Evaluating the Perceived Similarity Between Audio-Visual Features Using Corpus-Based Concatenative Synthesis

Augoustinos Tsiros
Centre for Interaction Design
Edinburgh Napier University
10 Colinton Road, EH10 5DT
Edinburgh, United Kingdom
a.tsiros@napier.ac.uk

ABSTRACT
This paper presents the findings of two exploratory studies. In these studies participants performed a series of image-sound association and detection tasks. The aim of the studies was to investigate the perceived similarity and the effectiveness of two multidimensional mappings, each consisting of three audio-visual associations. The purpose of the mappings is to enable visual control of corpus-based concatenative synthesis. The stimuli in the first study was designed to test the perceived similarity of six audio-visual associations, between the two mappings using three corpora resulting in 18 audio-visual stimuli. The corpora differ in terms of two sound characteristics: harmonic content and continuity. Data analysis revealed no significant differences in the participants' responses between the three corpora, or between the two mappings. However significant differences were revealed between the individual audio-visual association pairs. The second study investigates the effects of the mapping and the corpus in the ability of the participants to detect which image out of three similar images was used to synthesise the audio stimuli. Data analysis revealed significant differences in the ability of the participants to detect the correct image depending on which corpus was used. Less significant was the effect of the mapping in the success rate of the participants' responses.

Keywords
Concatenative Synthesis, Audio-visual Correspondence, Multimodal interaction, Embodied interaction

1. INTRODUCTION
Concatenative synthesis methods allow to synthesise sound from pre-recorded audio that have been spliced into small segments known as audio-units. 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 target 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. The input data stream 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 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.

NIME'14, June 30 – July 03, 2014, Goldsmiths, University of London, UK.

Recent research suