension vary simultaneously then subjects should not be able to selectively attend to either dimension and perform the classification task fast and accurate. However, if the feature dimensions being tested are well differentiated from the irrelevant dimensions that vary simultaneously, this is considered as an indication that differentiation of the feature dimension happens at early sensory processing and so the subjects are not confused by the irrelevant dimension, so Garner's interference is absent. Soon this paradigm was applied to the study of cross-modal attention. Initially it was assumed that there should be no Gerner's interference when classifying stimuli based on variation in one feature dimension while irrelevant features vary simultaneously. However it was soon realised that subject's Gerner's interference was also present across modalities. For example, when subjects were instructed to classify visual brightness while sound loudness or pitch vary simultaneously participants performance was decreased, which suggest interactions between feature dimensions across the different sensory modalities, (ibid: p. 89).

An even older methods used to study cross-modal correspondence were concerned non-speeded responses. There are a number of different methods of unspeeded tasks for studying cross-modal correspondence including pairwise similarity judgments, forced matching, multiple item arrangement and confusion task, (see Figure 29 for a description and the pros and cons of each method). In fact speeded methods are particularly suitable for studying correspondence in pre-attentive perception (correspondence at a psychophysical level). However correspondences as it was discussed earlier can occur at a structural, statistical and semantic levels. In unspeeded tasks conceptual properties and higher level cognitive processes (such as semantic or psycholinguistic levels) become more influential in the subjects' responses, (Goldstone, 1994). Therefore, unspeeded tasks are more suitable for studying higher level cognitive processes. Moreover, it could be argued that studying cross-modal correspondence in applied context (e.g. cross-modal mapping for musical interaction) could have different requirements than that of studying early stages of sensory processing. Therefore, it could be argued that speeded and unspeeded methods are not mutual exclusive but complementary. For example, in the section 2.4 we discussed a number of experiments that used the free-tracing paradigm, see (Amelynck, Maes, Martens, & Leman, 2014; Athanasopoulos & Moran, 2013; Caramiaux *et al.*, 2014; Caramiaux *et al.*, 2010a, 2010b; Godøy *et al.*, 2006b; Godøy, 2006; Kussner & Leech-Wilkinson, 2013; Küssner *et al.*, 2014; Küssner, 2014). Another recent study have used a multiple arrangement task to investigate audiovisual correspondence between musical timbre and visual shape, (Adeli, Rouat, & Molotchnikoff, 2014). The following section discusses a number of issues related to the study of similarity as a cognitive process for the categorisation of sensory stimuli.

3.6 Similarity and feature selection in categorisation

This section focuses mainly on similarity of form and structure as it relates to perception, cognition. Similarity as a process has many potential applications and it is important to various context and to a diverse range of disciplines. Similarity is often considered as a fundamental mechanism that enables humans as well as computers to grouping things together, learning, making comparison, and develop categories. As Quine said "there is nothing more basic to thought and language than our sense of similarity", (Ibid: 1969: p. 116). The notion of similarity appears fundamental to theories of perception, learning, and judgement, but it has also been argued that it has weaknesses. However many researchers and scholars have argued that similarity can only suffice to explain a limited number of mental phenomena. According to some similarity is (i) too vague and unconstrained to provide a complete explanation of how we form categories or make comparisons, (ii) not elaborate enough to account for all categories, and (iii) is highly context depended (Goldstone, 1994). It has been argued that more or less everything might be seen in a sense similar to everything else in some respect. Obviously that possess a serious challenge to similarity (ibid: 127).

When Hume introduced the term *resemblance* for discussing similarity he recognized that similarity as a process is of great importance in making comparison between objects and concepts, but he also noticed that not all properties have equal weight in assessing similarity. He argued that when a quality becomes very common amongst many objects or concepts these properties lose their significance in establishing links between two or more objects. According to Hume these properties are given less weight as a criteria in a similarity judgement, as we are inclined to nominate less attention to these features because the possibilities of choice become immense. Based on these it could be said that having common properties might not be adequate to explain similarity. Another approach to consider resemblance/similarity is the concept of likeness, were a number of objects are comparatively tested for similarity against each other. For example (a) might be considered to be more like object (b) rather than (c), (Gamboa, 2007)Another alternative to property based resemblance is that resemblances occur on higher-level attributes this involves abstract relations or causal-structural. In abstract resemblance we have emergent or contextual



	Description	Pros	Cons
(1) Pairwise similarity judgment	Each pair of items is presented in isolation and the subject rates the dissimilarity on a scale	• Each pair is independently rated (this is a pro, if set context is thought to distort judgments or a con, if set context is thought to anchor and inform judgments)	• Slow: $(n^2 - n)/2$ separate judgments* required, thus only feasible for small item sets • Interpretation of the dissimilarity scale may drift as previous judgments are not visible for comparison
(2) Free sorting	The subject sorts the items into a freely chosen number of piles (i.e., categories)	• Quick: requires only *n* placements*, thus has essentially linear time complexity (neglecting the time taken to decide the categories), thus feasible for large item sets	• Gives only binary dissimilarities (same pile, different pile) for a single-subject • Category definition might be dominated by the first items and might drift if piles are perceived to be represented by the item on top
(3) Single arrangement	The subject arranges the items in 2D with the distances taken to reflect the dissimilarities	• Relatively quick: each placement of an item communicates multiple dissimilarity judgments (superlinear, but subquadratic time complexity) • The relationships of multiple pairs are considered in context	• Restriction to 2D prevents communication of higher-dimensional dissimilarity structures
(4) Multi-arrangement (proposed method)	A generalization of (1), (2), and (3), in which multiple item subsets are arranged in a low-dimensional (e.g., 2D) space and the dissimilarity structure is inferred from the redundant distance information	• Includes methods (1)–(3) as special cases, so cannot do worse • Enables us to quickly acquire judgments reflecting higher-dimensional dissimilarity structures • Anytime behavior: process can be terminated anytime after a first trial containing all items (=single arrangement) • Addresses the cons of methods (1), (2), and (3)	• Requires a method for constructing subsets (which may involve assumptions that affect the results) • Requires a method for estimating the dissimilarity structure from multiple item-subset arrangements (which may involve assumptions that affect the results)
(5) Arrangement of pairs by dissimilarity (proposed here for comparison purposes)	Each item pair is represented by a visual icon, and the subject arranges the icons along a 1D dissimilarity scale	• Dissimilarities are judged in the context of all other pairwise dissimilarities • Each pair is independently rated	• Time-intensive: $(n^2 - n)/2$ separate judgments* required • Space-intensive: $(n^2 - n)/2$ pair icons need to fit along the scale • Only feasible for small item sets for the above reasons
(6) Implicit measures: confusions and discrimination times (not discussed here in detail)	Subject performs a task requiring discrimination among the items. If two items are more frequently confused or take longer to discriminate, they are considered more similar	• Reflects perceptual representations that might not be reflected in explicit judgments	• Slow: $(n^2 - n)/2$ separate trials required, thus only feasible for small item sets • Not informative about explicit judgments

Where n is the number of items.

Figure 29. Methods used to assess dissimilarity and the advantages and disadvantages of each method (Kriegeskorte & Mur, 2012).

property that establishes similarity or might be seen as resembling. For example empirical research findings show that subject judged a raccoon and a snake to be more similar when the word pet was presented above the two representation than when no context was provided, (Barsalou, 1999).

Goodman who did theoretical work in analysing similarity/resemblance suggested that similarity or likeness between two units such as X and Y cannot be established until a third contextual/psychological property Z defines in which respect X and Y are compared, (Goodman, 1972). Consequently, when a person is asked two make a similarity judgement between two units without defining Z property guessing what Z property/ies will be used in the judgement is hard to predict. Goodman suggested that determining what criterion is psychologically important when making a similarity judgements is a hard task (Goldstone, 1994). It is therefore reasonable to say that similarity does not only depend on structural and objective features but it also has a psychological dimension to it.

Figure 30 illustrates nicely the problems deciding which attribute will be psychological important in a similarity judgement. Similar problems arise in computational feature matching. For example when using criteria for sound selection/ identification based on similarity or dissimilarity features. In the example of the figure below we could distinguish between two features that would suggest different similarity classification, one based on the outline or the second based on texture. Determining which on will be psychologically important it is a hard task as the following experiment suggests. Students were asked to classify the images in Figure 30 in two group, were each group consisted of two images. One third of the students grouped A-B, C-D the one third A-D, B-C and the other one third did not answer the question, (Jehan, 2005). These demonstrates some of the difficulties that might be faced and the decisions that have to be made, when it is required to classify objects or other entities based on set of criteria.

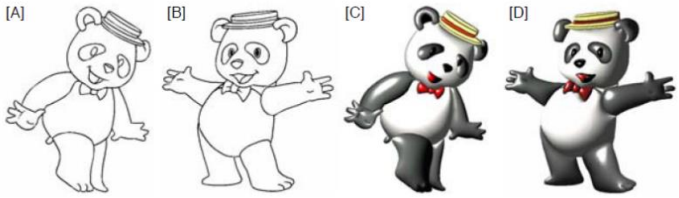


Figure 30. Simple example of the feature selection problem when making a similarity judgment in the visual domain (Jehan, 2005).

The selection criteria based on which a subset is selected is in a dialectic relationship with interest and intentions of the subject who makes the similarity judgement. Interests and intentions that define selection criteria might vary depending on context. As for example Barsalou findings show that subjects judged a raccoon and a snake to be more similar when the word pet was presented above the two representation than when no context was provided, (L. W. Barsalou, 1999), or similarly with the example we have seen earlier were a cat and dog are considered more similar if we introduce a bird in the comparison that if we just compare the two.

3.7 Conclusions

This chapter reviews literature that investigates the cognitive mechanism that mediate cross-modal binding and congruency effects. As it was discussed in this chapter, at first glance, sensory modalities might appear to be of such disparate nature. Nevertheless sensory modalities appear to converge and overlap in a number of ways. Understanding better the equivalences between the sensory modalities can deepen our knowledge of how we perceive and experience information in the environment. Research findings have shown that for a successful multimodal binding to occur, the relationship between two modal objects and their attributes must be plausible in terms of: (i) of time (i.e. synchrony), (ii) space (i.e. collocation), and (iii) prior experience of similar events and phenomena. The underlying principles that mediate cross-modal binding and congruency effects, particularly beyond spatiotemporal integration are poorly understood. The need for systematic empirical work to shed light on the multimodal nature of perception is evident.

Furthermore, as I have argued, different methods are better suited for studying cross-modal correspondence at different hierarchical levels of sensory information processing. Speeded methods are well suited for studying structural similarity and statistical correspondence, while unspeeded methods are more suitable for studying cross-modal feature correspondence at either the statistical or the semantic level. Hence, speeded and unspeeded methods are complimentary. Moreover, in applied context such as in human computer interaction and the design of cross-modal mappings both methods can be applied to assess the effectiveness of the mapping between the users' sensorimotor input and the system output.

4 Morpheme: a multidimensional sketching interface for the control of corpus-based concatenative synthesis

This chapter presents the design and the implementation of *Morpheme*[3]. *Morpheme* is a sketching interface for the control of concatenative synthesis. The main motivation for this investigation is (i) to explore appropriate models of interaction for efficient exploration of the audio corpus, and (ii) to develop perceptually meaningful mappings to enable practitioners to create novel sounds by specifying the perceptual characteristics of the sound they want to synthesise in visual terms.

According to the literature reviewed in Chapter 2 many researches support that defining a framework for visual representation of sound is important for pedagogical, analytical and compositional purposes (Blackburn, 2009; Couprie, 2004; Patton, 2007; Smalley, 1997). Furthermore, the literature suggests that mapping user sensorimotor input to sound parameters is a central element in the design of a music system and a determinant of the creative potential of the system. Although in the digital domain we have a lots of freedom in determining how the performer's sensorimotor inputs are associated to the sound synthesis or music parameters, it is exactly this freedom that often makes the performer feel confused about the affordances of the interface and affect the appreciation and level of control over the sound produced by the systems (Leman, 2008). Hence, there is need for a design framework that will allow instrument designers to make justified decisions about the mapping of users sensorimotor inputs and sound parameters.

In Chapter 3 we discussed a large body of research suggesting that humans exhibit consistent patterns of crossmodal correspondences between audio-visual features in different sensory modalities. Therefore a reasonable step forward would be to investigate whether crossmodal correspondences exhibited by humans can be successfully be applied in the design of crossmodal mapping for intuitive representation of and interaction with sound. In order to investigate whether cross-modal correspondences could be informative in the context of designing mappings for musical interaction, Morphemes' mappings used to render the sketches that are created by the practitioner into sound were informed by empirically validated audio and visual correlates. The mapping is described in detail in section 4.6.2.

Before I begin describing the design and implementation of *Morpheme* I would like to

[3] For a standalone version of Morpheme see appendix A and for a Max/MSP patch see appendix B on the DVD appendix which was submitted as part of this thesis.

discuss the reasons for choosing sketching as interaction paradigm and concatenative synthesis as the synthesis method.

4.1.1 Why Corpus-Based Concatenative sound synthesis?

After exploring a number of sound synthesis methods during my first exploratory year of this doctoral project concatenative synthesis attracted my attention primarily because unlike other synthesis methods were the sound is represented by low-level signal processing parameters related to the given sound generation or processing unit, in the context of concatenative sound synthesis, sounds are represented using sound descriptors related to perceptual parameters. This opens up new possibilities for user interaction, as users can synthesise sounds by describing the properties of sound in terms of perceptual related cues rather than through direct manipulation of synthesis parameters. The perceptually relevant computational representation abstracts the complexity of manually having to model the relationships between low level synthesis parameters to perceptual related auditory parameters, a problem that was discussed in Chapter 2, section 2.4 .

Secondly, another reason for choosing to base my research on user interaction with concatenative synthesis was that this synthesis method has not been as widely explored as other synthesis methods in the context of creative sound and music practice.

Thirdly, granular synthesis methods such as concatenative synthesis it is a good alternative to physics-based models for the synthesis of natural sounding auditory textures, as the source audio used for sound synthesis can derive from any recorded natural sounds. Hence, natural sounding sounds with complex timbre characteristics can be synthesised. The synthesis of natural auditory texture using physical modelling and statistical learning is possible (Dubnov & Bar-Joseph, 2002; McDermott & Simoncelli, 2011; Ren, Yeh, Klatzky, & Lin, 2013; Ren, Yeh, & Lin, 2013; Strobl, Eckel, & Rocchesso, 2006). A high degree of realism can be achieved using physics-based models, however the models tend to have high demands in terms of computational for executing the models, require complex parameter estimation and tuning and often real time control of the models parameters it is either not possible and /or requires advanced technical understanding of the synthesis method (i.e. specialist skills). Further the need to simulate many types of materials and interactions between material each with different characteristic and distinct mechanical and sound properties poses a challenge to the synthesis of natural sounding audio with computational means.

4.1.2 Why sketching was considered a good model of interaction with sound synthesis?

This section explains the reason for having considered sketching as the model for user interaction with concatenative synthesis. Our ability to express our thoughts and ideas in visual terms cannot be underestimated. As soon as a few marks are made on a piece of paper the brain reacts to them. These reactions develop naturally based on the experience of a lifetime of visual associations and a continuous effort of perceptual mechanisms to identify form, structure, and meaning. Auditory perception is not different in this respect. During these process unexpected thoughts occur about where and/or how to make the next actions. So it could be said that visualizing concepts through sketching has an inherent experimental and speculative nature.

The exploratory nature of sketching as well as its' ability to inspire imagination and support thinking makes sketching an appealing medium for interaction with sound. After all sound design and musical composition are exploratory processes that require imagination, creativity and ingenuity. As it was discussed in Chapter 2, evidence show that composers, particularly in electroacoustic music, use sketching and drawing as a medium for conceptualisation of ideas at different stages of the composition process. Many researchers have argued that creating tools for visual representation of sound is very important for pedagogical, analytical and compositional purposes (Blackburn, 2009; Couprie, 2004; Patton, 2007; Smalley, 1997). When a person is listening to sounds while seeing the sounds being illustrated by iconic means, perceptual links are established between sound and image that might have not been established otherwise. This in turn might help learners assimilate information quicker and better and inspire thought and imagination.

Sketching creates an artefact that the user can iteratively refine, providing immediate primary feedback. Another advantage of sketching as a paradigm for interaction with sound is that while sound is temporally bound, sketching breaks the temporal dependency of sound and allows sounds' temporality to be manipulated in spatial terms. It would also be possible to imagine that sketching could be used to control other time-based interfaces such as visual query of images. In comparison, gestural input retains the temporal dependency, making temporal editing without the use of graphical tools difficult. Therefore for live performance of music and real-time expressive control of sound parameters gestural input might be more appropriate. However, it could be argued that sketching as a model of interaction for the control of sound synthesis parameters is better suited to pre-production and conceptualisation purposes than gestural control. This is due to the fact that sketching creates a visual representation of the sonic composition that can be used to refine sound design and musical ideas and to discuss ideas between professionals. Due to its representational and communicative character sketching can be used for both analytical and pedagogical purposes. Hence, sketching is a well suited medium and model of interaction for communicating sound design and musical ideas to the computer and iteratively refine them. Animation is also a very promising direction for the control of sound synthesis. Animation combines the temporal dependency of sound with the pictorial dimensions of sketches or images in general. However unlike sketching, animation requires more training than creating simple sketches does.

Although sketching has been widely explore as a medium for interaction with sound synthesis and musical composition, Morpheme is (to my knowledge) the first attempt ever made to use sketching as a model of interaction for concatenative synthesis.

4.1.3 The analysis of the sketch

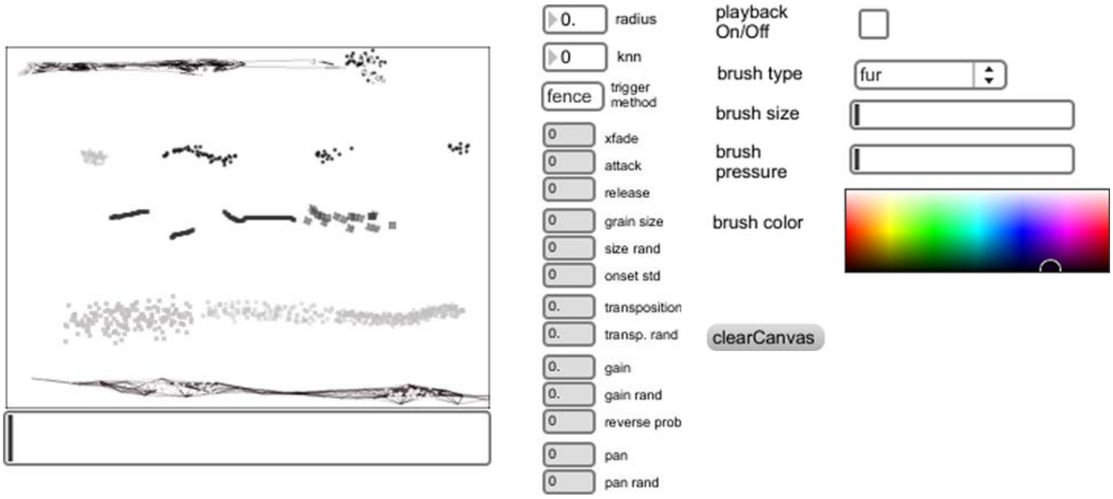
There were two options I considered for extracting information from the sketch. First, the system could use the vector data resulting from the brushstrokes which are performed on the canvas, similar to the Hyperscore and HighC systems that were discussed in Chapter 2. The second method considered was using optical and image analysis. Due to a number of reason discussed in this section I decided to focus on optical methods and the data that derive from the image analysis of the sketch as opposed to the history of the construction of the sketch (i.e. command list of brushstroke vector coordinates, and brush settings). This decision was made considering that the sketch is what is important and what the practitioner sees and works with. I consider that focusing on the analysis of the sketch would make it easier to interpret what is perceptually relevant for the practitioner as oppose to the history of the construction of the sketch. As the vector data that result from the act of sketching on a digital canvas leads to a way more fragmented account of what is perceptible in the sketch. Moreover trying to reconstruct what would be perceived by the practitioner from the sketch's vectors is considerably more difficult. Hence, optical analysis seems a more attractive route, although as we will discuss later there are limitations in what can be achieved using optical analysis.

4.2 Morpheme First Version: Design

Morpheme aims to extend the compositional and sound design potential of corpus-based concatenative synthesis. *Morpheme* allows the control of the concatenative synthesis CataRT[4] (see (Schwarz, 2004)) through the act of sketching on a digital canvas. *Morpheme* was developed

[4] http://imtr.ircam.fr/imtr/CataRT

using the graphical programming environment Max MSP. The implementation of concatenative synthesis *Morpheme* interfaces with (i.e. CataRT), works by segmenting a number of audio files into small units. The units are then analysed, tagged with the analysis data and stored in a database. Synthesis is accomplished by recombining the audio units from the database based on the input data. In the context of this project, the input data stream results from statistical analysis of the HSL matrix of the canvas which is scanned from left to right. Figure 31, shows the main user interface of the first version of morpheme. This includes the canvas, four controls for the brush attributes (i.e. type, size, pressure and color) and a number of sound processing controls of CataRT.



0. radius
0 knn
fence trigger method
0 xfade
0 attack
0 release
0 grain size
0 size rand
0 onset std
0. transposition
0. transp. rand
0. gain
0. gain rand
0 reverse prob
0 pan
0 pan rand

playback On/Off □
brush type: fur
brush size
brush pressure
brush color

Figure 31. Canvas, brush and other user interface controls of Morpheme.

Mapping audio and visual descriptors for audio-unit selection

This section discusses the approach used for determining the mapping between audio and visual descriptors for the selection of audio segments from the corpus. The canvas is scanned vertically stepping through each column of the pixel matrix from left to right. Following the score conversions, the x-axis represents time and the y-axis represents frequency. The features used for the retrieval of audio units include periodicity, spectral flatness, loudness and spectral centroid. The features extracted from the sketch include brightness, compactness, vertical position, and variance. Table 2 and Table 3 illustrate an analysis of all of the features used for association between visual and audio stimuli. The analysis aimed to identify the ontology of the features and their affordances. By ontology I am referring to the actual physical properties of the feature, while by affordances I refer to the polar perceived recognizable states/transitions that are afforded by the feature.

Table 2. Analytical cataloguing of canvas features in terms of parameters, domains and affordances.

Canvas features		
Visual Features	***Ontology***	***Affordances***
Color intensity	Texture energy	Dark-Bright
Compactness	Texture distribution	Dense-Sparse
Path	Spatial distribution of texture	Continuous-Discontinuous, Horizontal - Vertical
Variance	Spatial distribution	Invariant-Variant

Table 3. Analytical cataloguing of concatenative synthesis features in terms of parameters, domains and affordances

Audio-unit selection features		
Audio Features	***Ontology***	***Affordances***
Periodicity	Spectral distribution	Periodic – Non-periodic
Spectral flatness	Spectral distribution	Symmetric - Asymmetric
Loudness	Energy	Quiet- Loud
Spectral Centroid	Spectral distribution	Bright - Dark

Identifying the ontology of each parameter and its affordances can be informative in the process determining the associations of a cross-modal mapping, as it might expose any shared physical perceptual properties between auditory and visual dimensions. Taking a deeper look at the nature of the parameters in relation to the ontology, we can see that in both tables there are some recurrent themes such as energy, distribution and barycentre of distribution. Brightness and loudness are indicators of energy. Density and spectral flatness indicate morphological aspects of texture and spectral distribution. Path and spectral centroid both relate respectively to the perceived barycentre of texture and spectral distribution. Variance and periodicity are both parameters which depend on two domains. Sound periodicity depends on temporal and spectral organisation, while graphical periodicity depends on both coordinate axes and the texture organisation. Looking further at the affordances of the features can also be informative as to how sound and graphic descriptors could be associated. Table 4 illustrates the associations between audio and visual features which were developed based on the analysis described above. Figure 32 illustrates four examples that show some extreme cases of how different statistical distributions of the pixel data are associated to the features of the audio units in the database.

Table 4. Associations between graphical and audio-units features.

Visual Features	Audio Features
Variance	Periodicity
Compactness	Spectral flatness
Brightness	Loudness
Vertical position	Pitch
Duration	Horizontal length

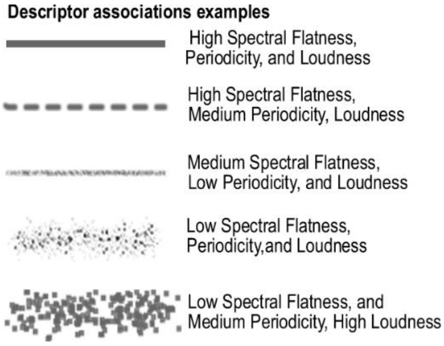

Descriptor associations examples
High Spectral Flatness, Periodicity, and Loudness
High Spectral Flatness, Medium Periodicity, Loudness
Medium Spectral Flatness, Low Periodicity, and Loudness
Low Spectral Flatness, Periodicity,and Loudness
Low Spectral Flatness, and Medium Periodicity, High Loudness

Figure 32. Five examples of the interpretation of the mapping association between visual and concatenative synthesis retrieval descriptors.

The mapping in the first version of Morpheme was devised using a very idiosyncratic approach. It soon became obvious that the mapping of Morpheme must be based on the empirical findings rather than try to devise a new theoretical framework and a heuristic method. In the second version of morpheme this issue was addressed by informing the association of the audio-visual mapping using empirically validated audio-visual feature correlates, for more information see section 4.6.3.

4.3 Morpheme First Version: Implementation

4.3.1 System Architecture

This section presents the implementation of the first version of Morpheme. Figure 33 illustrates the architecture of Morpheme. The digital canvas where the sketches are drawn is divided in five horizontal bands, as shown Figure 33. Statistical analysis is performed on each of the 5 sub-matrices using a window. The canvas is scanned vertically stepping through each column of the pixel matrix from left to right, for more information about the algorithm for visual feature extraction, please see section 4.3.2. The result of the analysis of each area results in five feature vectors what we will refer to as a *target vector* as they are used as targets for the retrieval of audio units from the corpus. Each target vector is used as an input to a selection algorithm that finds the best match between the vector and the descriptors of the pre-analysed audio-units in the corpus. Each selection algorithm selects audio units only from a limited part of the audio corpus which is equal to the one fifth of the overall frequency spectrum of the corpus as the vertical axis is mapped to pitch.

The decision to divide the canvas into five equally spaced frequency bands was made for two reasons. First, I had not yet identified an effective image analysis method to estimate the barycentre of an image (e.g. brushstrokes' vertical position on the canvas). The second reason that led to this decision was to enable practitioners to layer more than one sounds. As it will be discussed in section 4.4.1 informal evaluation with my supervisory team showed that the division of the canvas was confusing, hence in the second version of Morpheme the canvas consisted of a single band and the vertical centroid of brushstrokes was used to determine the target frequency for querying audio-units from the database.

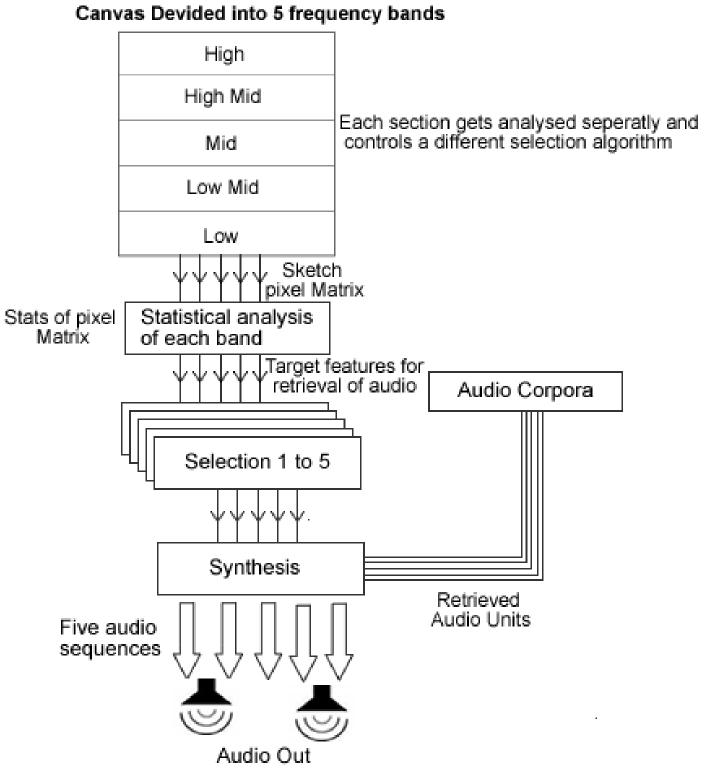

Canvas Devided into 5 frequency bands
High
High Mid
Mid
Low Mid
Low
Each section gets analysed seperatly and controls a different selection algorithm
Sketch pixel Matrix
Stats of pixel Matrix
Statistical analysis of each band
Target features for retrieval of audio
Audio Corpora
Selection 1 to 5
Synthesis
Retrieved Audio Units
Five audio sequences
Audio Out

Figure 33. System architecture for retrieval of a variable number of audio-units.

4.3.2 Audio-Visual Descriptor Extraction and Mapping

As mentioned earlier, synthesis is accomplished by recombining the audio units from the database based on an input data. In the context of this project, the input data stream results from statistical analysis of the intensities (i.e. greyscale) and HSL (Hue Saturation and Lightness) matrix of the canvas. During playback windowed analysis is performed on the greyscale version of the sketch. As mentioned earlier, a window scans the sketch from left to right one pixel at every clock cycle, the rate of which is determined by the user. Only the areas of the canvas that are within the boundaries of the window area are subjected to image analysis.

The window width is determined by the user, while the window height is estimated by dividing the window height by 5, because each canvas band is equal to the one fifth of the overall height of the canvas. As mentioned above the window's width can be determined by the user, however the default size of the analysis window is 5 pixel wide by 48 pixel height.

Table 5 shows all the features that are extracted from the canvas and explains how they are estimated. *Morpheme* uses two computer vision algorithms to detect the centre of the painted area and the texture variance (cv.jit.centroid, cv.jit.framesub), see[5].

Table 5. Shows how visual features are estimated.

Visual Features	Windowed analysis
Variance x-axis	Estimated based on the amount of change between consecutive frames.
Vertical position	Estimated based on the centroid of the pixels which are ON in a binary image.
Texture variance	Estimated based on variance of the pixel intensities.
Brightness	Mean lightness, occluding background pixels (i.e. white pixels)

Table 6. Associations between graphical and audio-units features.

Visual Features	Audio Features
Variance x-axis	Periodicity
Texture variance	Spectral flatness
Brightness	Loudness
Vertical position	Pitch
Duration	Horizontal length

[5] http://jmpelletier.com/cvjit/

4.4 Issues in the first version of Morpheme

The first version implementation of Morpheme was more a proof of concept rather than a complete application. After discussion and informal trials of the software and the user interface with the supervisory team and colleagues, a number of issues were identified which driven the design and implementation of a second version. We could distinguish between two types of issues, those related to the design and others to the implementation. These are discussed below.

4.4.1 Design Issues

The mapping in the first version of Morpheme was devised using a very idiosyncratic approach. As I continued exploring literature related to the question of cross-modal mapping, I started to discover a plethora of studies that provided empirical evidence that suggested that there is a consistency in the audio and visual feature dimensions humans perceive as good correlates. It soon became obvious that the mapping of Morpheme must be based on the empirical findings rather than try to devise a new theoretical framework and a heuristic method. Further, it was also realised that similar empirical methods used in to study cross-modal correspondences and similarity between structural features can be extremely useful for the evaluation of multimodal interfaces for musical interaction amongst other human computer interface applications.

A second issue is related to the division of the canvas into five sections. The division of the canvas in the first version of Morpheme allowed to layer multiple sounds that it could be considered a desired feature. However, this feature was at the expense of usability. The main usability issue related to the division of the canvas was that the user had no way of determining the boundaries of each section of and consequently which part of the sketch will be analysed together. Informal trials showed that the division of the canvas is confusing for the user as it was not related to either the way we would naturally perceive a canvas or it follows any conventions derive from the use of other mainstream audio interfaces and music notation. Two options were considered in order to solve this problem: (i) perform feature extraction on the entire sketch but at a blob level and for each blob use a different audio-unit selection and synthesis module, and (ii) perform feature extraction on the entire sketch and use one audio-unit selection and synthesis module. The second option was considered the most appropriate, given that the other approaches could be very demanding in terms of computational resources due to the potential of large number of simultaneous queries to the database for audio-unit selection. The first option was also rejected because it was deemed too complicated to work at an object level. First there are discrepancies between what a computer recognises as a blob and what humans recognise as objects, the mismatch between user and the computer's object recognition abilities could lead to user confusion. Second, imagine the scenario were seven blobs are aligned on the vertical axis, this would mean that image analysis will have to be performed separately for each blob, which will result in target for the selection of the audio-units and seven sound synthesis modules will be required in order to playback the retrieved units.

4.4.2 Implementation Issues

Due to the fact that the present version of Morpheme uses a series of five retrieval algorithms and synthesis modules and a set of feature extraction algorithms one for each area of the canvas running the system is extremely expensive from a computationally point of view. This is obviously an issue that has to be addressed as real-time interaction with the system is the main aim. A number of modifications were identified that could improve the computational efficiency of the system including: (i) reduce the number of retrieval and feature extraction and sound synthesis algorithms, and (ii) change the brushes methods used for sketching as they are implemented in javascript which is slower than Cycling 74 jitters native objects/methods which are developed using the C programing language. Further, in the second version of Morpheme, it will be necessary to implement a mapping for associating the distance between the target and selected feature vectors to the sound synthesis parameters. For example, when a target loudness is -20db and the selected audio-unit is -3db the amplitude of the selected audio-unit should be decreased by 17db in order to match the target.

For the selection of audio units from the corpus, CataRT uses k-Nearest Neighbour (k-NN) algorithm. k-NN works by estimating the shortest distance between the feature vector of the target (e.g. visual features extracted from the canvas) and the feature values of the units stored in the database (for more information on the algorithm, see Schwarz, (2004). A common phenomenon when an audio corpus consists of a relatively small number of audio units is that the distribution of the audio descriptions in the feature space is concentrated, forming dense clusters in some areas while the rest of the feature space is relatively empty. One problem that can be observed is that the target features values requested might be well outside the main clusters of the feature space and the target might not match any of the audio units in the corpus. In such case, the k-Nearest Neighbour (k-NN) algorithm will select the nearest unit that can be found in the feature space. On the other hand if the target feature vectors in a series of queries are well outside the main clusters, the selection algorithm will stay in the periphery of the clusters and will not access the clusters. This results in the following problems: (i) although the corpus might consist of many audio-units, it might be difficult to access the clusters, and (ii) two very different target queries (i.e. two brush strokes in the context of Morpheme), which both request feature vectors that are well outside the main clusters of the corpus, might retrieve the very same audio-unit. So in the new version of Morpheme a method should be devised to address this problem and improve the exploration of the audio corpus.

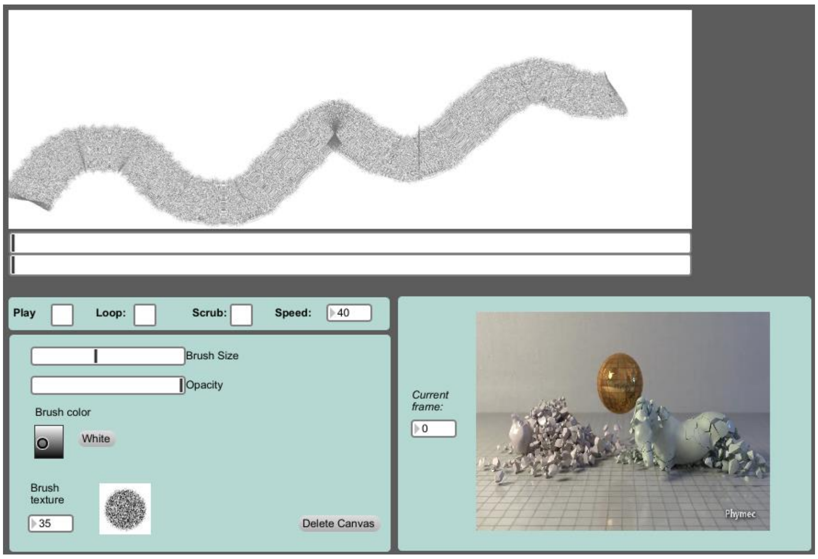
Finally, the mapping between audio and visual feature dimensions in the new version of Morpheme will be informed by empirically validated audio-visual correspondences. Consequently, the extraction of visual features will also need to be revised as algorithms used must be able to describe the required visual features.

4.5 Morpheme Second Version: Design

This section presents the second version of morpheme. A number of changes were made in the system architecture, the visual feature extraction, the mapping and the user interface. All of the changes aimed at improving the efficiency and the usability of the system. A number of algorithms were developed in order to improve the exploration of the audio corpus, see sections 4.6.6 and 4.6.7. In order to facilitate sound design applications such as sound for film and animation a number of features were developed. One of these features is the synchronisation between the canvas' timeline and duration of a video file. In the context of Morpheme, the duration of the sound being synthesised is determined by the length of the canvas along the x-axis, so we use the duration of the video file to resize the canvas so that the duration between the two is matched. Furthermore, the payback rate of audio and video playback are paired together, in order to synchronise the two media. This section provides an analytic description of the present system architecture and design of the graphical user interface of the second version of Morpheme.

4.5.1 The graphical user interface

Figure 34, shows a screenshot of Morphemes graphical user interface. We could distinguish between four main interface components in the second version of Morpheme's interface, the canvas, the timeline, the playback controls, the brush controls and the video display. Similarly to the first version of Morpheme the sketches drawn in the canvas are the main interface used for interacting with the audio sound synthesis engine.



Play □ Loop: □ Scrub: □ Speed: 40

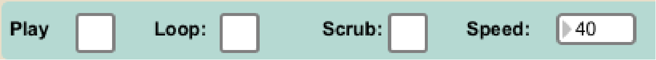
Brush texture 35

Current frame: 0

Figure 34. Morphemes' main graphical user interface.

The playback controls provide a number of function (see Figure 35) including:

1. **Play**: starts the analysis of the sketch which results in the data used to query the database and drive the sound synthesis engine.
2. **Loop:** repeats the entire length of timeline when the cursor reaches at the end of the timeline.
3. **Scrub:** functions freezes the cursor in a given location of the timeline. The users by clicking and dragging the cursor of the timeline can move the analysis window through the sketch to a desired position.
4. **Speed:** allows the user to determine the speed (in milliseconds). The speed controls the rate at which the analysis window moves from left to right though the timeline. This affects the rate at which Morpheme is sending target vectors to the selection algorithm for the retrieval audio-units from the database.



Play □ Loop: □ Scrub: □ Speed: 40

Figure 35. Screenshot of the user interface playback controls.

Brush Controls provide a number of function (see Figure 36) including:

1. **Brush size:** size of the brush
2. **Opacity:** opacity of the textured brush.
3. **Brush color:** color of the textured brush.

4. **White:** control can be used as an eraser. This is indeed used as a shortcut, since the colour of the canvas is white, it provides a quick means of selecting this colour means that this selection equates to the selection of an eraser brush.
5. **Brush selection:** by clicking and scrolling on the number box users can select from 41 different textured brushes.
6. **Clear Canvas:** erases the sketch from the canvas.

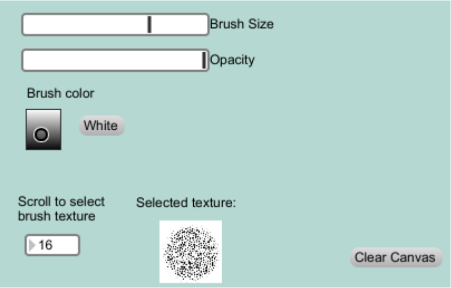


Figure 36. Screenshot of graphical interface for the control of the brush parameters.

Morpheme incorporates a number of features in order to enhance the usability of the interface for sound design applications. Using the menu, users can load a folder with video files, select the video file they desire from a dropdown menu, play the video and set loop points.

4.6 Morpheme Second Version: Implementation

4.6.1 System Architecture

Figure 37 illustrates the architecture of Morpheme. Statistical analysis is performed on the greyscale and HSL matrix of the canvas by scanning the canvas from left to right using a window, for more information on visual feature extraction, please see section 4.6.4. The decision to represent time in the vertical axis was driven by empirical findings that suggest that often humans associate the temporal dimension of sound to left to right visual trajectories (Athanasopoulos & Moran, 2013; Küssner, 2014). The analysis of the canvas matrix results in a feature vectors that describes visual attributes of the sketch and which is used as the target for querying audio-units from the database. The target vector is normalised to either the range of values defined by the minimum and maximum values of the descriptors of the audio-units stored in the corpus, or based on constrains defined by the user used to rescale the target, see Figure 38.

The interface for setting constrains is described in detail in section 4.6.7. The normalised target vector is then sent to the selection module that is part of the CataRT system (see (Schwarz, 2004)) which performs a k-NN search to find the best match between the target vector and the descriptor vectors of each audio unit which are organised in the database using k-d tree data structure. The distance between the descriptors of the audio-units and the target are computed by subtracting the two vectors, which are used to control the synthesis parameters, this process is described in detail in section 4.6.5

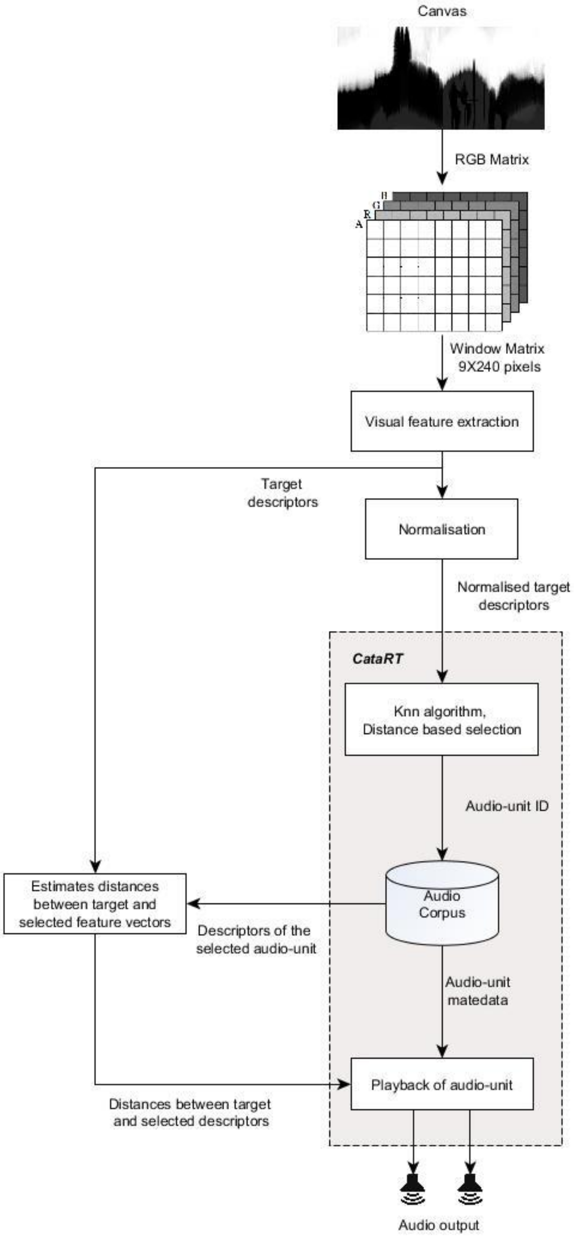


Figure 37. An overview of the architecture of Morpheme.

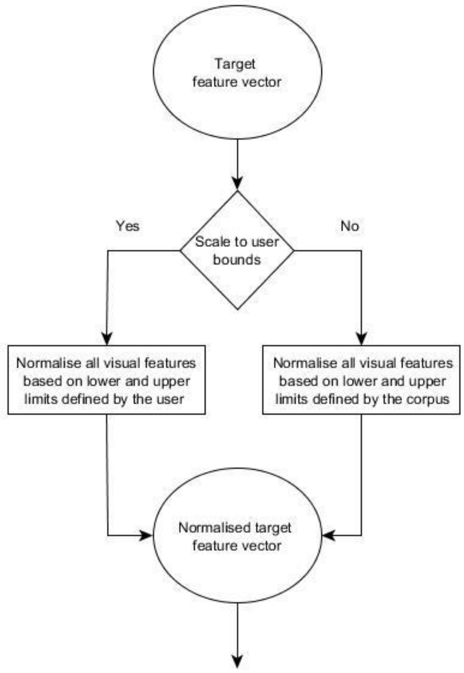

Target feature vector
Scale to user bounds
Yes
No
Normalise all visual features based on lower and upper limits defined by the user
Normalise all visual features based on lower and upper limits defined by the corpus
Normalised target feature vector

Figure 38. Normalising visual feature either based on the minimum and maximum value of the descriptors of the audio-units in the corpus or based on the limits defined by the user.

4.6.2 Parameter Mapping

In the current implementation of *Morpheme,* we can distinguish between two mapping layers. The first layer consists of a mapping between visual and auditory descriptors for the selection of audio units. The second layer consists of a mapping that associates the distances between audio and visual descriptors to the synthesis parameters.

4.6.3 Mapping Visual to Audio Features for Selection of Audio Units

A large body of research suggests that humans exhibit consistent patterns of crossmodal correspondences between stimuli features in different sensory modalities. Table 7 illustrates the findings from a number of studies that investigated the perceived correlation between auditory and visual features, which were discussed in 3.5. This consistency in the patterns of crossmodal correspondence humans exhibit is rather encouraging as it suggests that gathering more empirical could help in the formation of an empirical framework for the association between visual and aural structures. The most common audio-visual correlates according to the empirical findings presented in Table 7 were selected to develop two mappings that enable the selection of audio units from the corpus. After several informal trials where different feature set combinations were tested, these two multidimensional mappings were considered as the most reasonable ones. The distinction between the two mappings is that one is achromatic while the other is chromatic. The first mapping could be considered achromatic in the sense that the visual descriptors extracted from the sketch are estimated mainly based on volumetric and spatial attributes of the sketch. While the second mapping could be considered as chromatic due to the fact that all of the visual features extracted from the sketch are estimated based on color attributes. Table 8, shows the associations that each of the two mappings comprise.

Table 7. Empirically validated auditory and visual corresponding feature pairs.

Audio visual features	**Authors**
Size to Loudness	(Berthaut, Desainte-catherine, & Hachet, 2010; Küssner, 2014; Lipscomb & Kim, 2004; L. B. Smith & Sera, 1992; R. Walker, 1987)
Vertical position to Pitch	(Ben-Artzi & Marks, 1995; Bernstein & Edelstein, 1971; Chiou & Rich, 2012; Eitan & Timmers, 2010; Evans & Treisman, 2010; Hidaka, Teramoto, Keetels, & Vroomen, 2013; Marks & Eitan, 2012; Marks, 1989; Rusconi *et al.*, 2006)
Visual brightness to Pitch	(Bernstein & Edelstein, 1971; Hubbard, American, & Summer, 2007; Marks, 1983; Martino & Marks, 2000; Melara & O'Brien, 1987; Melara, 1989; Patching & Quinlan, 2002)
Visual repetitiveness – Sound dissonance	(Giannakis, 2001, 2006)
Texture granularity-Sound compactness -	(Giannakis, 2001, 2006)

Table 8. Achromatic and Chromatic mapping associations between audio and visual descriptors for the retrieval of audio.

Visual Features	**Audio Features**	**Visual Features**	**Audio Features**
Texture compactness	Spectral flatness	Color variance	Spectral flatness
Vertical position	Pitch	Color brightness	Spectral centroid
Texture variance	Periodicity	Brightness variance	Periodicity
Size	Loudness	Size	Loudness
Horizontal length	Duration	Horizontal length	Duration

4.6.4 Visual Feature Extraction

During playback windowed analysis is performed on the greyscale version of the sketch. As mentioned earlier, a window scans the sketch from left to right one pixel at every clock cycle, the rate of which is determined by the user. Only the areas of the canvas that are within the boundaries of the window area are subjected to statistical analysis. The window dimensions are determined by Window width by window height. The window width can be determined by the user, however the default size of the analysis window is 9 pixel wide by 240 pixel height.

Morpheme uses two computer vision algorithms to detect the centre of the painted area and the texture compactness (cv.jit.centroid, cv.jit.circularity), see[6]. Table 4 shows all the descriptors that are extracted from the canvas and explains how they are estimated. As we can see in Table 9 in the case of color brightness the white pixels are occluded so that the estimation of the mean is not affected by the brightness values of the background pixels.

Table 9. Shows how visual features are estimated

Visual Features	Method of analysis
Size/Thickness	Estimated by filtering the background and counting the number of the remaining pixels.
Vertical position	Estimated based on the centroid of the pixels which are ON in a binary image, see *Equation 1* & *Equation 2*.
Texture compactness	Estimated based on the ratio between the painted area and its' perimeter, see *Equation 3*.
Texture variance	Estimated based on the entropy of the histogram of intensities of a greyscale image, *Equation 4*.
Colour variance	Estimated based on the entropy of the histogram of HSL matrix (i.e. Hue, Saturation, and Lightness).
Brightness variance	Estimated based on the coefficient of variation of the histogram of the lightness matrix, see *Equation 5*.
Color Brightness	Mean lightness, filtering out background pixels (i.e. white pixels)

The perimeter of the painted area of the canvas which is used to estimate the circularity (i.e. compactness) of the sketch is estimated by counting the number of edge pixels in a binary image, that is, the number of ON pixels that have at least one OFF neighbour. The area is estimated by adding the intensities of all cells of a thresholded image (i.e. number of ON pixels) which is then divided by 255, in order to normalise the sum of all intensities to a number that has a maximum value of 255. Thresholding is method for partitioning the image into foreground

[6] http://jmpelletier.com/cvjit/

and background. Thresholding is performed to a greyscale version of the sketch (i.e. RGB to monochrome conversion resulting in a matrix with a single plane that contain the luminosity of the original image). In order to threshold an image, the difference between the pixel value and the average brightness of pixels within the distance set by the "radius" attribute is calculated. If the distance is greater than the "threshold" value, the pixel value is set to ON (i.e. luminosity value set to 255) otherwise the pixel is turned off (i.e. luminosity value set to 0), see *Equation 2*. *Figure 39* and *Figure 40*, show in detail the different steps followed for the analysis of the image and the extraction of visual features.

Equation 1. The following equation is used to estimate the centroid of the image.

$$Centroid = Cx = \frac{M10}{M00} \; and \; Cy \frac{M01}{M00}$$

Equation 2. Image thresholding to create a binary image.

$$\mathrm{dst(x,y)} = \begin{cases} \mathrm{maxVal} & \mathrm{if\, src(x,y) > threshold} \\ 0 & \mathrm{otherwise} \end{cases}$$

Equation 3. The following equation is used to estimate the compactness of the image.

$$Circularity = \frac{4 * \pi * Area}{Perimeter * perimeter}$$

Equation 4. The following equation is used to estimate the entropy of a histogram.

$$Entropy = -sum(p * log2(p))$$

Equation 5. The following equation is used to estimate the coefficient of variation of a histogram.

$$Coefficient\ of\ variation = \frac{Standard\ deviation}{Mean}$$

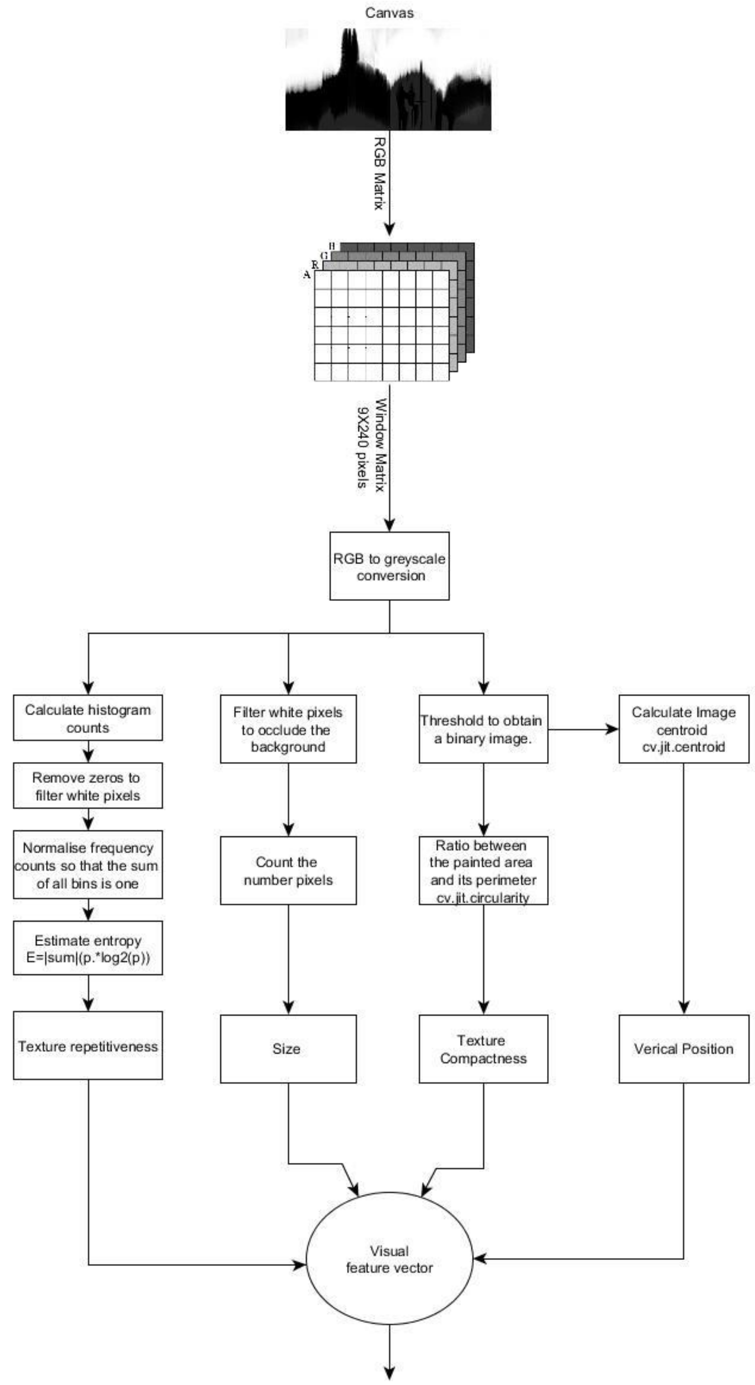

Canvas
RGB Matrix
B
G
R
A
Window Matrix
9X240 pixels
RGB to greyscale conversion
Calculate histogram counts
Remove zeros to filter white pixels
Normalise frequency counts so that the sum of all bins is one
Estimate entropy E=|sum|(p.*log2(p))
Texture repetitiveness
Filter white pixels to occlude the background
Count the number pixels
Size
Threshold to obtain a binary image.
Ratio between the painted area and its perimeter cv.jit.circularity
Texture Compactness
Calculate Image centroid cv.jit.centroid
Verical Position
Visual feature vector

Figure 39. Statistical analysis performed on the canvas matrix to extract visual features used as the target for the retrieval of audio-units for the achromatic mapping.

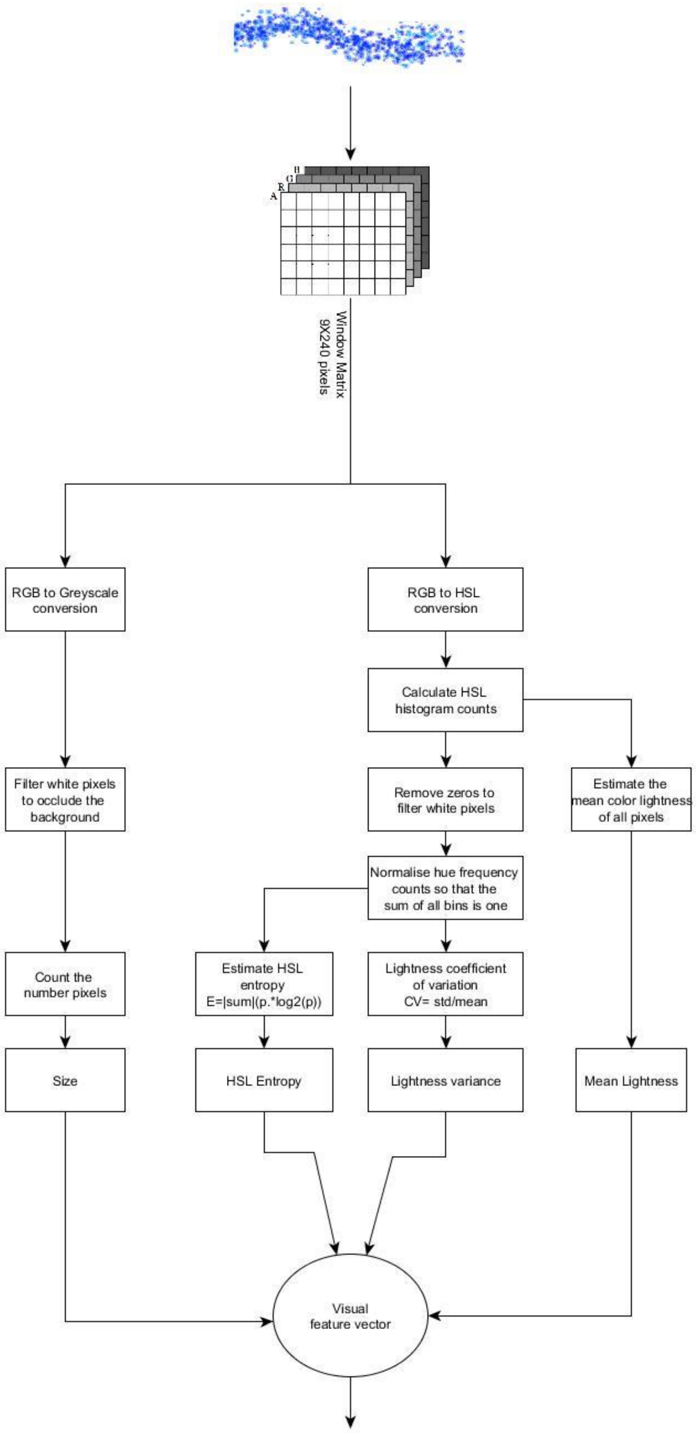

B
G
R
A
Window Matrix
9X240 pixels
RGB to Greyscale conversion
RGB to HSL conversion
Calculate HSL histogram counts
Filter white pixels to occlude the background
Remove zeros to filter white pixels
Estimate the mean color lightness of all pixels
Normalise hue frequency counts so that the sum of all bins is one
Count the number pixels
Estimate HSL entropy E=|sum|(p.*log2(p))
Lightness coefficient of variation CV= std/mean
Size
HSL Entropy
Lightness variance
Mean Lightness
Visual feature vector

Figure 40. Statistical analysis performed on the canvas matrix to extract visual features used as the target for the retrieval of audio-units for the chromatic mapping.

4.6.5 Mapping the Distances to the Synthesis Parameters

Table 5 shows how the distances between audio features and their respective visual features shown in tables 2 and 3 are associated with the synthesis parameters. The distances are estimated by subtracting the target and selected feature vectors, see Figure 41. Decisions regarding the mapping which are presented in table 5, were informed by the initial mapping for the retrieval of audio-units. For example, if the feature of size in the visual domain is mapped to loudness, then the distance between target thickness and selected loudness is mapped to control the amplitude parameter of the sound synthesis. However, other audio features such as spectral flatness, periodicity and spectral centroid require more careful consideration as they do not have a direct corresponding synthesis parameter. The decision on how to map the features which do not have a direct correspondence was made in an intuitive manner, by trying different combinations and assessing which correlations are plausible. However to answer these questions in objective terms, empirical work will have to be conducted, to test which correspondences are considered optimal.

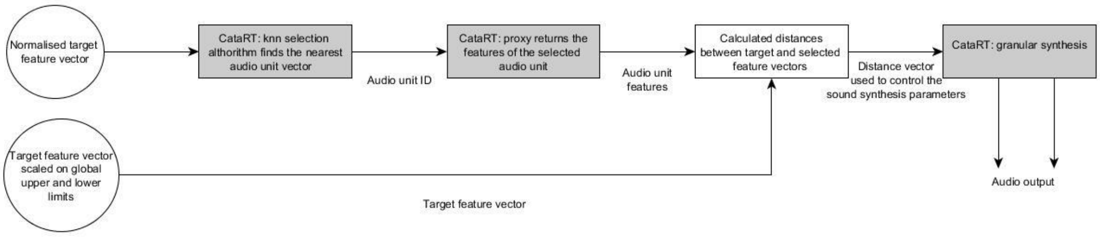

Normalised target feature vector
CataRT: knn selection althorithm finds the nearest audio unit vector
Audio unit ID
CataRT: proxy returns the features of the selected audio unit
Audio unit features
Calculated distances between target and selected feature vectors
Distance vector used to control the sound synthesis parameters
CataRT: granular synthesis
Audio output
Target feature vector scaled on global upper and lower limits
Target feature vector

Figure 41. Diagram explaining the way the distance between target and selected vectors are estimated in order to control the sound synthesis parameters.

Table 5. *Mapping the distances between audio and visual feature vectors to synthesis parameters.*

Audio features	Synthesis parameters
Spectral flatness	Transposition randomness
Periodicity	Grain size and amplitude randomness
Pitch, Spectral Centroid	Transposition
Loudness	Amplitude

4.6.6 Controlling the Weights for Feature Selection

For the selection of audio units from the corpus, k-Nearest Neighbour (k-NN) is used. k-NN works by estimating the shortest distance between the feature vector of the target (e.g. visual features extracted from the canvas) and the feature values of the units stored in the database (for more information on the algorithm, see Schwarz, (2004). Because the retrieval of audio units is based on multiple features, often the selected audio-units are not the best match for each individual feature value. Instead, it is the optimal match, taking into consideration all the distances between the target feature vectors and the values of the audio-units found in the database. To weights can be used on the selection algorithm in order to determine how dependent should the k-NN algorithm be on a particular descriptor when estimating the shortest distance between the target vector and the feature vectors of the audio units found in the corpus.

A simple method was devised to automatically adjust the weights of each audio feature dimension by counting the number of frequency bins that result from the histogram analysis of the feature vectors of the audio-units which are stored in the concatenated corpus. The number of bin of a histogram describes how wide is the distribution of the audio-units which indicates how different/similar the audio-units are in each of the feature dimensions. The weighting algorithm is based on the notion that when a feature value becomes very common between a set of objects, it becomes less salient for establishing links between two or more objects. Consequently it could be argued that descriptor dimensions that have high dispersion should be given more weight (i.e. be enforced) than feature dimensions that have low dispersion, as the latter are not distinct enough for assessing feature based similarity. As an example, the histogram of the spectral centroid presented in Figure 42 has relatively low dispersion while spectral flatness has higher dispersion; in this case the spectral centroid should be given more weight than spectral flatness. The aim is to weaken feature dimensions when the audio-units stored in the corpus are very similar. As the more similar the audio-units are, in a given feature dimension (e.g. loudness, pitch, periodicity), the less concerned we should be about which audio-unit will be selected by the algorithm. Conversely, when the distribution of the audio-unit in a given feature dimensions is wide that feature dimension should be given more weight as it indicates that the audio material is diverse.

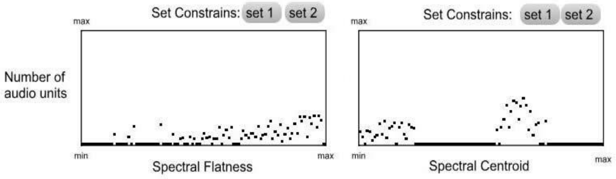


Figure 42. The histograms display the distribution of the audio units in a corpus for two features.

This approach helps improve the efficacy of the distance based algorithm (i.e. k-NN) in assessing feature-based similarity in a multidimensional context, by optimizing the selection algorithm in a corpus dependent manner. To achieve automatic weighting, the coefficient of variation for each feature dimension is estimated based on its histogram. The coefficient of variation is the ratio of standard deviation to the mean. It provides an estimate of the variability of the audio-units in the corpus which help compare the audio features that have different mean value. The percentage of dispersion given by the coefficient of variation is used to determine how dependent should the selection algorithm be when assessing the distances.

4.6.7 Constraining the Selection Algorithm

Constraining the selection algorithm to the areas of the corpus where audio units have formed clusters can improve the navigation of the feature space. A common phenomenon when an audio corpus consists of a relatively small number of audio units is that the distribution of the audio descriptions in the feature space is concentrated, forming dense clusters in some areas while the rest of the feature space is relatively empty see Figure 43. One problem that can be observed is that the target features values requested might be well outside the main clusters of the feature space and the target might not match any of the audio units in the corpus. As mentioned earlier in such case, the k-NN algorithm will select the nearest unit that can be found in the feature space. On the one hand this is very useful as a unit can be selected even if the target does not exactly correspond to any of the descriptions of the audio units found in the corpus. On the other hand if the target feature vectors in a series of queries are well outside the main clusters, the selection algorithm will stay in the periphery of the clusters and will not access the clusters (i.e. Figure 43 shows what is referred to as periphery). This results in the following problems: (i) although the corpus might consist of many audio-units, it might be difficult to access the clusters, and (ii) two very different target queries (i.e. two brush strokes in the context of Morpheme), which both request feature vectors that are well outside the main clusters of the corpus, might retrieve the very same audio-unit. Figure 43 demonstrates the problem; Target 1 and Target 2 are relatively distant in the feature space, however the same audio unit is selected as it is the nearest. This in a sense makes difficult to explore the feature space and it can create ambiguities for the user regarding the association between target and selected feature.

A method has been devised to address these issues and improve the navigation of the feature space. To achieve this, a histogram analysis is performed across each of the features dimensions which are used for retrieval, see Figure 42.The histogram shows the distribution of all the audio units of a particular feature. Currently Morpheme allow to set up two constraints in order to avoid empty or unwanted areas of the corpus (e.g. empty, silent or undesired audio units). This is accomplished by setting minimum and maximum bounds using a pointing device directly on the histogram's graphical representation, see Figure 44. The minimum and maximum bounds are then

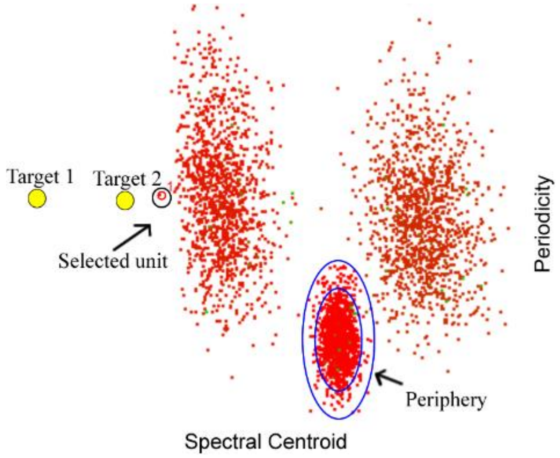
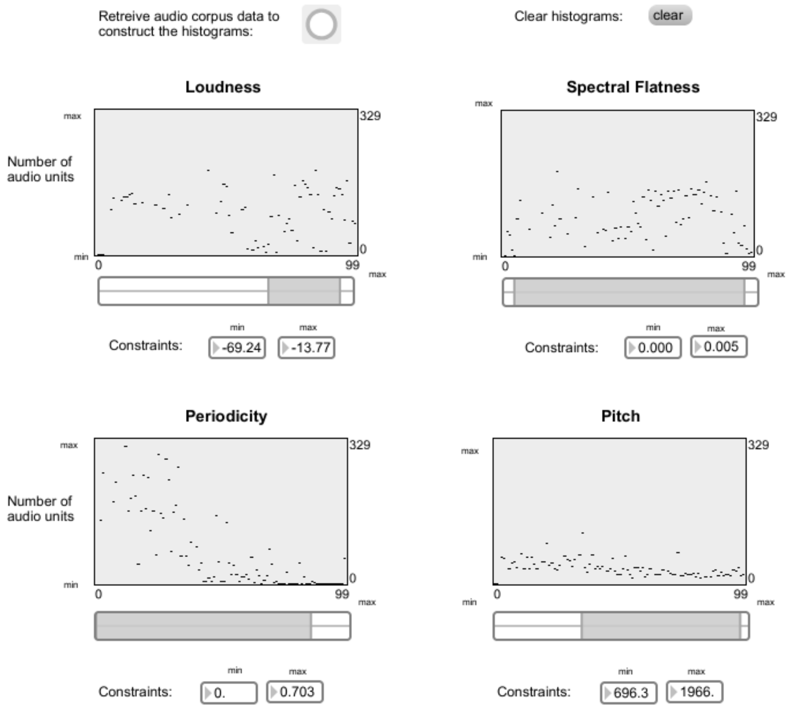


Figure 43. A two dimensional plots shows the distribution of the audio units in a corpus.

used to scale the target features and map them to the areas of the corpus defined by the constraints which were set by the user. Any target query that requests feature vectors from the Area out of bound 1 will be scaled to a corresponding value from Constrain area to bounds 1, while queries that request units from Area out of bound 2 are mapped to Constrain area to bounds 2.



Loudness Constraints: min -69.24 max -13.77

Spectral Flatness Constraints: min 0.000 max 0.005

Periodicity Constraints: min 0. max 0.703

Pitch Constraints: min 696.3 max 1966.

Figure 44. A two dimensional plots display the distribution of the audio units and the user interface that was developed for setting constraints.

4.7 Conclusions

In this chapter, the initial design and the redesign of Morpheme, a user interface for visual control of corpus-based concatenative synthesis was presented. Morpheme incorporates sketching as a model for interaction and image analysis techniques are deployed to extract visual features from the sketch which are used to query audio-units from the database. The mapping between audio and visual descriptors which are used for achieving visually driven audio retrieval are based on previous empirical finding from studies that investigated audio-visual feature correspondence, which were in Chapter 3. Various implementation details have been discussed including, the user interface, the audio-visual mapping, the visual feature extraction methods and the exploration of the audio corpus. The following, three chapters are concerned with the evaluation of the system. More specifically Chapters 5 and 6 present two experiments which were conducted to assess the effectiveness of the mapping. While Chapter 7, present a user study that was conducted in order to evaluate Morpheme's user interface.

5 Experiment 1 - An Investigation of the Effects of the Harmonicity of the Audio Corpus on the Perceived Correspondence of Audio-Visual Associations

In order to validate the Audio/Visual (A/V) mapping strategies devised during the design stage of Morpheme, three empirical studies were conducted, which are presented in this and the two subsequent chapters. This chapter investigates the perceived strengths of chosen A/V mappings when different audio corpora are used by the Morpheme concatenative synthesiser. The A/V associations tested in this study are according to previous studies some of the most highly rated feature correlates, (Eitan & Timmers, 2010; Evans & Treisman, 2010; Kostas Giannakis & Smith, 1993; Kohn & Eitan, 2009; Kussner & Leech-Wilkinson, 2013; Küssner, 2014; Lipscomb & Kim, 2004; Marks, 1989; Rusconi *et al.*, 2006; Walker, 1987). Previous research findings suggest that the degree of perceived correspondence between complementary polar features can be affected in dynamic multidimensional contexts. As discussed in Chapter 3, complementary polar features refer to pairs of complementary cues[7] that produce high levels of perceived congruency. For instance, an increase in tempo rate is often correlated to an increase in visual motion speed (a complimentary polar features), however when increase in tempo is accompanied by a decreasing loudness the correspondence between tempo and speed is weakened, (Eitan, 2013). Eitan suggests that the relationships between complimentary polar features can be affected by at least three factors the type of the stimuli (i.e. static-dynamic), the interactions between simultaneous variation of multiple feature dimensions. Moreover, research findings show that complimentary polar features can sometimes match equally well more than one features of different sensory modality (Eitan, 2013). For example, as discussed in Chapter 3, it has been observed that visual size corresponds equally well to both pitch and loudness, while pitch corresponds well to both vertical position and lightness.

Although previous studies have explored perceived correspondence between visual and auditory parameters (as it has been discussed in Chapter3) and the literature shows that interactions between different dimensions of auditory and visual parameters in multidimensional

[7] Complementary cues, refer to the features of the physical world which can be experienced only by one sensory modality (e.g. color and timbre). The opposite of complimentary cues is redundant cues, which refer to sensory information perceived through different sensory modalities which describe the same feature of the physical world (e.g. location or the shape of an object in close proximity can be experienced through vision and touch), see Parise & Spence, (2013).

context can occur (Eitan, 2013). No previous study has examined how the degree of perceived correspondence is affected by the harmonicity of the audio that is used to render the A/V feature associations. Further, the audio stimuli that were used for testing A/V correspondence by previous studies comprised simple synthetic stimuli such as synthesized sine tones, while sounds in nature consist of a multiple feature dimensions. Hence, it could be argued that there are evident ecological validity questions regarding the findings of most studies of cross-modal correspondence and that follow-up experiments are required in order to assess how this findings generalise/apply to different contexts. For example, how is the degree of perceived correspondence affected by the presence of the less salient features of the stimuli? In the context of this research, it is extremely important to answer these questions as corpus-based concatenative sound synthesis uses pre-recorded audio material in order to synthesise sound that can have complex timbre characteristics. In order to examine how the characteristics of timbre of the sounds used to render a complementary polar features affect the degree of perceived similarity, a pairwise similarity test was designed were subjects were presented with a series of audio-visual stimuli and rated the perceived similarly for each pair. This chapter describes in details the experimental design and the results from this study.

5.1 Methodology

5.1.1 Hypotheses

Three primary research questions were investigated.

Q1: Which audio-visual features are perceived as the best correlates?

H1: Size/Loudness and Vertical position/Pitch, Visual brightness and Spectral brightness are dominant A/V associations

Q2: Does the source audio that the corpus consists of have a significant effect on the perceived similarity of the audio-visual associations?

H2: Research findings suggest that the perceived correspondence of the A/V associations is susceptible to the presence of other features.

Q3 Does mapping (i.e. chromatic/achromatic) as a whole have a significant effect on the perceived similarity of the audio-visual associations?

H3: No, as both mapping have equal number A/V dimensions that are strongly correlated.

Q4: Will the degree of perceived correspondence of the A/V association differ between the two groups (i.e. sound practitioners and non-sound practitioner)?

H4: There will be no difference between the two groups particularly for the dominant A/V associations Size/Loudness and Vertical position/Pitch these associations have not formed due to musical/sound training.

5.1.2 Apparatus

The experiments took place in the Auralization room a sound proofed recording studio at Merchiston Campus of Edinburgh Napier University. Participants used Beyer Dynamics DT 770 Pro monitoring headphones to listen to the audio stimuli. A 17.3 inches laptop screen was used for viewing the visual stimuli, see Figure 45 for an illustration of the experimental setup. SurveyGismo was used for viewing the stimuli and for recording the participants' responses. Ethical approval of the experiment was obtained by the research ethics committee of Edinburgh Napier University, the relevant documents can be found in the Appendix C and D.

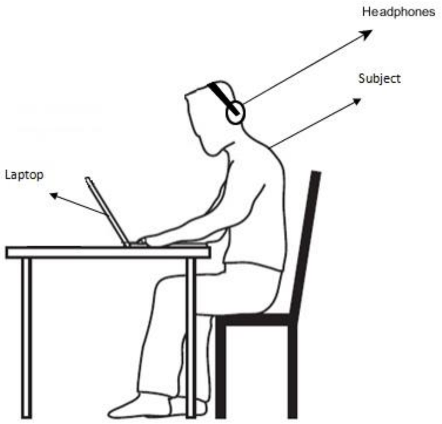


Figure 45. Experimental setup.

5.1.3 Procedures

Subject responses were collected independently. Each participant completed the following tasks. Participants were given a brief description of the task followed by a short demonstration of the apparatus used for viewing the stimuli and for capturing and storing their responses. Before beginning the main experiment, each participant had to fill a form with their demographic information. Prior to the main experiment participants did a warm-up trial consisting of three audio-visual examples to confirm their understanding of the task, become familiar with the apparatus and the response procedures. After the training session which lasted no more than a minute, participants continued with the actual experiments. The transition between the warm-up task and the main experiment was seamless without interruptions. Participants were asked to watch and listen to a series of brief simultaneous auditory and visual stimuli. After each individual audio-visual stimulus was presented, subjects expressed if they felt that the auditory and visual stimuli they experienced were similar or not. Subjects' responses were expressed by selecting between two on-screen button (i.e. labelled *Dissimilar – Similar)* using a computer mouse. After the first binary response, subjects also rated the degree of similarity between the auditory and visual stimuli. Subjects could indicate their responses by controlling an on-screen slider (i.e. numeric scale from 0 to 100). In total participants were exposed to 18 audio-visual stimuli, excluding the three stimuli used for the trial. Subjects could playback the stimuli as many times as they wished. The task takes approximately ten minutes to complete. The order in which the stimuli were presented to each participant was randomized to avoid a repetition effect.

The same experiment was also performed in an uncontrolled environment as an online study in order to obtain more responses. The survey was posted to five audio and music related mailing lists. I will be using the term supervised study to refer to the experiment that took place in a controlled environment and the term unsupervised study to refer to the online study.

5.1.4 Subjects

In the supervised study a total of 44 participants volunteered to take part. Sixteen were sound practitioners, and twenty-eight non-sound practitioners. All of the expert participants except two played a musical instrument and the self-reported level of expertise was 4 basic, 6 intermediate and 4 advanced. Except one, all of the other participants reported using analogue and digital equipment for sound synthesis, signal processing, sequencing and audio programming tools. Seven subjects reported an intermediate level of expertise regarding the use of digital and analogue audio equipment and eight reported advanced skills.

In the unsupervised study participants were similarly divided into two groups. There were a total of 66 participants. The expert group consisted of 39 participants and the non-expert group of 27. All of the expert participants except one played a musical instrument and the self-reported level of expertise was 11 basic, 14 intermediate and 14 advanced. Twenty eight of the participants had received formal music theory training for at least one year. All of the participants reported using analogue and digital equipment for sound synthesis, signal processing, sequencing and audio programming tools. Nineteen subjects reported intermediate level of expertise regarding the use of digital and analogue audio equipment and twenty reported advanced skills.

Participants filled a form with demographic information before the experiment. The main aim of this form was to assess the level of expertise of the participants in sound and music related areas and to detect participants that had severe hearing or visual impairments. The data from participants who reported having severe hearing or visual impairment were discarded. In order to distinguish between the expert and non-expert groups two criteria were used, participants needed to either play a musical instrument or use sound synthesis and processing equipment and rate their abilities regarding this two factors as intermediate or advanced. Sound practitioners and musicians were put into the same group because this study considers that what is important is the rich conceptual knowledge and listening skills that both sound practitioners and musicians have developed for identification, categorisation and description of sound as oppose to the instrument performance skills and/or their understanding of music theory. Of course musical instrument skills can lead to the development of different sensorimotor mental models of cross-modal features association. However this is also the case between musicians that play different type of instruments such as a woodwind instrument player and a drummer. In the context of this research what is important is not whether subjects play a musical instrument and have music theory education but instead whether or not subjects have rich conceptual knowledge and listening skills. There are procedures available, but not ideal for this type of study, see (Edwards, Challis, Hankinson, & Pirie, 2000; Hakinson, Challis, & Edwards, 1999; Wallentin, Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010).

5.2 Audio Stimuli

Three audio corpora were prepared for the study. Segmentation of the audio-units which each corpus consists of was set to 240 milliseconds. The decision to segment the sounds used in the corpus to audio-units of 240 milliseconds duration was primarily driven by the need to ensure that audio -units will overlap to create a continuous sequence even when the audio-units are transposed to a much higher pitch; which results in a reduction of the duration of the audio-units. For synthesis of the audio stimuli, audio-units are requested from the database every 40 milliseconds. The width of the canvas from where the target features are extracted is 100 pixels, this results in audio stimuli with a duration of 4 seconds. Each audio corpus was designed using sound recordings that vary in terms of two characteristics: harmonic content and continuity. For instance, the first corpus consists of audio-units that are very harmonic and continuous. The second corpus consists of audio-units that are moderately harmonic and continuous. While the third corpus consists of audio-units that are relatively dissonant, discontinuous and erratic such as impact/percussive sounds. Below follows a more detailed description of the source sounds that were used to generate the audio stimuli for this experiment.

5.2.1 String Corpus

The first corpus consists of audio units that have resulted from the segmentation of a 14 seconds audio recording of a bowed violin. The violin audio recording used in this corpus is very harmonic, periodic, with relatively low spectral flatness. Other characteristics of the source audio in this corpus include that the sound is continuous, invariant and sustained.

5.2.2 Wind Corpus

The second corpus consists of audio units that have resulted from the segmentation of a 60 seconds audio recording of wind. The source audio in this corpus has relatively high spectral flatness and low periodicity. Similar to the first audio corpus the second corpus consists of audio material that is continuous and sustained. However unlike the violin recording used in the first corpus the wind sound is less periodic and it has a flatter spectrum. For instance, string sound is closer to sine tone, while wind sound is closer to white noise.

5.2.3 Impacts Corpus

The third corpus consist of audio units that have resulted from the segmentation of 93 audio recordings of impact sound events such as smashing materials, shattering glass etc. The corpus has been prepared using source audio that have low spectral flatness, low periodicity and the sounds tend to be abrupt, discontinuous and dissonant.

5.3 Visual Stimuli

Six visual stimuli were designed for this experiment. The main criterion applied to design visual stimuli was that this study aimed at testing a single audio-visual feature association per audio-visual stimulus. For instance, in order to test the relationship between size and loudness, the visual stimuli must entail variations only in terms of size, while ensuring that all the other visual feature values remain unchanged, see next page Figure 47. However complete isolation of feature dimension was not always possible. For example in stimuli 3 and 6 shown in Figure 47, although we intend to manipulate only the granularity of the texture, small amount of difference in size will be detected by the respective visual extraction algorithm. Furthermore, after conducting a pilot study (for detail see (Tsiros, (2014)) that used visual stimuli that comprised a symmetric variation of each visual parameter (see Figure 48) it was realised that the symmetry of variation might result in a bias towards judging the A/V association as similar not because of the correspondence between the intentionally varied parameters, but due to the symmetry of the stimuli. So, we changed it to unidirectional one as show in Figure 47.

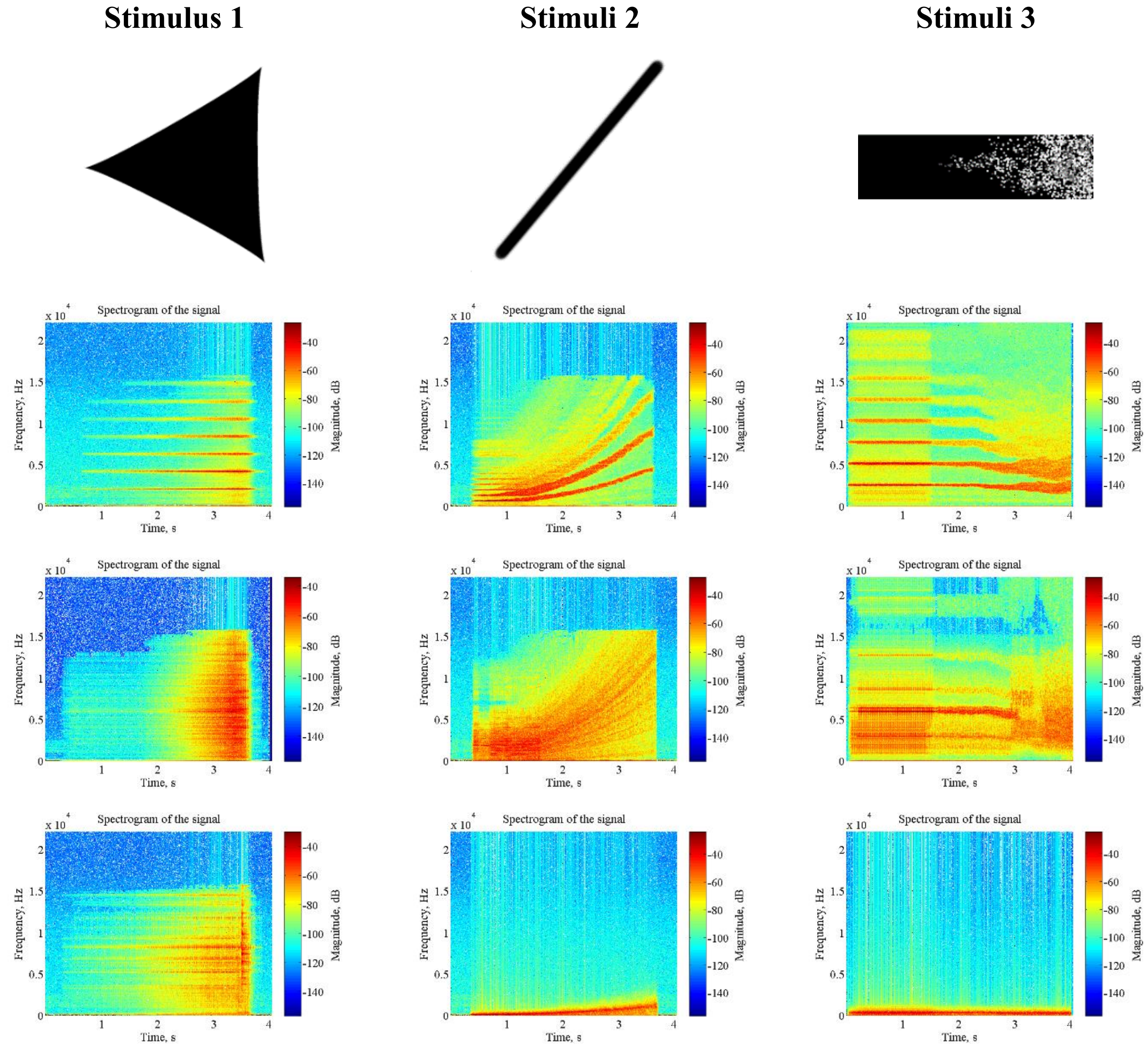


Figure 46. All the audio-visual stimuli used in this study. Each visual stimuli was used to generate three sound one for each of the audio corpus, see spectrograms in each column below the visual stimuli. The first row of spectrograms shows all the audio stimuli that was synthesised using each of the corpus. First row: string corpus. Second row: impacts corpus. Third row: wind corpus.

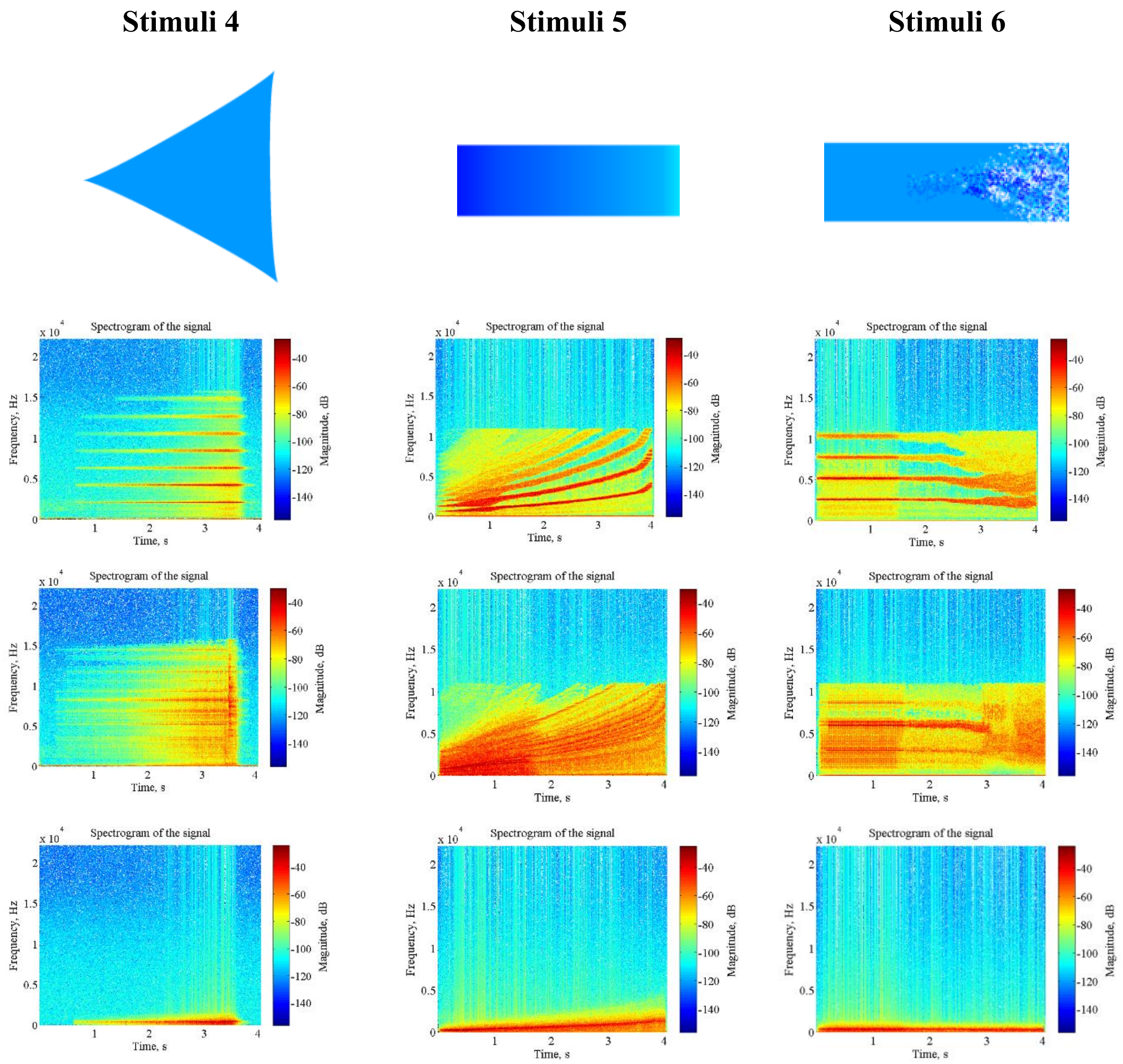


Figure 47. All the audio-visual stimuli used in this study. Each visual stimuli was used to generate three sound one for each of the audio corpus, see spectrograms in each column below the visual stimuli. The first row of spectrograms shows all the audio stimuli that was synthesised using each of the corpus. First row: string corpus. Second row: impacts corpus. Third row: wind corpus.

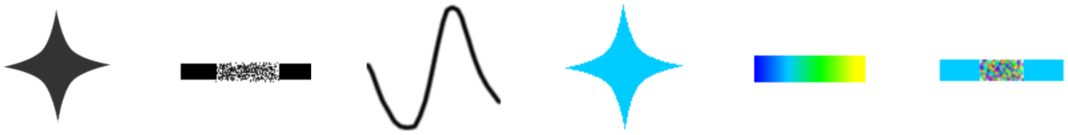


Figure 48. Stimuli used in the pilot study that varied each visual parameter in a symmetric manner.

Please note that in the present study, in each mapping two visual feature and their corresponding audio features were tested using a single stimuli (i.e. stimuli 3 and 6 shown in Figure 47). So texture compactness and texture repetitiveness are presented as texture granularity. Further color variance and brightness variance are presented as color variance. Similarly the audio features periodicity and spectral flatness are presented as Sound dissonance. Although in the context of the system these are separate feature associations (i.e. *Spectral flatness-Texture compactness, Periodicity-Texture repetitiveness, Spectral flatness-Color Variance, Periodicity-Brightness variance*), these associations were tested using a single audio-visual stimuli. These feature pairs were considered as a single feature dimension because they are tightly linked to the complexity of the texture in both modalities, describing the transitions between simple-complex, order-disorder (in the visual domain) and tone-noise, harmonic dissonant (in the auditory domain). Furthermore these feature pairs are somewhat dependent. For instance a periodic sound can have more or less flat spectrum, however if the spectrum of a sound is very flat the sound is likely to have low periodicity and vice versa. The stimuli, the data gather from this study and the statistical analysis tables are available in the appendix C, please see Chapter 9.

5.4 Experiment-1 Data Analysis

5.4.1 Question 1: results

This section presents the results obtained from the first question of the experiment, were expert and non-expert subjects in the supervised and unsupervised condition (i.e. in controlled environment and online study) made a series of pairwise similarity judgments between pairs of audio and visual stimuli. A Pearson chi-square test of independence was performed with the similarity judgments as within subjects factor and the mapping, the corpora and the A/V associations as between subject factors. The data analysis of the non-expert group in the supervised condition shows that there was a statistically significant relationship between perceived similarity and A/V associations: X2(5, N = 504) = 46.19, p < .001. While no correlation was identified between perceived similarity and the audio corpora $X^2(2, N = 504) = 1.42, p = .49$, or the mapping $X^2(1, N = 504) = 1.34, p = .246$.

For the expert group in the supervised condition, data analysis shows that there was a statistically significant relationship between perceived similarity and A/V associations X^2(*5, N = 288)* = 35.02, *p* = .0. While no association were revealed between perceived similarity and the audio corpora X^2*(2, N = 288)* = 1.76, *p* =.41; or the mapping X^2*(1, N = 288)* = .423, *p* =.51.

For the non-expert group in the unsupervised condition, data analysis shows that there was a statistically significant relationship between perceived similarity and A/V associations X^2*(5, N = 486)* = 50.79, *p* = .0. While no association were revealed between perceived similarity and the audio corpora X^2*(2, N = 486)* = 1.24, *p* = .53; or the mapping X^2*(1, N = 486)* = .232, *p* =.63.

For the expert group in the unsupervised condition, data analysis shows that there was a statistically significant relationship between perceived similarity and A/V associations X^2(5, N = 702) = 71.28, p =.0; and the mapping X^2(1, N = 702) = 4.57, p =.033. While no association were revealed between perceived similarity and the audio corpora X^2(2, N = 702) = 4.22, p = .12.

Post-hoc testing of each group/condition chi-square results was performed for each level of the factor A/V associations shown in Figure 49 in order to determine which factor levels contribute to the statistical significance that was observed. The results of the post-hoc test are presented in Table 10. Post-hoc testing was achieved by estimating p values from the chi-square residuals for each level of the factor and comparing these to an adjusted benferonni corrected p-value. The benferonni corrected p-value is estimated by dividing .05 by the number of analyses which corresponds to the number of levels of the factors involved. For example there are six A/V associations and two potential responses (i.e. Dissimilar/ Similar) consequently the benferroni adjusted p-value for the present analyses is estimated by dividing .05 by 12 which is equal 0.0041. For more information about the approach followed to perform post-hoc testing refer to (Beasley & Schumacker, 1995; Garcia-perez & Nunez-anton, 2003).

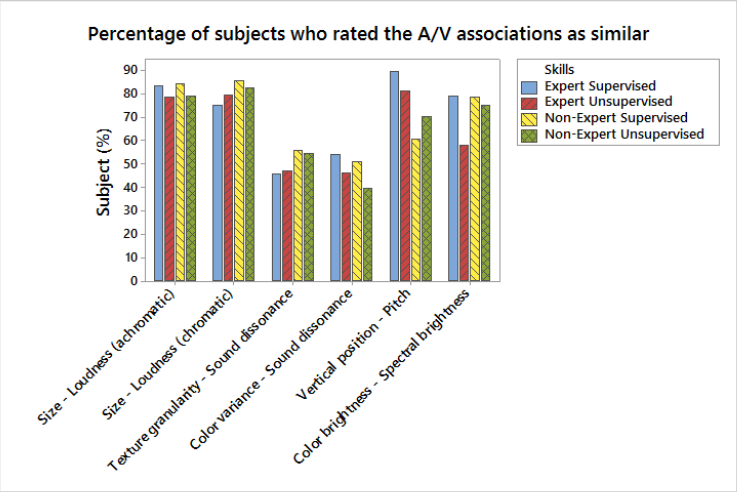


Figure 49. Percent of participants who rated the A/V Associations as similar.

Table 10. Post-hoc testing of each group/condition chi-square results for each A/V associations. An asterisk (*) indicates statistical significance at adjusted alpha level of .0041.

	Experts (s)		**Experts (u)**		**Non-experts(s)**		**Non-experts(u)**	
A/V Associations	***X^2***	***p***	***X^2***	***p***	***X^2***	***p***	***X^2***	***p***
Size - Loudness (a)	4.1	.04	11.3*	<.0041	10.8*	<.0041	6.4	0.01
Size - Loudness (c)	0.4	.5	12.7*	<.0041	12.5*	<.0041	11*	<.0041
Texture granularity - Sound dissonance	18*	<.0041	20.2*	<.0041	8.6*	<.0041	6.9	0.008
Color variance - Sound dissonance	8.1*	<.0041	22.1*	<.0041	15.8*	<.0041	32.8*	<.0041
Vertical position - Pitch	9.5*	.002	16*	<.0041	3.6	0.05	0.5	0.4
Color brightness - Spectral brightness	1.79	.18	3.	.08	3.9	0.04	3.1	0.07
N	48		117		84		81	

s=supervised / u=unsupervised

Benferroni adjusted p value 0.0041

This section presents the results obtained from the comparison between the expert and the non-expert groups in the controlled and uncontrolled conditions. Pearson's chi-square test of independence was performed on four between subject variable to test the null hypothesis of no association between participants' similarity judgments and the skills of the subject and between the four groups. Data analysis shows that there were no significant relationship between perceived similarity and subject training X^2*(1, N = 792)* = .263, *p* = .608 in the controlled condition or the uncontrolled condition X^2*(1, N = 1188)* = .401, *p* = .526, nor between the two groups in the two conditions X^2*(3, N = 1980)* = 4.57, *p* = .205.

Since there were no significant differences between the two subject groups in the two conditions we merged the results from all groups/conditions in order to perform a chi-square test on the merged dataset for the A/V association factor that was the only factor in which we found statistically significant differences in the participant responses. The data analysis of the merged dataset shows that there was a statistically significant relationship between perceived similarity and A/V associations: X2(5, N = 1980) = 169, p < .001. Post-hoc testing was performed to identify which levels of the factor A/V associations were contributing to this statistical difference. Table 11 shows that the A/V associations: *Size-Loudness* both chromatic/achromatic, *Texture granularity-Sound dissonance* and *Color variance Sound dissonance* are the levels that appear to have significant statistical differences. As it can be seen in Figure 49, the A/V association *Size-*

Loudness was much higher rated for all groups and conditions that the other two associations Texture granularity-Sound dissonance and Color variance Sound dissonance.

Table 11. Post-hoc testing of the merged (group/condition) chi-square results. An asterisk (*) indicates statistical significance at adjusted alpha level of .0041.

A/V associations	X^2	**p**
Size – Loudness(achrom)	32.35*	<.004
Size – Loudness(chrom)	33.83*	<.004
Texture granularity - Dissonance	49.86*	<.004
Color variance - Dissonance	76.30*	<.004
Vertical position - Pitch	8.90	.113
Color brightness – Spectral Centroid	1.71	.887
N	330	
Total	X^2*(5, N =1980) =169, p <.001*	

benferroni adjusted p value 0.0041

5.4.2 *Question 2: results*

Here we present the results obtained from the second question of the experiment where subjects rated the degree of similarity between a series of audio and visual stimuli. A three way analysis of variance was performed using a general linear model with the similarity ratings as within subjects factor and the mapping, the corpora and the A/V associations as between subject factors. Further the model comprised two crossed factors: the mapping and the corpus and one nested factor A/V associations (mapping). Each mapping consists of three A/V associations and has two levels (i.e. chromatic and achromatic, (2x3)). The general linear model was computed to test (i) the null hypothesis of no statistically significant difference between the response variable 'similarity ratings' and the factors A/V association, mapping and audio corpora, and (ii) to test for interactions between the factors A/V associations-corpus, mapping-corpus. The results of the analyses are shown in Table 12 and Table 13 respectively (i.e. main effects and interactions). Data analysis revealed statistically significant difference in the participant responses as a result of the different audio-visual associations for all the subject groups in the two conditions (i.e. non-expert/expert, unsupervised/supervised) but not as a result of the mapping or due to the harmonicity of the audio corpus, see Table 12.Significant interactions between the factors A/V associations and audio corpus were revealed for all the subject groups in the two conditions (i.e. non-expert/expert, unsupervised/supervised), see Table 13. While no interactions between the factors mapping and audio corpus with exception for the expert group in the supervised condition were a significant interaction was revealed, also see Figure 50.

Table 12 General ANOVA model computed to investigate the effect of the A/V associations, corpus and mapping on the perceived similarity reported by the subjects. An asterisk (*) indicates statistical significance.

	A/V associations		**Corpus**		**Mapping**		
Group	***F***	***P***	***F***	***P***	***F***	***P***	***Total***
Non-experts (s)	12.96*	<.001	2.04	.13	0.73	.39	503
Non-experts (u)	13.95*	<.001	.76	.46	2.69	.1	287
Experts(s)	15.27*	<.001	1.91	1.9	0.26	.6	465
Experts(u)	27.28*	<.001	1.84	.16	8.26*	.004	701
df	4		2		1		

s=supervised / u=unsupervised

Table 13. ANOVA model created to investigate the interaction between A/V association- Corpus and Mapping-Corpus. An asterisk (*) indicates statistical significance.

	A/V associations **Corpus**		**Mapping** **Corpus**		
Group	***F***	***P***	***F***	***P***	***Total***
Non-experts (s)	3.94*	<.001	0.75	.4	503
Non-experts (u)	3.39*	<.005	0.25	.7	287
Experts(s)	5.18*	<.001	4.31*	<.05	465
Experts(u)	5.84*	<.001	.09	.9	701
df	8		2		

s=supervised / u=unsupervised

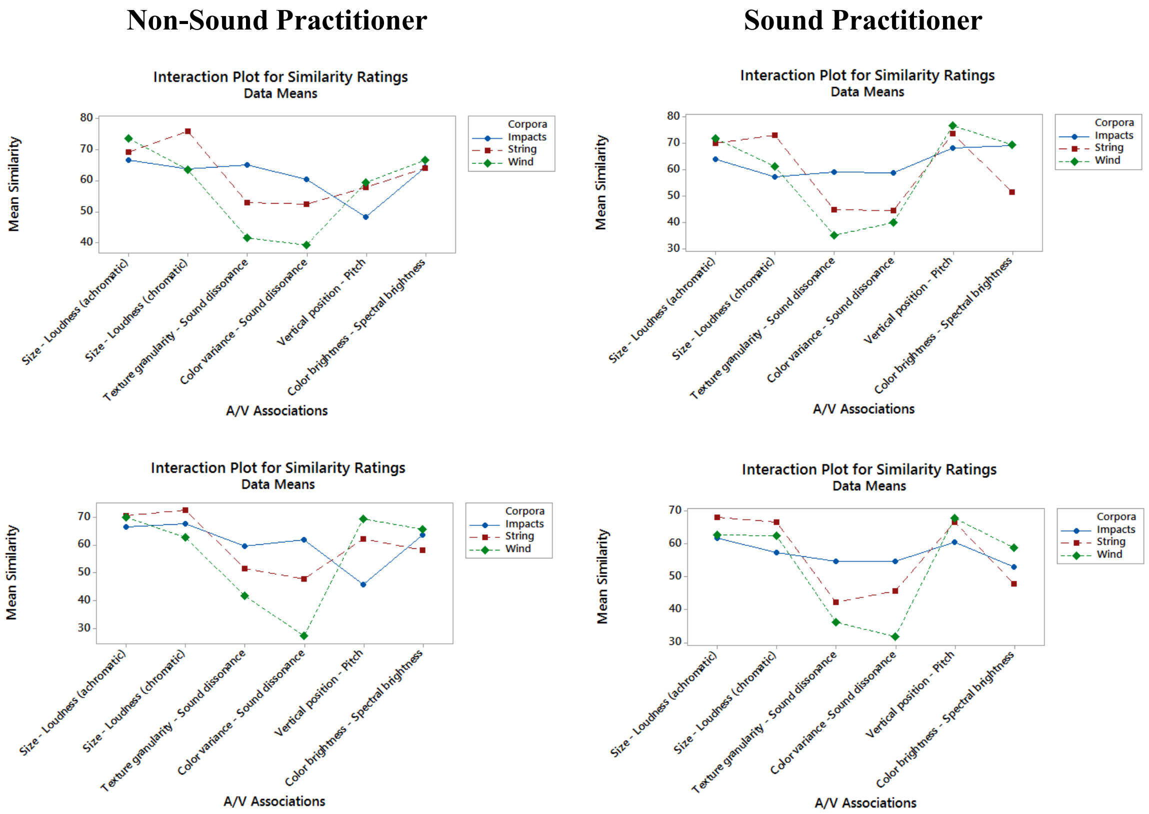


Figure 50. Interactions between subject responses for each A/V associations and the audio corpora.

This section presents the results obtained from the comparison between the expert and the non-expert groups for the second question of the experiment where subjects rated the degree of similarity between a series of audio and visual stimuli. A four way analysis of variance was performed using a general linear model with the similarity ratings as within subjects factor and the subjects' skills, the mapping, the corpora and the A/V associations as between subject factors. A nested model was prepared for analysis of variance between the two groups, using the following hierarchical order: skills(mapping(A/V associations(corpus))). The general linear model was computed to test the null hypothesis of no statistically significant difference between the similarity ratings of the expert and non-expert groups. Data analysis between the two groups showed statistically significant differences between the two groups due to the A/V associations, while no due to the corpora, or the mappings, see

Table 14. Post-hoc testing using benferroni adjustment with 95% confidence interval show that the difference in the perceived similarity between the expert and non-experts of the supervised group were significant only for the A/V association Vertical position- Pitch. In the unsupervised study post-hoc testing show none of the associations' means were significantly different. Figure 51, Figure 52 and Figure 53 summarise the results of all the groups and

conditions.

Table 14. General ANOVA model computed to investigate the effect of the participant skills on the perceived similarity of the A/V associations, corpus and mapping reported by the subjects. An asterisk (*) indicates statistical significance.

		A/V associations		**Corpus**		**Mapping**		
Skills		***F***	***P***	***F***	***P***	***F***	***P***	***Total***
Non-experts(s)	Experts(s)	12.62*	<.001	1.43	2.53	0.73	.39	791
Non-experts(u)	Experts(u)	2.62*	<.05	.10	.9	1.7	.19	1187
df			5		2		1	

s=supervised / u=unsupervised

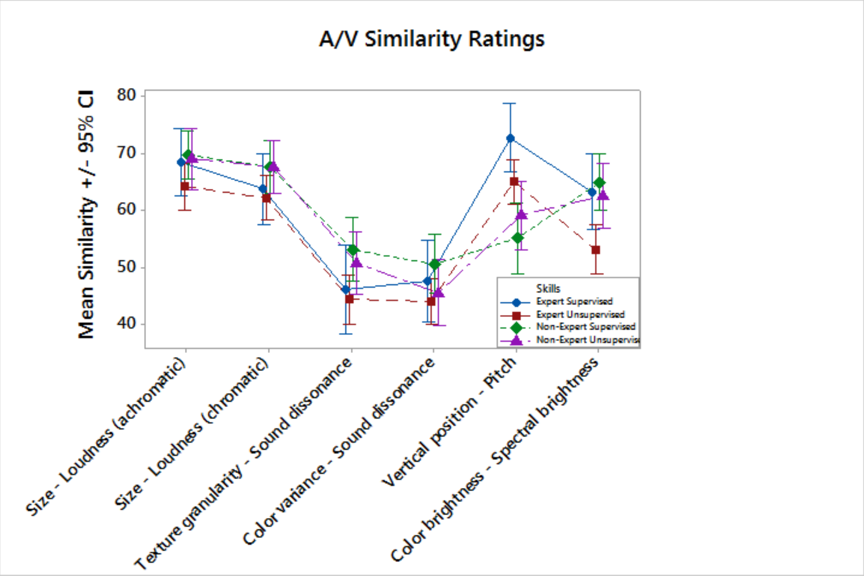


Figure 51. Data means and confidence intervals of participants' similarity ratings for A/V associations by all the groups. If intervals do not overlap the corresponding means are statistically significant.

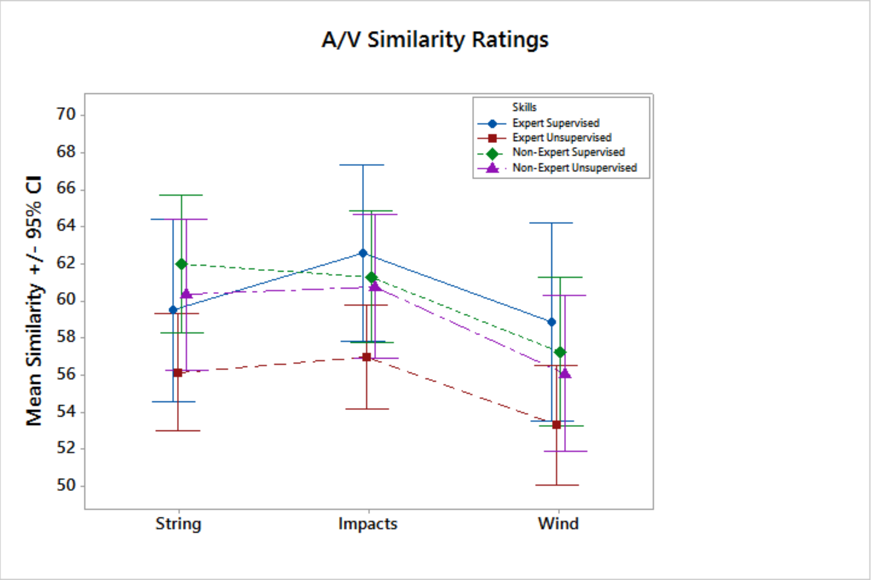


Figure 52. Data means and confidence intervals of participants' similarity ratings for the corpus factor by all the groups. If intervals do not overlap the corresponding means are statistically significant.

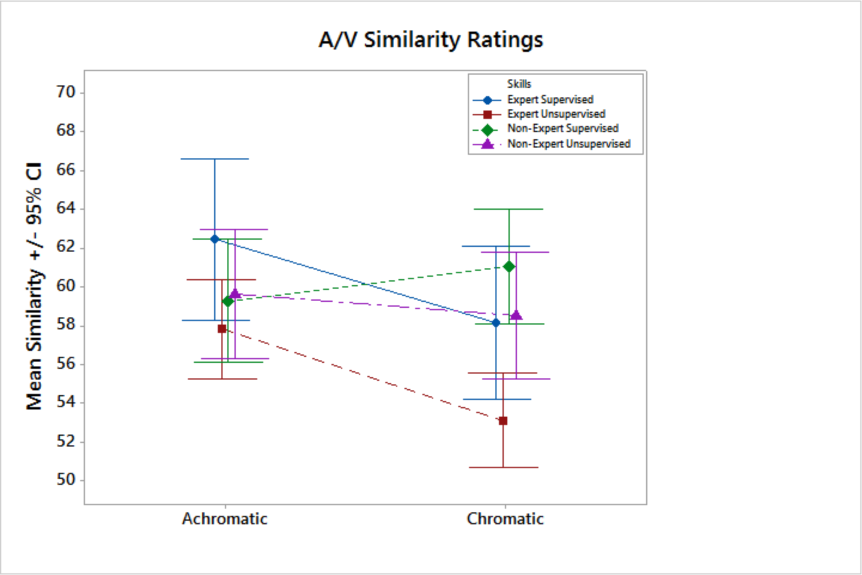


Figure 53. Data means and confidence intervals of participants' similarity ratings for the mapping factor by all the groups. If intervals do not overlap the corresponding means are statistically significant.

5.5 Discussion

The present experiment tested the effects of the corpus on the perceived correspondence of six audio-visual feature associations (i.e. 6x3). The primary purpose of the six A/V associations tested in the present experiments is to enable visual interaction with corpus based concatenative synthesis for creative applications. In an overview of the data gathered in this experiment it is worth noting that overall there are consistencies in the subjects' responses across the groups and conditions. The fact that there are consistency between the two subject groups (i.e. sound practitioner and non-sound practitioner) suggests that musical/sound training was not a significant factor; this is discussed in more detail later in this section. The fact that there are consistencies in the subjects' responses between the two experimental conditions (controlled environment and online) suggests that both approaches for gathering data are equally well suited for this experiment. Moreover, consistencies could also be observed in the participant responses between the first and the second question of the experiment (i.e. pairwise similarity judgment and rating). This consistency suggests that both data gathering methods are appropriate for measuring perceived similarity of cross-modal stimuli and could be used interchangeably.

In agreement with the first hypothesis, the experimental results from the first study revealed differences in the degree of perceived correspondence reported by the subjects between the individual A/V feature associations that were tested. The present study confirms the results revealed by previous studies (Eitan & Timmers, 2010; Evans & Treisman, 2010; Kostas Giannakis & Smith, 1993; Kohn & Eitan, 2009; Kussner & Leech-Wilkinson, 2013; Küssner, 2014; Lipscomb & Kim, 2004; Marks, 1989; Rusconi et al., 2006; R. Walker, 1987) which found strong relationships between the audio-visual feature associations of size – loudness, vertical position-pitch, color brightness– spectral brightness. Weaker were the relationships between texture granularity – sound dissonance and color complexity- sound dissonance, similar to the findings of (Giannakis, 2006). The weak correspondence reported by the subjects between these features of the auditory and the visual stimuli could be interpreted in two ways. Either the features that were tested (i.e. between texture granularity – sound dissonance and color variance- sound dissonance) are not a good match, or the synthesis parameter that was used to map the distance between visual texture granularity and sound dissonance (i.e. transposition randomness, selection randomness) are not the most appropriate parameters. Moreover, it is worth noting that auditory and visual textures are more difficult to define in computational and statistical terms due to the fact that both auditory and visual texture are higher dimensional features that consist of multiple lower level auditory and visual parameters. This is true, particularly if we compare auditory and visual textures to other auditory and visual features such as auditory pitch, loudness, brightness and visual size, position and brightness. Further research will be required to investigate which set of parameters are the most appropriate for mapping the distances between texture granularity – sound dissonance and color complexity- sound dissonance.

Contrary to the second hypothesis, the results showed that subjects' responses did not vary significantly as a result of the harmonicity of the source audio which was used to synthesise the stimuli. These findings suggest that the strength of the perceived correspondence between the A/V associations prevails over the timbre characteristics of the sounds used to render the complementary polar features. Hence, the empirical evidence gathered by previous research is generalizable/ applicable to different contexts and further the overall dimensionality of the sound used to render should not have a very significant effect on the comprehensibility and usability of an A/V mapping. An interesting trend can be observed in the interaction between the factors A/V associations and Corpus. The data show that the interactions between these two factors were greater in the case of the A/V association where perceived correspondence was weak (i.e. texture granularity– sound dissonance and color complexity- sound dissonance) than A/V association where perceived correspondence was strong (i.e. Size-Loudness, Vertical position-Pitch, Color and spectral brightness). An interpretation of the difference in the strength of the interactions between A/V associations and the harmonicity of the corpus, is that in the case of the strongly correlated A/V associations, the strength of the correspondence prevails to less important features of the harmonicity of the audio corpus. The strength of the A/V association dominates the subjects' judgement. While, in the case of weakly correlated A/V associations, the less important features of harmonicity become more influential in the subjects' similarity judgment. This would lead to the conclusion that the influence of the harmonicity of the audio corpus when making a similarity judgement is relative to the strength/dominance of the A/V association being tested.

Another interpretation related to the variation of the interaction between the A/V associations and Corpus could be based on the observation that for the weakly correlated feature dimensions the most highly rated corpus was the impacts corpus followed by string corpus, while the lowest rated was the wind. This could be attributed to the fact that the texture granularity– sound dissonance and color complexity- sound dissonance A/V associations are both related to a transition from consonance to dissonance that rely on textural/timbral feature of the sound. Hence, the wind corpus being the least harmonic corpus in comparison to the strings and the impacts corpora could render less well the transition from consonance to dissonance, Figure 54. The string corpus, although it was the most harmonic, was rated consistently lower than the impacts. This observation could be attributed to the fact that the string corpus was more homogeneous than the impact corpus in terms of periodicity and spectral flatness and brightness. If this is true then it could be argued that both the heterogeneity of the corpus as well as the harmonicity of the audio are important for rendering the transition from audiovisual consonance to dissonance. However

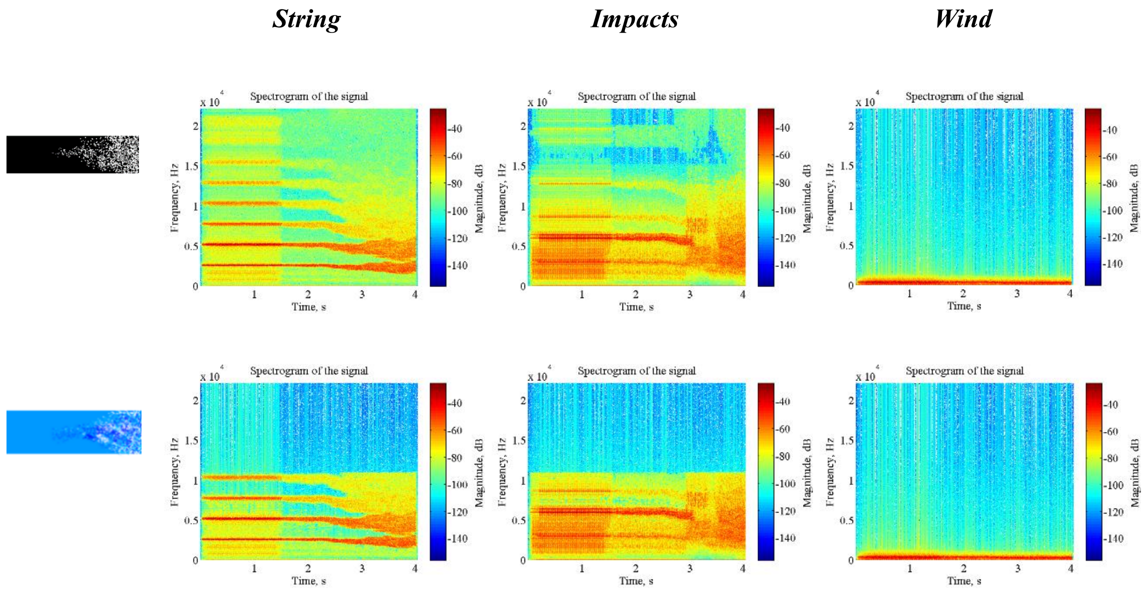


Figure 54. Shows that the wind corpus being the least harmonic corpus in comparison to the string and impacts corpora could render less well transitions between consonance and dissonance.

further research will be necessary to support this claim.

The fact that there was no significant differences between the expert and the non-expert group suggests that the cross-modal correspondences tested in this study are not dependent on the level of music/sound training of the subjects, which is in agreement with the third hypothesis. This finding are in agreement with (Lipscomb & Kim, 2004) findings and oppose the findings of (Kussner & Leech-Wilkinson, 2013; Küssner, 2014; R. Walker, 1987). The fact that musical/sound training was not a very significant factor suggests that the correspondences might be underpinned by either psychophysical, structural similarity between audio-visual feature dimensions and/or other cultural conventions, but not specifically to conventions related to the acquisition of musical/auditory skills. An interesting trend in the data is that the expert subjects appear to perceive a stronger correspondence between vertical position-pitch than color brightness-spectral brightness, while non-expert subjects appear to perceive a stronger correspondence between color brightness-spectral brightness than vertical position-pitch. In my interpretation, the expert subjects are more accustomed to the association vertical position- pitch as this association is very widely used in music software and digital audio workstation (e.g. sequencer) as discussed in Chapter 2.

5.6 Conclusions and future work

This experiment has verified and extended the analysis of results from previous research discussed in the introduction of this chapter, showing very strong correlation between size-loudness, vertical position-pitch, color brightness-spectral brightness. Furthermore, the analysis suggests that there is a non-linear interaction between the harmonicity of the corpus and the perceived correspondence of the audio-visual associations that were tested in this experiment. Strongly correlated dimensions such as size-loudness or vertical position-pitch are affected less by the harmonicity of the audio corpus in comparison to weaker correlated dimensions (e.g. texture granularity-sound dissonance). However further research will be necessary to determine whether the no-linear interactions between the harmonicity of the audio corpus and the A/V association is relative to the strength/dominance of the association or an effect that occurred in this specific experimental set-up. For future directions of this work, a number of follow-up experiments, altering the data gathering methods and the materials are required in order to assess to which extent our findings are generalizable. More specifically it will be necessary to test alternative combinations for the audio – visual feature associations for the feature associations that were not highly rated in experiment one. More studies will be conducted to evaluate the distance mapping between target and selected feature vectors to the synthesis parameters. Future studies will examine whether there are interactions between participants' similarity ratings and the statistical correlation of the audio-visual feature vectors of each A/V stimuli across the feature dimensions of the mappings.

6 Experiment 2- An Investigation of the Effects of the Mappings and Harmonicity of the Audio Corpus on the Subjects' Ability to Discriminate Between Complex Audio-Visual Stimuli

6.1 Introduction

Cross-modal correspondence is often studied by isolating the feature dimensions. Isolating feature dimensions enables testing dependency between one-to-one correspondences between cross-modal feature dimensions. However cross-modal sensory stimuli in the physical environment are often more complex than the stimuli used for experimental testing, entailing dynamic variation across multiple sensory parameters simultaneously. Hence, it could be argued that isolating cross-modal feature dimensions can obviously lead to ecological validity issues. As discussed in Chapter 3, research findings suggest that the relationships between corresponding A/V associations can be affected by two factors the interactions between multiple A/V feature dimensions that vary simultaneously and the overall dimensionality of the mapping (number of corresponding A/V associations). The first issue is related to the type of stimuli used, as according to research findings depending on whether the stimuli is static or dynamic can affect the perceived correspondence between cross-modal dimensions. For example, Eitan *et al.*, (2011) showed that although large visual objects are usually associated with low pitched sounds and small objects with high pitched sound, when the stimuli pitch is dynamically manipulated this effect can be reversed. For instance, it was found that rising pitch contour is associated with an increasing in size visual object and falling pitch contour with decreasing in size object. The second issue is related to interactions that occur between cross-modal features in dynamic multidimensional contexts. For instance, an increase in tempo rate is often correlated to an increase in visual motion speed, however when increase in tempo is accompanied by a decreasing loudness the correspondence between tempo and speed is weakened (Eitan, 2013).

This suggests that examining cross-modal correspondence at feature set level as oppose to isolated feature dimensions it is a more ecologically valid. Further, many researchers have argued that there is certainly need for the development of new techniques that enable to study cross-modal correspondence in multidimensional context (Eitan *et al.*, 2011; Küssner *et al.*, 2014; Marks & Eitan, 2012) where multiple dynamic co-variation of cross-modal features happen simultaneously.

In the context of the present research investigating how the entire feature set involved in the mapping affects the subject ability to discriminate A/V stimuli when multiple parameters are manipulated simultaneously is extremely important as it can affect the comprehensibility and effectiveness of the mapping. Further, as the synthesis method used in the interfaces presented in this thesis (i.e. Morpheme) is corpus depended (i.e. uses source audio material in order to synthesise sound that can vary both quantitatively and qualitatively) assessing whether the harmonicity of the audio corpus affect the saliency of the A/V corresponding dimensions is also very important as it can affect usability of the interface. The aim of the present study is to test whether the subjects' audio-visual discrimination ability is affected by the harmonicity of the audio used to render the A/V associations and the mapping (i.e. two mapping each consisting of three A/V feature associations). The two mappings (described in sections 4.6.3 and 4.6.5) consists of A/V associations that are, according to previous studies, some of the most highly rated A/V feature correlates. While previous studies have explored perceived congruency between A/V feature dimensions, no previous study has examined how the perceived congruency is affected by the harmonicity of the sound used to render the A/V association and very few studies have investigated A/V correspondence in multidimensional contexts. Below follows the description of a discrimination experiment using the discrimination ability of the subject as an indicator of the comprehensibility and effectiveness of the mapping. Musical and sound training factors were also measured in this experiment.

6.2 Hypothesis

Four primary research questions were investigated.

Q1: Will the harmonicity of source audio that the corpus consists of have a significant effect on the ability of the participants to discriminate the A/V stimuli?

H1: There will be a positive correlation between the harmonicity of the audio-units and the number of correct detections, as the harmonic contain can cause a reduction in the salience of the A/V associations involved in the mapping. Reduction in the salience of the association is expected to increase the level of difficulty of the task and as a result the likelihood of failing to respond correctly is higher.

Q2 Does mapping (i.e. chromatic/achromatic) as a whole have a significant effect on the ability of the participants to discriminate?

H2: No, as both mapping have equal number two A/V dimensions that are strong correlates.

Q3: Will the confidence of the participants vary as a result of whether they have selected or not the correct image?

H3: There should be a positive correlation between the participants' confidence ratings and whether they have detected the correct image or not.

Q4: Will the ability to discriminate audio-visual stimuli differ between the two groups (i.e. sound practitioners and non-sound practitioner)?

H4: Yes, as it is expected that training will have a positive effect on the ability of the participants to discriminate complex and dynamic A/V stimuli.

6.3 Procedures

The same number of subjects and subject groups as in the experiment that was described 5.1.4, also apply in this experiment. This experiment was performed almost immediately after the first experiment which was described in section 5. The two experiments were very different, therefore there was no need to randomise the order between the two experiments. Conversely, due to the fact that the second task was difficult, the sequence of the two experiments was appropriate. Subject responses were collected independently. In each session a single participant completed the following tasks. Instruction about the task were displayed on the screen. After reading the instruction and before beginning the main experiment, each participant did a warm-up trial consisting of three audio-visual examples to confirm their understanding of the task, become familiar with the apparatus and the response procedures. After the training session which lasted no more than a few minutes, participants continued with the actual experiments. The transition between the warm-up task and the main experiment was seamless without interruptions. In this study participants were presented with three images per audio stimulus. Subjects were told during the training session that only one of the three images displayed on the screen was used to generate the sounds they would hear. By clicking an on-screen button using a computer mouse, subjects could playback the audio files. After listening to each audio stimulus, participants could use an on-screen radial button to indicate the image they thought was used to generate the sound. After participants indicated their responses, they were asked to rate how confident they were about their decision. Subjects could playback the stimuli as many times as they wished. The task takes approximately five minutes to complete. The order that the stimuli was presented to each participants was randomized to avoid effects on their responses due to biases.

6.4 Audio Stimuli

Eight audio stimuli were synthesised for this study. Four audio stimuli were tested for each mapping, i.e. four sounds per mapping, resulting in a total of eight audio stimuli. Each sound was synthesized using one of the three corpus (i.e. String, Wind, Impacts) which were described in detail in the section 5.2 above. The resulting sounds used as stimuli have a duration of 8 seconds. An additional corpus was used in this experiment that consisted from a large number of bird sounds. The four corpora differ in term of size (i.e. number of audio units) and harmonicity. The string and the birds' corpora consisted of manly harmonic and periodic sounds, which the wind and the impacts corpora consisted of mainly unpitched, non-harmonic sounds. The string and the wind corpora consisted of a small number of audio-units that are very homogenous (i.e. all audio-units had very similar characteristics (loudness, periodicity, pitch)), while the impacts and the bird corpora consisted of audio-units that were non- homogeneous (i.e. more variation in terms of loudness, periodicity and frequency contain).

6.5 Visual Stimuli

Six visual stimuli were designed for this experiment. Unlike the first study where the aim was to test one feature association at a time (i.e. isolate features dimensions), here the aim is to test all the associations used in the mapping and to manipulate multiple features values simultaneously. As mentioned earlier in this study, participants had to select one out of three images. The three images presented for each audio stimuli have been designed to be similar and vary only in some respects, see Figure 55. For instance, in the achromatic mapping images 1 and 2 are similar in terms of pitch contour, while images 1 and 3 are similar in terms of texture granularity. In the chromatic mapping, images 1 and 2 are more similar in terms of pitch contour, while the images 2 and 3 are more similar in terms of texture granularity of color variance. This decision aimed at increasing the level of difficulty and the probability of confusing the participants. The stimuli, the data gather from this study and the statistical analysis tables are available in the appendix D, please see chapter 9.

6.6 Experiment-2 Data Analysis

6.6.1 Question-1: results

This section presents the results obtained from the first question of the experiment, were non-expert and expert subjects in the supervised and unsupervised condition (controlled environment and online study) performed a series of A/V detection tasks. Figure 56, shows the percentages of the subjects who detected correctly the image (out of three images) used to synthesise the audio stimulus. Pearson's Chi-square tests of independence were performed to examine the relationships and test the null hypothesis of no association between the factors (i) subjects' detection success rate and harmonicity of the corpora, and (ii) Subjects' detection success rate and A/V mappings.

Data analysis of non-expert group in the supervised condition shows that there is a strong relationship between the ability of the subjects to detect the correct image and the harmonicity of the audio corpora $X^2(3, N = 224) = 19.22, p = .0$. Overall no significant relationship was revealed between detection success rate and the mapping $X^2(1, N = 224) = 1.143, p = .28$.

Data analysis of the expert group in the supervised condition show that there is strong relationship between the ability of the subjects to detect the correct image and the harmonicity of the audio corpora was observed $X^2(3, N = 128) = 9.86, p = .02$. Overall no significant relationship was revealed between subjects success rate and the mapping X^2*(1, N = 128)* = 3.33, *p* = .068.

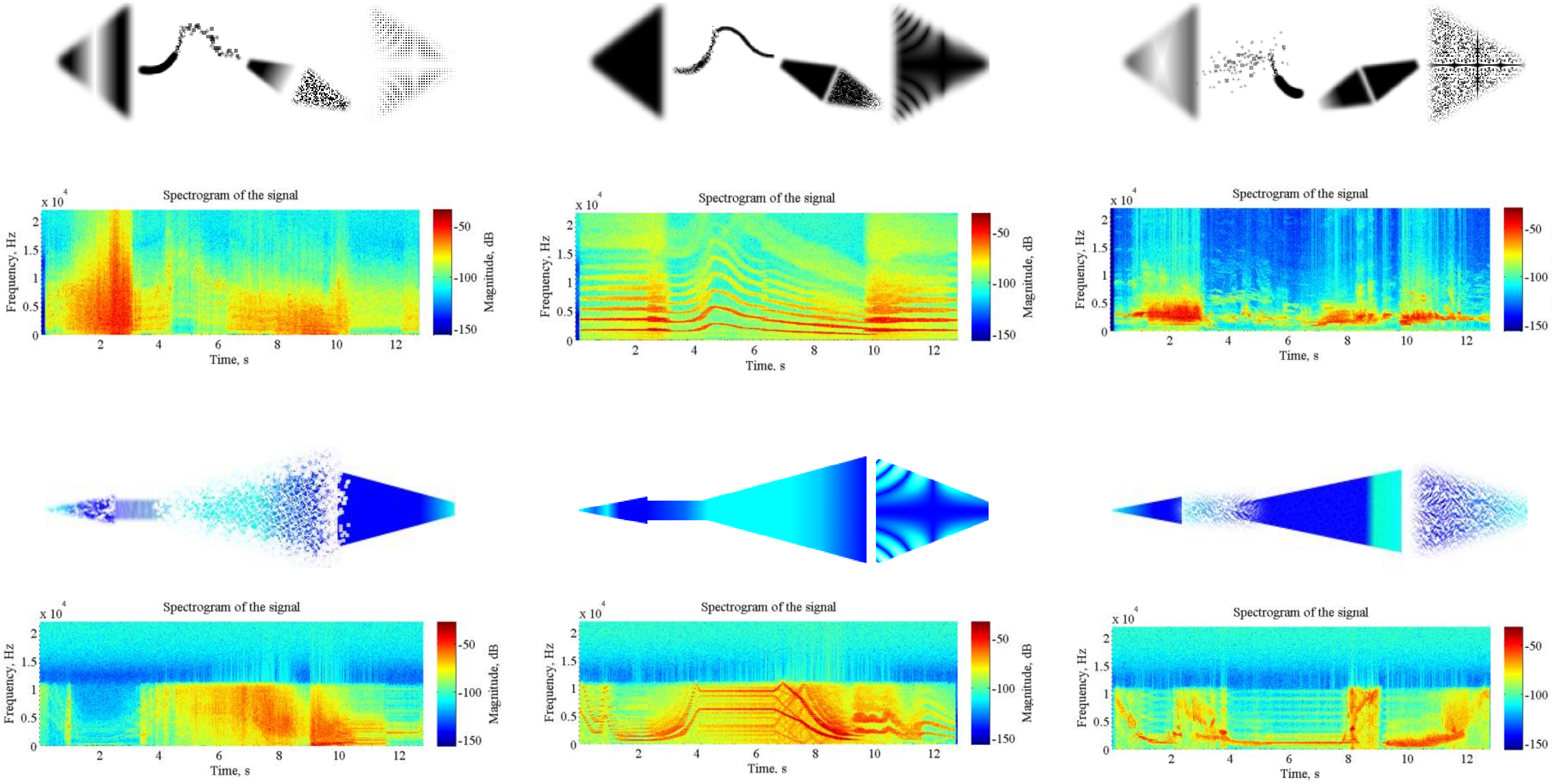


Figure 55. Example of visual and audio stimuli used in the second study for testing.

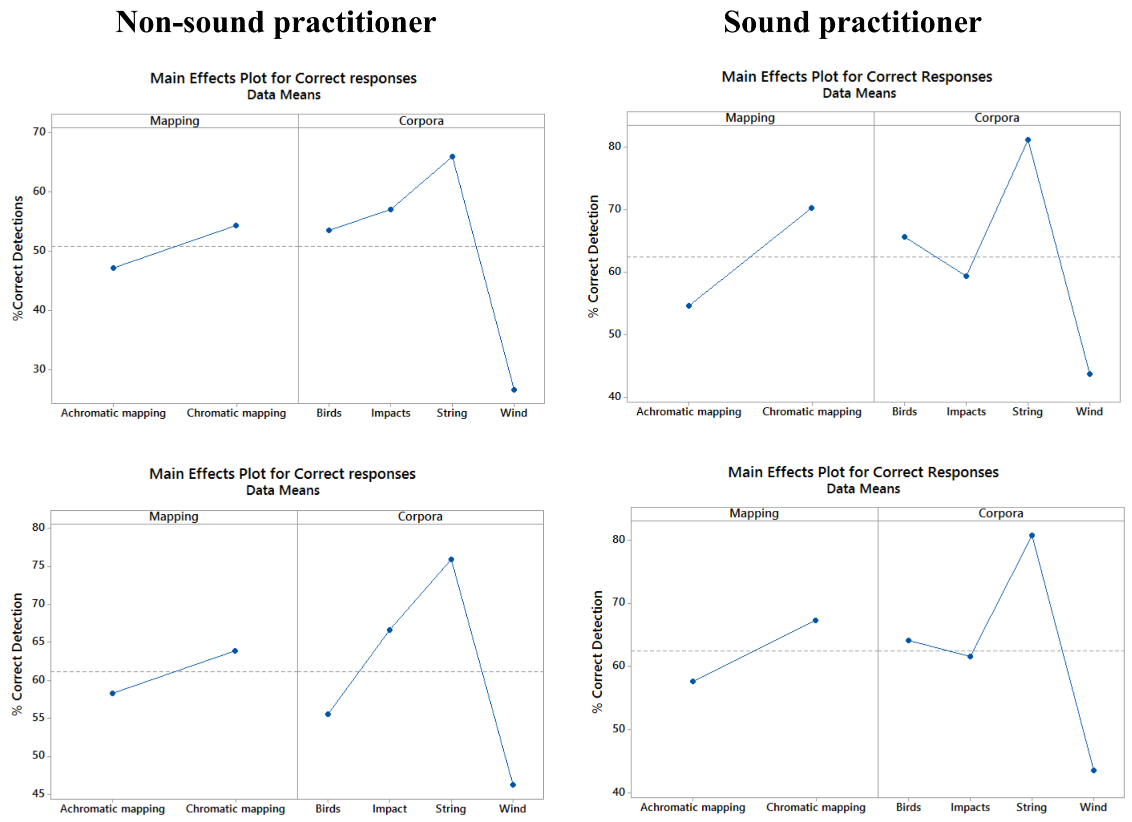


Figure 56. Overall effects of the mappings and the corpus on the subjects' ability to detect the correct stimuli. The dotted line represents the overall average of the subjects who rated the A/V associations as similar. Upper row shows the results for the supervised condition and lower row for the unsupervised (i.e. study conducted as online survey).

Data analysis of the non-expert subjects in the unsupervised condition shows that there is a strong relationship between the ability of the subjects to detect the correct image and the harmonicity of the audio corpora $X^2(3, N = 216) = 11.37, p = .010$. Overall no significant relationship was revealed between the variables detection success rate and the mappings $X^2(1, N = 216) =.70, p = .40$.

Data analysis of the expert subjects in the unsupervised condition shows that there is a strong relationship between the ability of the subjects to detect the correct image and the harmonicity of the audio corpora $X^2(3, N = 312) = 23.15, p < .001$. Overall no significant relationship was revealed between the variables detection success rate and the mappings $X^2(1, N = 312) =.3.07, p = .079$. Figure 57 show that the *String* and *Wind* corpus contribute more to the strength of the relationship between corpus and perceived similarity in all of the groups and conditions. Post-hoc testing of each group/condition chi-square results was performed for each level of the factor corpus in order to determine which factor levels contribute to the statistical significance that was observed. The results of the post-hoc test are presented in Table 15.

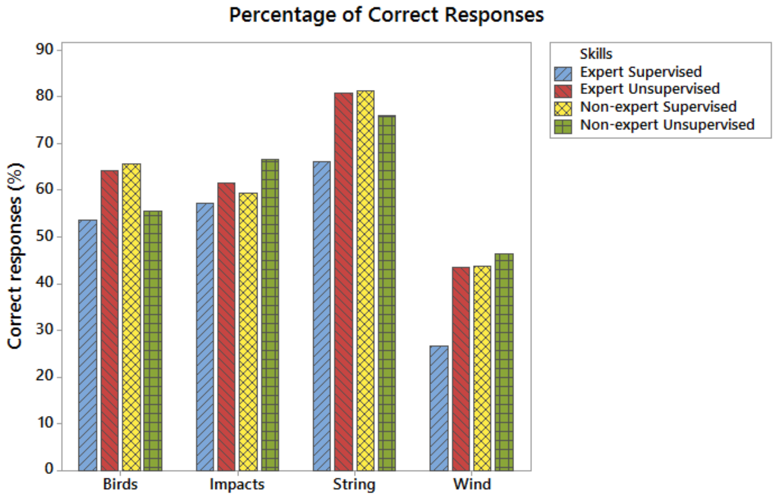
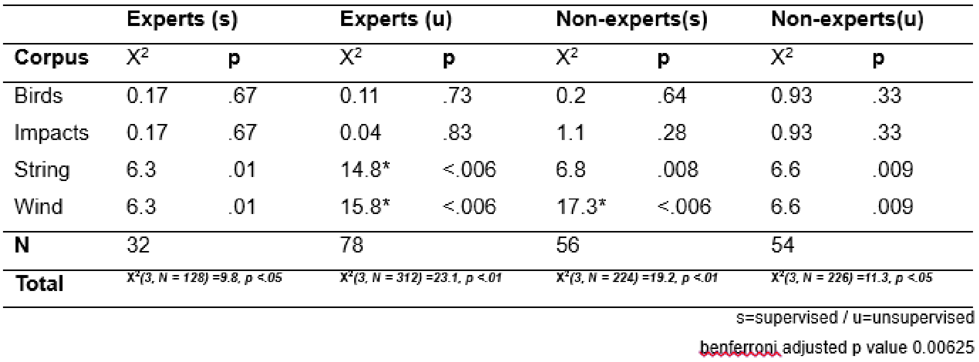


Figure 57. Percent of correct detection for each corpus.

Table 15. Post-hoc testing of each group/condition chi-square results by estimating p values from the chi-square residuals for each level of a factor and comparing these to an adjusted benferonni corrected p value. An asterisk (*) indicates statistical significance at adjusted alpha level of .006.



	Experts (s)		**Experts (u)**		**Non-experts(s)**		**Non-experts(u)**	
Corpus	X^2	p	X^2	p	X^2	p	X^2	p
Birds	0.17	.67	0.11	.73	0.2	.64	0.93	.33
Impacts	0.17	.67	0.04	.83	1.1	.28	0.93	.33
String	6.3	.01	14.8*	<.006	6.8	.008	6.6	.009
Wind	6.3	.01	15.8*	<.006	17.3*	<.006	6.6	.009
N	32		78		56		54	
Total	X^2*(3, N = 128) =9.8, p <.05*		X^2*(3, N = 312) =23.1, p <.01*		X^2*(3, N = 224) =19.2, p <.01*		X^2*(3, N = 226) =11.3, p <.05*	

s=supervised / u=unsupervised

benferroni adjusted p value 0.00625

This section presents the results obtained from the comparison between the expert and the non-expert groups in the supervised and unsupervised condition. Pearson's chi-square test of independence was performed on three between subject variable to test the null hypothesis of no association between: (i) the subjects' skills and ability for correct discrimination of the A/V stimuli, and (ii) subjects' skills and the frequency of correct detection between the two mappings, and (iii) between the four corpora.

The comparison between the expert and the non-expert groups in the supervised condition shows that there is an association between the variables subjects skills and discrimination ability X^2*(1, N = 352)* = 4.43, *p* = .035. Examination of the cell frequencies showed that about 70% (45 out of 64) of the non-expert subject responses were correct, while the percentage of expert subjects who detected the correct image was 54% (61 out of 112). Further testing of the relationship between the variables correct detection and skills was performed by layering the participant responses based on the two levels of the mapping factor. The results of the test show that the association between the subjects' skills and correct discrimination was stronger when the chromatic mapping was tested $X^2(1, N = 176) = 4.27, p = .039$ and not strong when the achromatic mapping was tested $X^2(1, N = 176) = .88, p = .347$, see Figure 58. However, Post-hoc testing show that the relationship between skills and conditions was not significant, see Table 16.

The comparison between the expert and the non-expert groups in the unsupervised condition show no association between the variables subjects skills and discrimination ability $X^2(1, N = 528) = 0.10, p = .747$. Further testing of the relationship between the variables correct detection and skills was performed by layering the participant responses based on the two levels of the mapping factor. The results of the test show that there was no association between the subjects' skills and correct discrimination for none of the mappings: achromatic $X^2(1, N = 264) = .011, p = .917$; chromatic mapping $X^2(1, N = 264) = .332, p = .564$.

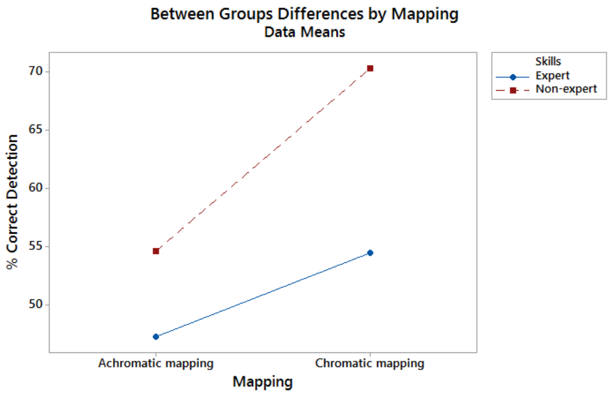


Figure 58. Differences between the expert and non-expert subjects in the supervised condition' discrimination ability as a result of the mapping.

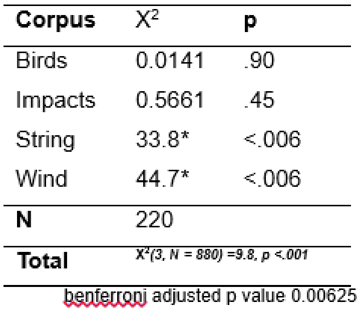
Table 16. Post-hoc testing of each group/condition chi-square results for subjects' skills factor.

Skills	**chi-square**	**p**	**N**
Expert (s)	.6	.41	128
Expert (u)	2.1	.14	312
Non-Expert (s)	8.5	.0033	224
Non-Expert (u)	.43	.51	216
Total	$X^2(3, N = 880) = 8.711, p = .033$		

Benfertoni adjusted p value 0.00625

Since there were no significant differences between the two subject groups in the two conditions, the results from all groups/conditions were merged in order to perform a chi-square test on the merged dataset, for the corpus factor that was the only factor in which we found statistically significant differences in the participant responses. The data analysis of the merged dataset shows that there was a statistically significant relationship between correct detection and corpus: X2(3, N = 880) = 9.8, p < .001. Post-hoc testing was performed to identify which levels of the factor corpus were contributing to this statistical difference. The results show that the string and the wind corpus were the factor levels that contributing to the association between participants' discrimination ability and corpus, see *Table 17.*

Table 17. Post-hoc testing of the merged (group/condition) chi-square results for the corpus factor. An asterisk (*) indicates statistical significance.



Corpus	X^2	p
Birds	0.0141	.90
Impacts	0.5661	.45
String	33.8*	<.006
Wind	44.7*	<.006
N	220	
Total	$X^2(3, N = 880) = 9.8, p < .001$	

benferroni adjusted p value 0.00625

6.6.2 Question-2: results

This section presents the results obtained from the second question of the experiment, were expert and non-expert subjects in the supervised and unsupervised conditions reported on their confidence levels regarding their decision (i.e. corresponding image for each audio stimuli). A repeated measure ANOVA using a general linear model was computed with the confidence ratings as within subjects factor and the corpora and the mapping as between subjects factors. The analysis aimed to test the null hypothesis of no association between the subjects' confidence ratings and the mappings, or the audio corpora.

Analysis of the subjects' confidence ratings revealed no significant differences due to the corpus with an exception for the non-experts in the unsupervised condition. Significant differences were revealed in the data means of the expert group in both conditions for the factor mapping, while no significant were the differences for the expert participants in the two conditions. While no interactions between the factors mapping and audio corpus were revealed. The results of the analyses are shown in Table 18. Figure 59 and Figure 60 shows the data means for all of the participants confidence ratings for the corpus and the mapping factors respectively. An additional ANOVA model was computed to examine the relationship between reported confidence and correct detections which was found to be significant for all the groups with only exception the expert group in the supervised condition. The results of the analyses are shown in Table 19.

Table 18. General ANOVA model computed to investigate the effect of the A/V associations, corpus and mapping on the perceived similarity reported by the subjects. An asterisk (*) indicates statistical significance.

	Corpus		**Mapping**		**Interactions**		
Group	***F***	***P***	***F***	***P***	***F***	***P***	***Total***
Non-experts (s)	2.25	.1	.9	.4	1.89	.1	223
Non-experts (u)	.28	.8	14.14*	<.001	1.93	.1	215
Experts(s)	.70	.5	2.26	.1	0.75	.5	127
Experts(u)	4.74*	<.005	15.51*	<.001	1.7	.1	311
df	3		1		1		

s=supervised / u=unsupervised

Table 19. General ANOVA model computed to investigate the effect of correct detection on the reported confidence level. An asterisk (*) indicates statistical significance.

	Correct detection		
Group	***F***	***P***	***Total***
Non-experts (s)	7.53*	<.005	223
Non-experts (u)	6.15*	<.05	215
Experts(s)	.34	.5	127
Experts(u)	16.73*	<.001	311
df	3		

s=supervised / u=unsupervised

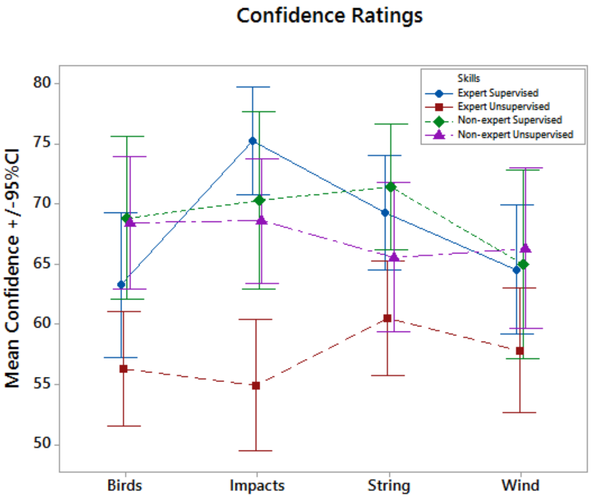


Figure 59. Data means and confidence intervals of participants' confidence ratings for the corpus factor for all of the groups. If intervals do not overlap the corresponding means are statistically significant.

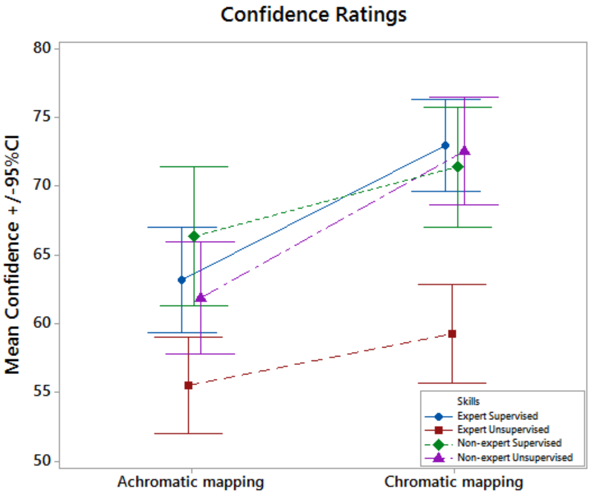


Figure 60. Data means and confidence intervals of participants' confidence ratings for the mapping factor for all of the groups. If intervals do not overlap the corresponding means are statistically significant.

6.7 Discussion

The present experiment tested the effects of the corpus and the mapping on the ability of the subject's to discriminate audio-visual stimuli. This experiment used the discrimination ability of the subject as an indicator of the comprehensibility and effectiveness of the mappings. The primary purpose of the mappings that were tested in the present experiment is to enable interaction with corpus-based concatenative synthesis for creative applications Musical and sound training factors were also measured in this experiment. In an overview of the data gathered in this experiment it is worth noting that overall there are consistencies in the subjects' responses across the groups and conditions. The fact that there are consistencies between the two subject groups (i.e. sound practitioner and non-sound practitioner) suggests that musical/sound training was not a significant factor. Furthermore, the fact that there are consistencies in the subjects' responses between the two experimental conditions (controlled environment and online) suggests that both approaches for gathering data are equally well suited for this type of experiment.

In agreement to my hypothesis, experimental results from this study revealed that participants' success rate in detecting the correct image did not vary significantly as a result of the mapping. However, some general trend that can be observed in the dataset is that both chromatic and achromatic mapping enabled participants to detect images well above chance levels, chance level being 33.3% given that there were three potential image matches for each audio stimulus. Overall the chromatic mapping enabled participants to correctly detect more images than the achromatic mapping, although this difference was not statistically significant.

In agreement to my hypothesis, the harmonicity of the audio corpus used to synthesise the audio stimuli from the images has significant effect in the subjects' ability to detect the correct image. Furthermore, as it was predicted, the detection success rate was higher when the string corpus was used, followed by the birds and the impact corpus, while the lowest success rate was observed when the wind corpus was used. The analysis shows that the corpora that had the strongest effect in the subjects' successful detection rate were the string and the wind. Moreover, it is worth noting that the non-sound practitioners' correct detection rate in the supervised condition when the wind corpus was used was 27.6% well below chance levels (i.e. 33.3%). While for the rest of the groups and conditions, the success rate was not far above chance levels.

Contrary to the results from the first study, in the context of the second study the harmonicity of the source audio which the audio corpus consists of appears to be important. A first interpretation of the effect of the harmonicity of the audio corpus in the ability of the subjects to detect the correct image is that when the sound corpus is not harmonic and continuous, the resulting sounds can be noisy and lack clarity. This could affect the effectiveness of the mapping by causing a reduction in the salience of the audio-visual associations. This in turn weakens the ability of the participants to pay attention to the causal relationship between the image and the sound. However further research will be necessary to support this claim.

Another interpretation of the divergence in the results between the two studies regarding the influence of the harmonicity of the audio corpus, is that in the first study the task was easier and less demanding from a cognitive point of view. In the second task, multiple audiovisual parameters were manipulated simultaneously and the images were very similar and due to these factors the decision which subjects were asked to make was by far more complex in comparison to the first task. In the second study there is a greater demand to detect subtle differences, forcing the participant to actively seek for cues to determine which image is the correct one. As a result, the clarity of the sounds which the corpus consists of became an important factor. So, our conclusion is that, in the context of corpus-based synthesis, the salience and efficacy of the cross-modal associations involved in a multidimensional mapping are to a degree dependent on the typological features of the source audio which the corpus consists of. Hence, the effectiveness of mappings that link user sensorimotor actions to audio parameters for the control of sound and music is subject to the qualitative characteristics of the sound used for testing the mapping. However further research will be necessary to assess the degree of this effect.

Contrary to my hypothesis, non-sound practitioner subjects performed overall better than expert subjects in both the supervised and the unsupervised experiment, however the difference between the two groups was not significant. As in the previous experiment, my interpretation of the fact that there was no significant differences between the expert and the non-expert group is that the cross-modal correspondences tested in this study are not dependent on the level of music/sound training of the subjects. These findings are in agreement with (Lipscomb & Kim, 2004) findings and oppose the findings of (Kussner & Leech-Wilkinson, 2013; Küssner, 2014; Walker, 1987).

Finally, the participants' confidence ratings revealed no statistically significant correlation between confidence levels and correct detection. Although overall a trend can be observed between correct detection and confidence levels reported by subjects (i.e. confidence levels on average are higher when participants have responded correctly rather than incorrectly). However the confidence levels reported on incorrect responses are very high (i.e. in most cases well above 50%), which does not indicate that participants were aware that their responses were incorrect. Overall, participants felt more confident when the chromatic mapping and the string corpus were used. Furthermore, the participants' confidence ratings show that when the wind corpus was used participants felt least confident, however for the other three corpora the results were not following a strong correlation pattern.

6.8 Conclusion and future work

This experiment has verified and extended the analysis of results from previous experiments discussed in the introduction of this chapter, showing a correlation between the harmonicity of the corpus and the ability of participants to discriminate audio-visual stimuli. The results suggest that sound/musical training had no significant effect on the discrimination ability of the subjects. My conclusion is that the audio-visual correspondences tested in the experiment (particularly the strongly correlated ones) are a product of either psychophysical, structural similarity or linked through other cultural conventions/factors, but not specific conventions or learned relationships that are related to the acquisition of musical/auditory skills. In the experiments presented in this and previous chapters, two methods were explored, non-speeded pairwise similarity judgments and multiple item discrimination. For future directions in this field of research, I would propose to conduct more experiments exploring other methods for obtaining similarity data such as free sorting and spatial arrangement methods, speeded classification and discrimination tasks.

7 Morpheme evaluation

This study aims to detect usability issues of the Morpheme interface and gather participants' opinions regarding cognitive, experiential and expressive aspects of the interaction with the interface developed in this study (i.e. Morpheme). The evaluation was performed after the participants had completed the experiments described in Chapters 5and 6.

7.1.1 Participants

One group was recruited that consisted of eleven musician/sound practitioner volunteers. All of the participants played a musical instrument and the self-reported level of expertise was five intermediate and six advanced. Seven of the participants had received formal music theory training at least for six months. All of the participants reported using analogue and digital equipment for sound synthesis, signal processing and sequencing. Four participants self-reported a level of expertise regarding the use of digital and analogue equipment as intermediate and seven reported advanced skills. None of the participants in this study reported having hearing or visual impairments. All participants had first participated in the experiments described in the chapters five and six prior to taking part in the present one. All participants were male while the age group ranged from 18-64.

7.1.2 Apparatus

The experiments took place in the Auralization room at the Merchiston Campus of Edinburgh Napier University. Participants used Beyer Dynamics DT 770 Pro monitoring headphones with 20db noise attenuation to listen to the audio stimuli. An HP ENVY dv7 laptop with 17.3 inch screen was used. For sketching on Morpheme's digital canvas a bamboo fun tablet was used. However participants were allowed to use a computer mouse if they preferred. SurveyGismo was used to record the participants' responses after the sound design task was completed.

7.1.3 Procedures

In this study participants were asked to design two soundscapes using the *Morpheme* interface for two video footages. Subject responses were collected independently. In each session a single participant completed the following tasks. Participants were given a brief description of the task followed by a short demonstration of Morpheme's graphical user. After a short training session were participants were shown how to use the graphical user interface of Morpheme in order to synthesize sounds, participants were instructed to proceed with the tasks. There were two eight minutes sessions (one for each video footage) during which participants were free to produce a soundscape that best suited the video using Morpheme. At the end of the sessions, participants were asked to complete a questionnaire. The questions aimed at assessing experiential, cognitive and expressive aspects of the interaction as well as to detect usability issues and gather ideas regarding usability improvements of the interface, please see Appendix E.

7.2 Stimuli

7.2.1 Video Footage

Two videos have been selected for this task. The first video footage has been captured in Bermuda during the recent hurricane Igor. The duration of the hurricane video is one minute. The camera shots included in the video have been captured from several locations during the hurricane. The second footage is a 3D animated scene that last for 4 seconds which represents a simulation of two porcelain objects been shattered on a tilled floor, see bottom row Figure 61. Both video footage require a relatively high precision in the way the sound is synced to the video sequence. However the second video sequence is slightly more challenging in this respect in comparison to the hurricane scene.

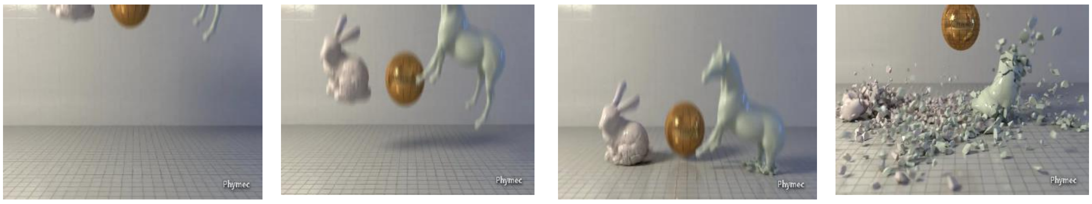


Figure 61. Four screenshots from the two video footage used in the study.

7.2.2 Audio Corpora

The audio corpus that participants had to use to synthesize the sound effects for the shattering scene consists of four audio recordings of glass shattering events. The corpus that is used to synthesize the soundscape for the hurricane scene consists of four audio recordings of windy acoustic environments. All eight audio files have been segmented to audio-units with durations of 242 milliseconds. The selection of the audio files used to prepare the two corpus was predominately determined by the theme of the video footage. However these two video footage were selected to allow testing the mapping in two very different auditory contexts. For example the shattering scene requires a corpus that consists of sounds that are relatively dissonant, non-periodic, and abrupt such as impact/percussive sounds. The second hurricane scene requires a corpus that contain moderately harmonic, slightly periodic and continues sounds. The stimuli, the data gather from this study are available in the appendix E, see Chapter 9.

7.3 Results

Figure 62 shows the mean and standard deviation of the participant Likert scale responses. The first question aimed at assessing participant satisfaction of the sounds designs created using Morpheme, see Table 20. Participant average response shows that participants were neutral regarding this question. Participants' responses show that there was a strong correlation between the user input (i.e. sketch) and the outcome sound, and that it was easy to understand the mapping. Although the degree of correlation was not as strong at all times. Participants' responses indicate that Morphemes' sketching interface help them to articulate their sound design ideas in visual terms, and that they felt they had control over the sound synthesis parameter. However the responses also indicate that more precise control of the audio parameters would be desired. Further participants felt equally in control using either corpus (i.e. wind and impacts) while there was indication that there was a stronger preference in working with the impacts corpus. Finally, participants agreed that Morpheme offers an interesting model for interaction with sound synthesis parameters and that it would be a useful addition to the sound synthesis tools they already use.

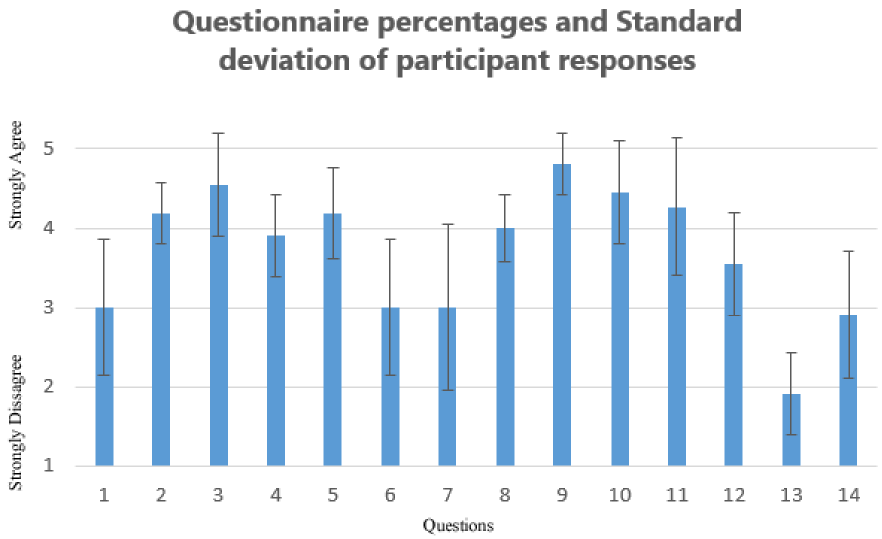


Figure 62. Results obtained by the questionnaire during the evaluation of Morpheme.

Table 20. The statistics of the Likert type questions that were answered by participant answers

	Questions	Average	STD
1	I am satisfied with the sound I designed using this mapping.	3	0.85
2	I felt there was a strong correlation between the sketch and the sound that was synthesised by the system.	4.18	0.38
3	I felt I understood how attributes of the sketch were associated to attributes of the sound.	4.54	0.65
4	I felt I could articulate my creative intentions using this mapping.	3.9	0.51
5	I felt I had control over the synthesis parameters while using the system.	4.18	0.57
6	I am satisfied with the level and precision of the control I had over the audio parameters while using the system.	3	0.85
7	I felt confused in several occasions about how my drawing affected the audio output.	3	1.04
8	Overall, I am satisfied with Morpheme's Graphical User Interface.	4	0.42
9	I believe that Morpheme offers an interesting approach to interacting with sound synthesis.	4.81	0.38
10	I believe that Morpheme would be a useful addition to the audio tools I currently use.	4.45	0.65
11	I felt Morpheme helped me think about sound in visual terms.	4.27	0.86
12	I felt equally in control while using the two sound corpora.	3.54	0.65
13	I felt frustrated about certain aspects of the interface/interaction.	2.9	0.79
14	I felt that Morpheme was complicated and difficult to use.	1.9	0.5

An analysis of the data gathered by the open ended questions was performed manually. Every time a new theme was encountered in the answers, it was used to form a new *category*. Then the frequency of these categories was recorded to identify which are the most prominent issues and desired technical features. An analysis of the open-ended questions revealed the following usability improvements, see Table 21 and Table 22. Table 21 presents the participants suggestions for improving the user interface, while Table 22 presents which aspects of the interaction participants found most frustrating.

Table 21. Participant’s answers to the question: What changes to the User Interface would you suggest to improve it?

Participants suggestions for user interface improvements	**N**
Image processing tools for refinement of the sketch	1
Timestamps navigation of the timeline	2
Edit the position of graphics based on timestamps	1
Larger canvas	5
Canvas zoom-in function	3
Temporal looping function based on user defined loop points	1
Undo function	1
Latency between graphics and audio timeline	1
Non-linear sketch exploration	1
Enable layering of multiple sounds/sketches and ability to shift between layers	1

N=Number of participants

Table 22. Participant's answer to the question: Which aspects of the interface/interaction you think were the most frustrating?

Participants’ usability complains	**N**
Audio interruptions when performing a brush stroke	1
Canvas Size	3
Latency between graphics and audio timeline	4
Larger canvas	5
Zoom-in function	3
Lack of precision using the brush	1
Eraser would not completely erase the sounds	2

N=Number of participants

7.3.1 Discussion

Based on the results presented above, it can be concluded that overall Morpheme achieve satisfactory levels of performance. The subjective level of control of the sound parameters through sketching, and the participants level of satisfaction with the sounds they designed was average. These results might be attributed to three factors. The first factor it might be related to the user’s unfamiliarity with sketching as a model of interaction with sound synthesis parameters. The second factor might be related to the unfamiliarity of the participants with the way concatenative synthesis works. This view is further supported from the average Mean=3 Std Div=1 responses in the question ‘I felt confused in several occasions about how my drawing affected the audio output'. This is also reflected in some of the user comments, for example:

"Unpredictable results at times",

"it wasn't always easy to be precise",

"Navigating was complicated at times, as so it was to identify the correlation between the pitch and the type of sounds played".

The way concatenative synthesis works is different to that of other synthesis methods. For example while in other synthesis methods increasing the amplitude or the pitch synthesis parameter results to changes only the parameters that were control, in the context of concatenative synthesis controlling the amplitude or the pitch may result in selecting different audio units that have very different timbre characteristics. This sudden and discreet changes are likely to have confused practitioners that are not familiar with the synthesis method. As it was mentioned earlier in section 7.1.3 the information that was provided to the participants prior to the experimental task was mainly about how to user interface and minimal information was provided about the synthesis method. This decision was made primarily to avoid the development of positive biases towards the system due to enthusiasm about the way the system synthesises sound. The third factor might be related to the usability issues discussed in the next paragraph. Overall, the perceived correlation between the visual and sound features were satisfactory. Participant responses showed that Morpheme is easy to use, offers an interesting approach to interacting with sound synthesis and supported that the interface helped them think about sound in visual terms. Furthermore, the majority of participants thought that Morpheme would be a useful addition to the audio tools they currently use. Participants responses were not conclusive as to whether the corpora that was used affected their perceived level of control over the system as participants response was Mean= 3.5 Std Div=0.6, while seven out of eleven participants seem to prefer working with the impacts corpus, three preferred the wind corpus and one neither. One of the differences between the impacts and the wind corpora is that the former is much larger. Based on the findings from the evaluation it appears that larger corpus can result in both positive and a negative effects. Some of the negative effects became evident from some of the participants comments discussed above such as more unpredictable results, because the probabilities of getting a sequence of audio-units with very distinct timbre is higher when there is a large nonhomogeneous corpus (e.g. impacts corpus used for the evaluation) than when a small and homogeneous corpus (such as the wind corpus) is used.

Many usability issues were also revealed, mainly related to the lack of standard controls found in other image processing applications (e.g. photoshop) such as zooming in and out, resize canvas and undo function. Further, participants also pointed out the lack of other functions that tend to be standard functionality in time-based media production applications such as setting loop and cue points on the timeline, having a precise transport panel and a sequencer were sounds can be layered. Moreover, several participants complained about latency between the timeline and the output sound. Latency depends on two factors: the size of the audio corpus (i.e. how many audio units are stored in the corpus) and how many comparisons the algorithm has to perform until it finds the audio-unit that its features best match the target. Another factor that might cause the perception of latency is that in the present version of morpheme, the current position of the analysis window is indicated by a slider that does not reflect well the actual position of the window, see left image in Figure 63. The problem is that the window is 9 pixels wide while the current cursor used to the representation of the position of the analysis window suggest that the window is smaller. A better solution would be to use a cursor as shown to the right of Figure 63. Figure 64 and Figure 65 shows a few examples of the images drawn by the participants for the hurricane and shuttering scene respectively. Figure 64 and Figure 65 shows a few examples of the images drawn by the participants for the hurricane and shuttering scene respectively.

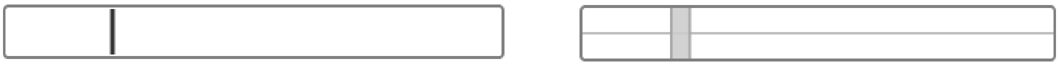


Figure 63. The left figure shows the current visual feedback for the representation of the position of the analysis window. The right figure shows a more precise visual feedback.

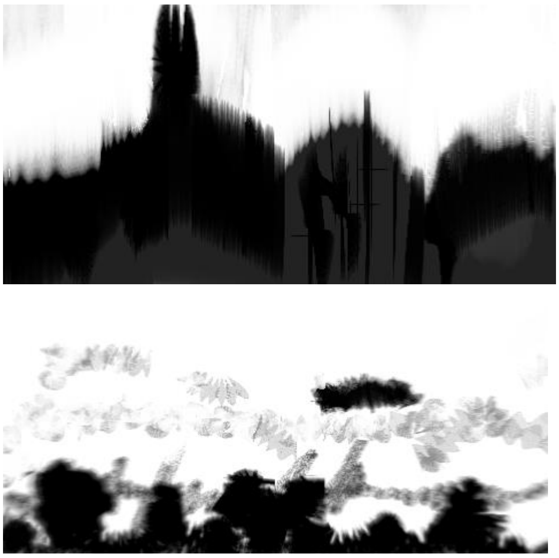


Figure 64. Three sketches drawn by the participants in response to the hurricane video footage.

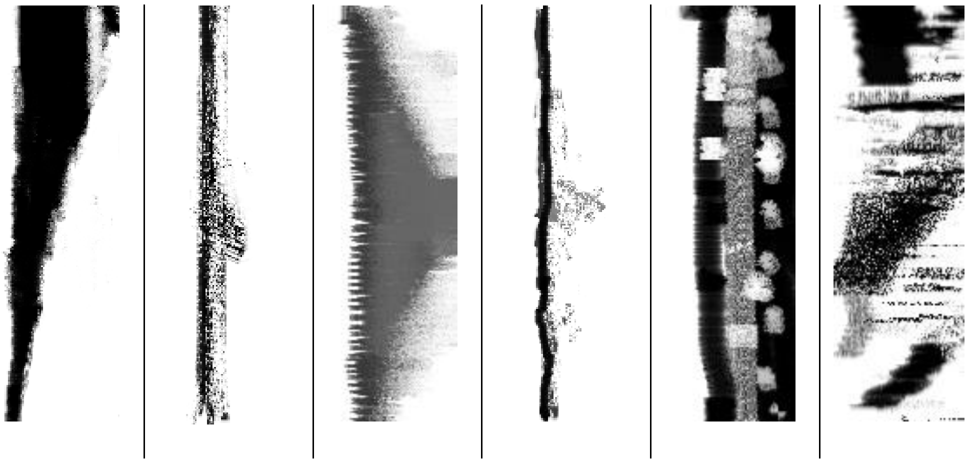


Figure 65. Six sketches drawn by the participants in response to the shuttering animation.

7.4 Conclusions

The evaluation of the present version of Morpheme showed that much functionality has to be implemented for the application to accommodate the type of functional features found in commercial application. However, the performance of Morpheme was satisfactory and participants seemed to recognise the creative potential of the interface. From the analysis of the results, we could distinguish between two types of issues. The first type were issues were related to the user interface. Most of the usability and functionality features that the participants noted can relatively easy be addressed with further implementation of the interface. The second type were issues which were related to the type of sound synthesis (i.e. target based automatic selection synthesis using low and high level descriptions). Some of issues involve the unexpected transition between audio-units that sounded very different, which gave participants the impression of lack of control. As it was discussed at the beginning of Chapter 4 in order to create sounds that are a plausible variations of the original audio used in the corpus a degree of awareness not only of the micro but also of the meso and the macro levels of the sound is required. This is not an issue that can be fixed as easily as the usability issues that were revealed by the evaluation. Through the evaluation of Morpheme many issues were identified, which will form the basis for future development of interface.

8 Conclusions Chapter

This chapter, presents a summary of my research efforts delineating the conclusions drawn from each chapter. Additionally, the contributions and limitations of the present work are discussed, together with the conclusions derived and suggestions leading to future work.

8.1 Summary of the thesis

The primary goal of this research was to investigate the issues related to visual interaction with corpus-based concatenative synthesis methods and to develop a graphical user interface that facilitates the expression of sound synthesis ideas in visual terms, using a set of perceptually meaningful audio-visual feature associations. Chapter 2 reviews previous and related work and identifies a number of issues of concern. Firstly, previous research has emphasised that defining approaches to visual representation of sound is extremely important for analytical, compositional and pedagogical purposes. Secondly, there is need for conceptualisation tools that are capable of bridging the gap in the creative process between the phases of conceptualisation and production. Thirdly, graphical representations used for interaction with sound synthesis for sound design and musical applications should focus more on perceptually relevant sound attributes and less on low level attributes of the given synthesis methods. As digital technologies allow us to make arbitrary associations between any type of data (i.e. modal or amodal). The question of how we can approach the design of such associations in objective terms poses a challenge to a number of disciplines including computer music, information display, and sensory substitution amongst others. Finally, there is need for empirically supported design framework for cross-modal representation and interaction that respect users' intuitions and expectations by maintaining cross-modal correspondences which are based on perceptual knowledge, when possible.

In many respects, issues that were identified in Chapter 2 formed the basis of the present research, and led to a more in-depth investigation of the similarities and differences between audition and vision, presented in chapter 3. In this chapter, works in audio-visual correspondence and metaphoric congruence were reviewed in order to:

- Identify which audio-visual feature dimensions are the best correlates and inform the Morpheme's audio-visual mapping.
- Understand better the mechanisms that mediate cross-modal integration and non-linguistic congruence effect.
- Emphasise the importance of enacting and applying prior perceptual in the context of multimodal interaction design as well as to demarcate the limitations.

The above reviews resulted in the development of two empirically validated audio-visual mappings and informed the experimental design of the studies which were conducted to validate the mappings. In many respects, the knowledge resulted from the literature review formed the basis for the design of Morpheme, a novel interface for the interaction with corpus-based concatenative sound synthesis through the act of sketching on a digital canvas. The first version of Morpheme which was described in section 0 was a proof of concept prototype, also see (Tsiros, Leplâtre, & Smyth, 2012). Hence, further development of the first version of Morpheme was required. These led to: (i) redesign of the user interface, (ii) the development of a number of algorithms which aimed at improving the level of user control over the navigation of the audio corpora, see sections 4.6.4, 4.6.6 and 4.6.7), and (iii) a number of changes aimed to decrease the computational resources required to run the application. For more information please see thesis section 4.5 and Tsiros (2013).

The third stage of this research involved the evaluation of the Audio-Visual (A/V) mappings and of Morpheme's user interface. The evaluation comprised two controlled experiments, an online study and a user study. Participants in the first two experiments and the online study were asked to perform a series of image-sound similarity assessments and discrimination tests. The effect of four main factors were measured in these two experiments: the A/V mappings, the A/V associations, the harmonicity of source audio and the subjects' level of music/audio training. The findings show that the degree of perceived A/V congruency varies significantly between the individual A/V associations, but less as a result of either the audio corpus or the A/V mapping. However the data also show that the influence of the harmonicity of the audio corpus when making a similarity judgement is non-linear and relative to the strength/dominance of the A/V association being tested. Strongly correlated dimensions such as size-loudness or vertical position-pitch are affected less by the harmonicity of the audio corpus in comparison to weaker correlated dimensions (e.g. texture granularity-sound dissonance). Additionally, the results suggest that sound/musical training had no significant effect on: the perceived similarity or preferred A/V association and the discrimination ability of the subjects. My conclusion is that the audio-visual correspondences tested in the experiment (particularly the strongly correlated ones) are a product of either psychophysical, structural similarity or linked through other cultural conventions/factors, but not specific conventions or learned relationships that are related to the acquisition of musical/auditory skills. Our conclusion is that, in the context of corpus-based synthesis, the salience and efficacy of the cross-modal associations involved in a multidimensional mapping are to a degree dependent on the typological features of the source audio which the corpus consists

The third study consisted of an evaluation of *Morpheme's* user interface where participants were asked to use the system to design a sound for a given video footage. The usability of the system and participants' appreciation of the system was found to be satisfactory while a number of usability issues were revealed, which will form the basis for future development of the prototype.

8.2 Contributions

The contributions of the present research can be summarised as follows:

1. Morpheme, an interface for high level control of the retrieval and signal processing algorithms of corpus-based concatenative sound synthesis was developed. Morpheme is (to my knowledge) the first attempt ever made to use sketching as a model of interaction for concatenative synthesis.
2. This thesis adopted two methods (originally used in experimental psychology) for empirically driven assessment of the comprehensibility and effectiveness of audio-visual mappings for visually driven sound synthesis. This methods have not previously been used to in the context of designing multimodal interfaces for musical interaction. These methods could be applied to a wide range of contexts in order to inform the design of cognitively useful multi-modal interfaces and representation and rendering of multimodal data.
3. Moreover this research contributes to the broader understanding of multimodal perception by gathering empirical evidence about the correspondence between auditory and visual feature sets and by investigating how the harmonicity of the sound used to render audio-visual association affect the perceived congruency and effectiveness of the mapping.

8.3 Limitations

One of the limitations of the present empirical investigation is the limited number of audio-visual feature dimensions tested. This is a consequence of having spent a significant part of this research developing Morpheme and the fact that the empirical work could not be carried out before the implementation of the system. However, the fact that the A/V associations tested were based on previous empirical findings, reduced the need for testing a large number of alternatives A/V pairs.

A second limitation relates to the visual feature extraction particularly for the high level features of texture such as texture granularity and color variance. The descriptors obtained by the statistical analysis of these properties provide only a low level description of the otherwise high-level attributes texture. Hence, the reader should be aware that the descriptive ability of these two visual descriptors is limited. Moreover, it should be noted that descriptive statistics are used at both ends to describe the qualities of the audio-units which the corpus consists of and for visual feature extraction. However, high level features such as auditory timbre, and visual texture are hard to fully describe in statistical terms. Digitally extracted features provide a relevant but however reductionist interpretation of what it is being perceived when listening to music or sound at an experiential level. As Hoffman & Cook, (2006) suggest, this is evident from the continually improved features for music information retrieval, and it could also be argued that this is true for texture analysis and visual feature extraction for computer vision. Therefore the conclusions that can be drawn from the present empirical studies are valid within the limits imposed by these factors. The use of pattern recognition algorithms could yield better results for the detection of typological features of texture and it will certainly be considered in future development of the interface.

8.4 Future Work

The next step in this research will be to incorporate the finding from the evaluation of Morpheme to improve the usability of the interface as well as to use the findings from the experiments to inform the current mappings. A number of follow-up experiments, altering the data gathering methods and the materials are required in order to assess to which extent our findings are generalizable. More specifically it will be necessary to test alternative combinations for the audio–visual feature associations for the feature associations that were not highly rated in experiment one. More studies will be conducted to evaluate the distance mapping between target and selected feature vectors to the synthesis parameters. Future studies will examine whether there are interactions between participants' similarity ratings and the statistical correlation of the audio-visual feature vectors of each A/V stimuli across the feature dimensions of the mappings. Furthermore, in the experiments presented in Chapters 5 and 6 only two methods were explored, non-speeded pairwise similarity judgments and multiple item discrimination; other methods for obtaining similarity data are available such as free sorting and spatial arrangement methods, speeded classification and discrimination tasks. I believe that these techniques could have a wide range of applications in the evaluation of multimodal interfaces for computer based sound synthesis and musical interaction design. Currently I am designing experiments to test the performance of Morpheme and gather evidence of cross-modal correspondence in multidimensional dynamic environments.

Moreover, further elaboration of visual extraction algorithms in conjunction with empirical testing will be necessary, particularly for testing the estimation of visual attributes of texture such as repetitiveness, granularity and coarseness. Currently I am trying to port Haralick’s

Java texture feature analysis libraries[8] to Max MSP for analysis based on grey level co-occurrence matrix analysis. In addition, I am working on an adaptation of Morpheme named *AniMorph*. Some examples of the animation driven can be found in the following page[9].

Furthermore, a non real-time version of Morpheme will be implemented in order to synthesise sounds from images by layering. Currently, the audio-units in the corpus are segmented in the temporal domain only. However, preprocessing can be performed on the audio before temporal segmentation in order to segment sounds in the frequency domain as well. This can be achieved using a bandpass cochlear filter to decompose the sound waveform into acoustic frequency bands. After temporal segmentation is performed, the audio-units in the corpus will represent only a tiny portion of the original signal. As descriptive statistics are used to represent and retrieve audio from the corpus, when both frequency and temporal segmentation have been performed in the source audio the representational accuracy of the audio features that will be extracted in order to retrieve audio will be significantly higher, as there is a trade off when using summary statistics between the sample size and the descriptive accuracy of the estimated statistical properties of the audio units. However, this means that the audio units will have to be layered not only in the temporal domain but also in the frequency domain, which translates into a much larger number of queries in order to create the same duration of audio as when only temporal segmentation is performed on the audio signal. Hence, the system that uses this method has been envisaged as a non real-time. The aim of experimenting with frequency segmentation is to investigate the potential of resynthesis of natural sounding auditory textures.

Finally, the Morpheme can easily be re-appropriated to sonify visual features as a sensory substitution system. The main idea is to develop a headset that will consist of a set of cameras and a pair of headphone which will be connected to a raspberry PI that will run the software. The system will only sonify changes in the visual scene in an attempt to create a simple mechanism that emulates the cognitive mechanisms such as selective attentions for regulation of cognitive load. For example, it could be argued that when we are looking at a visual scene, if the scene is relatively static (i.e. not significant changes occur) humans tend to pay attention to the scene only for a limited amount of time. Then, our attention tends to shift from the visual scene and start to wander, engaging with other thoughts and problems, a process guided by the default mode network[10], (Buckner, Andrews-Hanna, & Schacter, 2008; Raichle & Snyder, 2007; Raichle *et al.*, 2001). If a significant change has occurred in the visual scene, the attention shifts back from the

[8] Haralick texture analysis library

[9] https://avrenderstudy.wordpress.com/

[10] The default mode network is an interconnected system of the brain that is activated when individuals engage in internal tasks such as daydreaming, thinking about the future, retrieving memories and examines different points of view.

wandering mode to attend to the new event. Similarly, when humans receive a high load of information from a ‘cluttered’ visual or auditory scene there is only a small portion of the information which we receive from the auditory and visual signal that can be attended/processed at any given moment, (Lavie, Hirst, de Fockert, & Viding, 2004). In fact, selective attention filter out irrelevant information in order to reduce the cognitive load while guiding perception to the information that is relevant to the organism in a given context. This interplay between attention and perception could be modelled in sensory substitution system to filter out irrelevant information. A simple way to achieve that is to focus on sonifying only the changes in the visual scene while ignoring anything that stays constant. This will help to decrease the amount of information that is sonified by the system leaving space for the auditory information from the environment. Finally, other, more elaborate models of selective attention and switching will be explored.

9 Appendix A-F

The appendices of thesis are available from the research repository, please see http://researchrepository.napier.ac.uk/9654/. The appendix lists the:

Appendix A - Morpheme standalone: This folder contains two standalone applications using the chromatic and achromatic mapping. To run the application you will need to install Max MSP/Jitter 5 Runtime[11] and the FTM 2.6.0 library[12].

Appendix B - Morpheme Max MSP patch: This folder contains two standalone applications using the chromatic and achromatic mapping. To run the application you will need to install Max MSP/Jitter 5 full application which can be downloaded from, see footnote 11. Finally you will need to install the FTM library 2.6.0 beta and place the CataRT folder in a folder that is included in the path of Max/MSP

Appendix C - Experiment 1: This folder contains the stimuli used in this experiment. The data that were gathered from the study. All the statistical analysis tables. The instructions used to explain the experimental task. The questioner that was used to gather the participants' demographic information and assess their expertise.

Appendix D - Experiment 2: This folder contains the stimuli used in this experiment. The data that were gathered from the study. All the statistical analysis tables. The instructions used to explain the experimental task. The participants' comments.

Appendix E - Experiment 3: This folder contains the stimuli used in this experiment. The sketches and the sounds the participants created in response to the task. The data that were gathered from the questioner which was filled after the experimental tasks. All the statistical analysis tables. The instructions used to explain the experimental task. The participants' comments.

Appendix F – Electronic version of the thesis manuscript.

Appendix G - Video instruction: Videos explaining how to use the software (i.e. Morpheme, Appendix A and B) and a video explaining the system in the DVD will be available from.

[11] https://cycling74.com/downloads/older/
[12] http://ftm.ircam.fr/index.php/Download

10 References

Adeli, M., Rouat, J., & Molotchnikoff, S. (2014). Audiovisual correspondence between musical timbre and visual shapes. *Frontiers in Human Neuroscience*, *8*(May), 352.

Amelynck, D., Maes, P., Martens, J., & Leman, M. (2014). Expressive Body Movement Responses to Music Are Coherent, Consistent, and Low Dimensional. *IEEE Transactions on Cybernetics*, *44*(12), 2288–2301.

Arfib, D., Couturier, J. M., & Kessous, L. (2002). Gestural Strategies for Specific Filtering Processes. In *proc. of DAFX - Digital Audio Effects* (pp. 1–6).

Armontrout, J. a, Schutz, M., & Kubovy, M. (2009). Visual determinants of a cross-modal illusion. *Attention, Perception & Psychophysics*, *71*(7), 1618–27.

Athanasopoulos, G., & Moran, N. (2013). Cross-cultural representations of musical shape. *Empirical Musicology Review*, *8*(3), 185–199.

Barry, B. (1977). Contemporary Music as Represented in Stockhausen's Plus-Minus. *College Music Symposium*, *17*(2), 42–46.

Barsalou, L. W. (1993). Challenging assumptions about concepts. *Cognitive Development*, *8*(2), 169–180.

Barsalou, L. W. (1999). Perceptual symbol systems. *The Behavioral and Brain Sciences*, *22*(4), 577–609; discussion 610–60.

Barsalou, L. W., Kyle Simmons, W., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*(2), 84–91.

Baveli, M.-D., & Georgaki, A. (1994). Towards a decodification of the graphical scores of Anestis Logothetis (1921-1994) . The graphical space of Odysee (1963). In *Sound and Music Computing Conference* (pp. 39–43).

Beasley, T. M., & Schumacker, R. E. (1995). Multiple Regression Approach to Analyzing Contingency Tables: Post Hoc and Planned Comparison Procedures. *Journal of Experimental Education*.

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics*, *57*(8), 1151–1162.

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology*, *87*, 241–247.

Bertelson, P., & de Gelder, B. (2004). The Psychology of Multimodal Perception. In C. Spence & J. Driver (Eds.), *Crossmodal space and Crossmodal Attention* (pp. 141 – 178). Oxford University Press.

Berthaut, F., Desainte-catherine, M., & Hachet, M. (2010). Combining audiovisual mappings for 3d musical interaction. *International Computer Music Conference*.

Blackburn, M. (2009). Composing from spectromorphological vocabulary: proposed application, pedagogy and metadata. *Electroacoustic Music Studies*.

Blackburn, M. (2011). The Visual Sound-Shapes of Spectromorphology: an illustrative guide to composition. *Organised Sound*, *16*(01), 5–13.

Bogaards, N. (2005). Analysis-assisted sound processing with audiosculpt. *Proc. of DAFX2005 - Digital Audio Effects*, 20–23.

Bond, B., & Stevens, S. S. (1969). Cross-modality matching of brightness to loudness by 5-year-olds. *Perception & Psychophysics*, *6*(6), 337–339.

Brandwein, A. B., Foxe, J. J., Russo, N. N., Altschuler, T. S., Gomes, H., & Molholm, S. (2011). The development of audiovisual multisensory integration across childhood and early adolescence: a high-density electrical mapping study. *Cerebral Cortex (New York, N.Y. : 1991)*, *21*(5), 1042–55.

Bregman, A. S. (1994). *Auditory scene analysis: the perceptual organization of sound.* Cambridge, MA: MIT Press.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The Brain's Default Network: Anatomy, Function, and Relevance to Disease. *Annals of the New York Academy of Sciences*, *1*, 1 – 38.

Caramiaux, B., Bevilacqua, F., & Schnell, N. (2010a). Study on Gesture-Sound Similarity. *3rd Music and Gesture Conference*, 1–2.

Caramiaux, B., Bevilacqua, F., & Schnell, N. (2010b). Towards a gesture-sound cross-modal analysis. In I. Kopp, S., Wachsmuth (Ed.), *Gesture in Embodied Communication and Human-Computer Interaction* (pp. 158 – 170). Springer, Heidelberg.

Caramiaux, B., Francoise, J., Bevilacqua, F., & Schnell., N. (2014). Mapping Through Listening. In *Computer Music Journal* (Vol. 38, pp. 1–30).

Chiou, R., & Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception*, *41*, 339 – 353.

Christensen, T. (2002). *The Cambridge history of western music theory*. Cambridge University Press.

Climent, R. (2001). Applications of Typomorphology in Acute; Scoring the Ideal and its Mirror. In *Sound and Music Computing Conference*.

Comajuncosas, J. (2011). Nuvolet: 3d gesture-driven collaborative audio mosaicing. *New Interfaces for Musical Expression (NIME)*, (June), 252–255.

Couprie, P. (2004). Graphical representation: an analytical and publication tool for electroacoustic music. *Organised Sound*, *9*(01), 109–113.

Dubnov, S., & Bar-Joseph, Z. (2002). Synthesizing sound textures through wavelet tree learning. *Computer Graphics and Applications, IEEE*, (August), 38–48.

Dutoit, T. (2008). Corpus-based speech synthesis. In Y. Benesty, Jacob, Sondhi, M. M., Huang (Ed.), *Springer Handbook of Speech Processing* (pp. 437–455).

Edwards, A., Challis, B. P., Hankinson, J. C. K. & Pirie, F. L. (2000). Development of a Standard Test of Musical Ability for Participants in Auditory Interface Testing. In *Internation Conference on Auditory Display*. Atlanta.

Eitan, Z. (2013). How pitch and loudness shape musical space and motion: new findings and persisting questions. In S.-L. Tan, A. J. Cohen, S. D. Lipscomb, & R. A. Kendall

(Eds.), *The Psychology of Music in Multimedia* (pp. 161–187). Oxford University Press.

Eitan, Z., & Granot, R. Y. (2011). Listeners' images of motion and the interaction of musical parameters. In *10th Conference of the Society for Music Perception and Cognition*.

Eitan, Z., Schupak, A., Gotler, A., & Marks, L. E. (2011). Lower pitch is larger, yet falling pitches shrink: Interaction of pitch change and size change in speeded discrimination. *Proceedings of Fechner Day*.

Eitan, Z., Schupak, A., & Marks, L. E. (2008). Louder is higher : Cross-modal interaction of loudness change and vertical motion in speeded classification. In *Proceedings of the 10th International Conference on Music Perception and Cognition (ICMP10)* (p. 2008).

Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: cross-domain mappings of auditory pitch in a musical context. *Cognition*, *114*(3), 405–22.

Emerson, S. (1986). The Relationship of the Language to Materials. In S. Emmerson (Ed.), *The Language of Electroacoustic Music* (pp. 18–39). MacMillan Press.

Emmerson, S. (2007). *Living electronic music*. Ashgate Publishing.

Evans, K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*, 1–12.

Farbood, M., Kaufman, H., Line, H., & Jennings, K. (2007). Composing with Hyperscore: An Intuitive Interface for Visualizing Musical Structure. In *International Computer Music Conference* (pp. 111–117).

Fauconnier, G. (1997). *Mapping in thought and Language*. Cambridge University Press.

Fels, S. (1994). *Glove-TalkII: Mapping Hand Gestures to Speech Using Neural Networks An Approach to Building Adaptive Interfaces*. University of Toronto.

Fox, C. (2000). Stockhausen’s plus minus, More or Less: Written in Sand. *The Musical Times*, *141*(1871), 16–24.

Gamboa, S. (2007). Hume on Resemblance, Relevance, and Representation. *Hume Studies*, *33*(1), 21–40.

Garcia-perez, M. a., & Nunez-anton, V. (2003). Cellwise Residual Analysis in Two-Way Contingency Tables. *Educational and Psychological Measurement*, *63*(5), 825–839.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45–56.

Giannakis, K. (2001). *Sound Mosaics*. Middlesex University.

Giannakis, K. (2006). A comparative evaluation of auditory-visual mappings for sound visualisation. *Organised Sound*, *11*(03), 297.

Giannakis, K., & Smith, M. (1993). Imaging Soundscapes : Identifying Cognitive Associations between Auditory and Visual Dimensions.

Glenberg, a M. (1997). What memory is for. *The Behavioral and Brain Sciences*, *20*(1), 1–19; discussion 19–55.

Godøy, R., Haga, E., & Jensenius, A. (2006). Exploring music-related gestures by sound-tracing: A preliminary study. In *2nd ConGAS International Symposiumon Gesture Interfaces for Multimedia Systems*.

Godøy, R. I. (2006). Gestural-Sonorous Objects: embodied extensions of Schaeffer’s conceptual apparatus. *Organised Sound*, *11*(02), 149.

Godøy, R. I., Haga, E., & Jensenius, A. R. (2006). Playing “ Air Instruments ”: Mimicry of Sound-producing Gestures by Novices and Experts. *Gesture in HumanComputer Interaction and Simulation*, *3881*(6801), 256–267.

Goldschmidt, G. (1991). The dialectics of sketching. *Creativity Research Journal*, *4*(2), 123–143.

Goldstone, R. L. (1994). The role of similarity in categorization: providing a groundwork. *Cognition*, *52*(2), 125–57.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585–612.

Goodman, N. (1972). Seven strictures on similarity. In *Probles and Projects* (pp. 437–446). Indianapolis/ New York: Bobbs-Merrill.

Granot, R. Y., & Eitan, Z. (2011). Musical tension and the interaction of dynamic auditory parameters. *Music Perception*, *28*, 219–245.

Grice, H. (1962). Some remarks about the senses. In R. J. Butler (Ed.), *Analytical Philosophy, First series*. Oxford University Press.

Hakinson, J. C. K., Challis, B. P., & Edwards, A. (1999). *MAT: A Tool for Measuring Musical Ability*.

Hampe, B. (2005). Image schemas in Cognitive Linguistics : Introduction. In R. Dirven, R. W. Langacker, & J. R. Taylor (Eds.), *From Perception to Meaning: Image Schemas in Cognitive Linguistics* (pp. 1–15). Walter de Gruyter press.

Handel, S. (2006). *Perceptual Coherence: Hearing and Seeing*. Oxford University Press.

Hidaka, S., Teramoto, W., Keetels, M., & Vroomen, J. (2013). Effect of pitch-space correspondence on sound-induced visual motion perception. *Experimental Brain Research*, *231*(1), 117–26.

Hoffman, M., & Cook, P. (2006). Feature-based synthesis: Mapping acoustic and perceptual features onto synthesis parameters. In *Proceedings of the 2006 International Computer Music Conference* (pp. 536–539).

Hubbard, T. L., American, T., & Summer, N. (2007). Synesthesia-like mappings of lightness , pitch , and melodic interval. *The American Journal of Psychology*, *109*(2), 219–238.

Hunt, a., & Kirk, R. (1999). Radical user interfaces for real-time control. *Proceedings 25th EUROMICRO Conference. Informatics: Theory and Practice for the New Millennium*, 6–12 vol.2.

Hunt, A., Wanderley, M., & Kirk, R. (2000). Towards a Model for Instrumental Mapping in Expert Musical Interaction University of York Analysis-Synthesis Team. In *International Computer Music Conference*.

Janer, J., & Boer, M. De. (2008). Extending voice-driven synthesis to audio mosaicing. In *5th Sound and Music Computing Conference*.

Jehan, T. (2005). *Creating Music by Listening by*. PhD Thesis, MIT.

Karkoschka, E. (1973). Notation in New Music: A Critical Guide to Interpretation and Realisation. *Journal of Music Theory*, *17*(1), 168–171.

Keeley, B. L. (2011). Making Sense of the Senses: Individuating Modalities in Humans and Other Animals. In F. Macpherson (Ed.), *The Senses Classic and Contemporary Philosophical Perspectives* (pp. 220 – 240). Oxford University Press.

Klingbeil, M. (2005). Software for spectral analysis, editing, and synthesis. *Proceedings of the International Computer Music Conference*.

Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.

Kohn, D., & Eitan, Z. (2009). Musical Parameters and Children ' s Movement Responses. In *7th Triennial Conference of the European Society for the Cognitive Sciences of Music* (Vol. 42, pp. 233–241).

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring Dissimilarity Structure from Multiple Item Arrangements. *Frontiers in Psychology*, *3*(July), 245.

Kubovy, M., & Schutz, M. (2010). Audio-visual objects. *Review of Philosophy and Psychology*, *7*, 49–72.

Kubovy, M., & Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition*, *80*(1-2), 97–126.

Küssner, M. B. (2014). *Shape, drawing and gesture: cross-modal mappings of sound and music*. King's College London.

Kussner, M. B., & Leech-Wilkinson, D. (2013). Investigating the influence of musical training on cross-modal correspondences and sensorimotor skills in a real-time drawing paradigm. *Psychology of Music*, *42*(3), 448–469.

Küssner, M. B., Tidhar, D., Prior, H. M., & Leech-Wilkinson, D. (2014). Musicians are more consistent: Gestural cross-modal mappings of pitch, loudness and tempo in real-time. *Frontiers in Psychology*, *5*(July), 789.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. The university of Chicago press.

Lakoff, G., & Johnson, M. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. The university of Chicago press.

Landy, M. S., Banks, M. S., & Knill, D. C. (2012). Ideal-Observer Models of Cue Integration. In and M. S. L. Julia Trommershäuser, Konrad Kording (Ed.), *Sensory Cue Integration* (pp. 1–35). Oxford Scholarship Online.

Larkey, L. B., & Markman, A. B. (2005). Processes of similarity judgment. *Cognitive Science*, *29*(6), 1061–76.

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology. General*, *133*(3), 339–54.

Leman, M. (2008). *Embodied Music Cognition and Mediation Technology*. MIT press.

Leslie, G., Schwarz, D., Warusfel, O., Bevilacqua, F., Zamborlin, B., Jodlowski, P., & Schnell, N. (2010). Grainstick: a collaborative, interactive sound installation. In *Proceedings of the International Computer Music Conference (ICMC)* (p. 4).

Levin, G. (2000). *Painterly interfaces for audiovisual performance*. MIT.

Levin, G. (2005). A Personal Chronology of Audiovisual Systems Research. In *NIME '05 Proceedings of the 2005 conference on New interfaces for musical expression* (pp. 2–3).

Levin, G. (2006). The table is the score: An augmented-reality interface for real-time, tangible, spectrographic performance. In *proc. of DAFX - Digital Audio Effects* (pp. 151–154).

Lipscomb, S., & Kim, E. (2004). Perceived match between visual parameters and auditory correlates: An experimental multimedia investigation. In *International Conference on Music Perception and Cognition* (pp. 72–75).

Lohner, H. (1986). The UPIC System: A User's Report. *Computer Music Journal*, *10*(4), 42–49.

Ludwig, V. U., Adachi, I., & Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (Pan troglodytes) and humans. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(51), 20661–5.

Marks, L. E. (1974). On the Associations of Light and Sound. The Mediation of Brightness, Pitch, And Loudness. *American Journal Of Psychology*, *87*(1-2), 173–188.

Marks, L. E. (1983). On associations of light and sound: The mediation of brightness. *Journal of Psychology*, *87*, 173–188.

Marks, L. E. (1989). On Cross-Modal Similarity : The Perceptual Structure of Pitch , Loudness , and Brightness. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 586–602.

Marks, L. E. (2004). Cross-modal interactions in speeded classification. In Gemma Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 85 – 106). MIT press.

Marks, L. E., & Eitan, Z. (2012). Garner's paradigm and audiovisual correspondence in dynamic stimuli: Pitch and vertical direction. *Seeing and Perceiving*, *25*, 70–70.

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, *28*, 903–923.

Martino, G., & Marks, L. E. (2000). Cross-modal interaction between vision and touch: the role of synesthetic correspondence. *Perception*, *29*(6), 745–754.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, *71*(5), 926–40.

Medin, D., Goldstone, R., & Gentner, D. (1993). Respects for similarity. *Psychological Review*.

Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology and Human Perception and Performance*, *15*, 69 –79.

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology:General*, *116*, 323 – 336.

Mollon, J. (1995). Seeing Colour. In *Colour: art & science* (pp. 127–151). Cambridge University Press.

Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(2), 133–136.

Mulder, A., Fels, S., & Mase, K. (1997). Empty-handed Gesture Analysis in Max/FTS. In *Kansei, The Technology of Emotion. Proceedings of the AIMI International Workshop* (pp. 87–91).

Navab, N., Nort, D. Van, & Wei, S. (2014). A Material Computation Perspective on Audio Mosaicing and Gestural Conditioning. In *New Interfaces for Musical Expression (NIME)*.

Neilson, I., & Lee, J. (1994). Conversations with graphics: implications for the design of natural language/graphics interfaces. *International Journal of Human Computer Studies*, *4*, 509–541.

Nelson, P. (1997). The UPIC system as an instrument of learning. *Organised Sound*, *2*(01).

Norman, D. A. (1999). Affordances, Conventions and Design. *Interactions*, *6*(3), 38–42.

Nuckolls, J. B. (2010). The Case for Sound Symbolism. *Annual Review of Anthropology, 28*(1999), 225–252.

Parise, C., & Spence, C. (2013). Audiovisual cross-modal correspondences in the general population. *The Oxford Handbook of Synesthesia*, 790–815.

Parise, C. V, & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research*, *214*(3), 373–80.

Patching, G. R., & Quinlan, P. T. (2002). Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology and Human Perception and Performance*, *28*, 755 – 775.

Patton, K. (2007). Morphological notation for interactive electroacoustic music. *Organised Sound*, *12*(02), 123.

Raichle, M. E., MacLeod, A., Snyder, A., Powers, W., Gusnard, D., & Shulman, G. (2001). Inaugural Article: A default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(2).

Raichle, M. E., & Snyder, A. Z. (2007). A default mode of brain function: A brief history of an evolving idea. *NeuroImage*, *37*(4).

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 11800–6.

Ren, Z., Yeh, H., Klatzky, R., & Lin, M. C. (2013). Auditory perception of geometry-invariant material properties. *IEEE Transactions on Visualization and Computer Graphics*, *19*(4), 557–66.

Ren, Z., Yeh, H., & Lin, M. (2013). Example-guided physically based modal sound synthesis. *ACM Transactions on Graphics (TOG)*, *32*(1).

Roads, C. (1978). Automated granular synthesis of sound. *Computer Music Journal*, *2*(2), 61–62.

Roads, C. (1988). Introduction to Granular Synthesis. *Computer Music Journal*, *12*(2), 11–13.

Roads, C. (1996). *The Computer Music Tutorial*. *Computer Music Journal* (Vol. 12). MIT press.

Roads, C. (2001). *Microsound*. MIT press.

Rovan, J. B., Wanderley, M. M., Dubnov, S., & Depalle, P. (1997). Instrumental Gestural Mapping Strategies as Expressivity Determinants in Computer Music Performance. In *Kansei-The Technology of Emotion Workshop. Proceedings of the AIMI International Workshop* (pp. 68–73).

Rusconi, E., Kwan, B., Giordano, B. L., Umiltà, C., & Butterworth, B. (2006). Spatial representation of pitch height: the SMARC effect. *Cognition*, *99*(2), 113–29.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology*, *12*, 225–239.

Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(6), 1791–810.

Schwarz, D. (2004). *Data-driven concatenative sound synthesis*. *Phd. Thesis*. Universite Paris 6.

Schwarz, D. (2005). Current Research in Concatenative Sound Synthesis. In *International Computer Music Conference* (pp. 9–12).

Schwarz, D. (2012). The sound space as musical instrument: Playing corpus-based concatenative synthesis. In *New Interfaces for Musical Expression (NIME)*.

Schwarz, D., & Caramiaux, B. (2014). Interactive Sound Texture Synthesis through Semi-Automatic User Annotations. *Lecture Notes in Computer Science.*, 1–21.

Schwarz, D., & Hackbarth, B. (2012). Navigating variation: composing for audio mosaicing. In *International Computer Music Conference* (pp. 1–4).

Schwarz, D., & Schnell, N. (2010). Descriptor-based sound texture sampling. In *Sound and Music Computing Conference*.

Shams, L., & Beierholm, U. (2012). Humans' Multisensory Perception, from Integration to Segregation, Follows Bayesian Inference. In and M. S. L. Julia Trommershäuser, Konrad Kording (Ed.), *Sensory Cue Integration* (pp. 1–16). Oxford Scholarship Online.

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411–7.

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 505–509.

Smalley, D. (1986). Spectromorphology and the Structuring Process. In S. Emmerson (Ed.), *The Language of Electroacoustic Music* (pp. 61–93). MacMillan Press.

Smalley, D. (1997). Spectromorphology: explaining sound-shapes. *Organised Sound*, *2*, 107–126.

Smith, L. B., & Sera, M. D. (1992). A Developmental Analysis of the Polar Structure of Dimensions. *Cognitive Psychology*, *142*, 99–142.

Smith, L., & Heise, D. (1992). Perceptual similarity and conceptual structure. *Advances in Psychology*.

Somers, E. (1998). A Pedagogy of Creative Thinking based on Sonification of Visual Structures and Visualization of Aural Structures. In *Internation Conference on Auditory Display*.

Somers, E. (2000). Vocabulary, Abstract Sound Objects to Expand the of Sound Design for Visual and Theatrical Media. In *Internation Conference on Auditory Display*.

Sorabji, R. (2011). Aristotle on Demarcating the Five Senses. In F. Macpherson (Ed.), *The Senses: Classic and Contemporary Philosophical Perspectives* (pp. 64 – 82). Oxford University Press.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology*, *28*(2), 61–70.

Spence, C. (2011). Crossmodal correspondences: a tutorial review. *Attention, Perception & Psychophysics*, *73*(4), 971–95.

Squibbs, R. (1996). Images of Sound in Xenakis's Mycenae-Alpha. *Les Cahiers Du GREYC4*, (1978).

Stevens, J. C., & Marks, L. E. (1965). Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Sciences*, *54*, 407–411.

Stevens, S. A., Doug, L. J., & Marschner, S. (2012). Motion-driven concatenative synthesis of cloth sounds. *ACM Transactions on Graphics*, *31*(4), 1–10.

Strobl, G., Eckel, G., & Rocchesso, D. (2006). Sound Texture Modelling : A Sound: SURVEY. In *Sound and Music Computing Conference* (pp. 3–7).

Talmy, L. (2000). Fictive motion in language and "ception." In *Towards Cognitive Semantics Vol. 1* (Vol. 1, pp. 99– 175). MIT press.

Thiebaut, J., Bello, J., & Schwarz, D. (2007). How musical are images? From sound representation to image sonification: an eco systemic approach. *Proc. of ICMC*.

Thoresen, L. (2010). Form-Building Patterns and Metaphorical Meaning. *Organised Sound*, *15*(02), 82–95.

Thoresen, L., & Hedman, A. (2007). Spectromorphological analysis of sound objects: an adaptation of Pierre Schaeffer's typomorphology. *Organised Sound*, *12*(02), 129.

Truax, B. (1990). Composing with Real-time Granular Sound. In *Perspectives of New Music* (pp. 120–134). USA: Hamilton Printing Company.

Tsiros, A. (2013). A Multidimensional Sketching Interface for Corpus Based Concatenative Synthesis. In *International Conference on Auditory Display* (pp. 279–282).

Tsiros, A. (2014). Evaluating the Perceived Similarity Between Audio-Visual Features Using Corpus-Based Concatenative Synthesis. In *New Interfaces for Musical Expression (NIME)*.

Tsiros, A., Leplâtre, G., & Smyth, M. (2012). Sketching Concatenative Synthesis: Searching For Audiovisual Isomorphism In Reduced Modes. In *International conference of sound and music computing*.

Vickery, L. (2014). The Limitations of Representing Sound and Notation on Screen. *Organised Sound*, *19*(03), 215–227.

Vinet, H. (1999). Concepts d ' interfaces graphiques pour la production musicale et sonore. In s H. V. et F. Delalande (Ed.), *Interfaces homme-machine et création musicale* (pp. 97 –121).

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, *21*(1), 21–5.

Walker, R. (1987). The effects of culture , environment , age , and musical training on choices of visual, *42*(5), 491–502.

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The Musical Ear Test, a new reliable test for measuring musical competence. *Learning and Individual Differences*, *20*(3), 188–196.

Walsh, V. (2003). A theory of magnitude: Common cortical metrices of time, space and quality. *Trends in Cognitive Sciences*, *7*, 483–488.

Wanderley, M. (1998). Escher-modeling and performing composed instruments in real-time. *Systems, Man, and Cybernetics, 1080-1084.*

Wanderley, M. M., & Depalle, P. (2004). Gestural Control of Sound Synthesis. *Proceedings of the IEEE*, *92*(4), 632–644.

Wei, K., & Körding, K. P. (2012). Causal Inference in Sensorimotor Learning and Control. In and M. S. L. Julia Trommershäuser, Konrad Kording (Ed.), *Sensory Cue Integration* (pp. 1–23). Oxford Scholarship Online.

Wishart, T. (1986). Sound Symbols and Landscapes. In S. Emmerson (Ed.), *The Language of Electroacoustic Music* (pp. 41–60). MacMillan Press.

Xenakis, I. (1971). *Formalized Music*. Bloomington: Indiana University Press.

Zaltman, G. (2002). Eliciting Mental Models through Imagery. In A. M. Galaburda, S. T. Kosslyn, & Y. Christen (Eds.), *The languages of the brain.* (pp. 363–375). Harvard University Press.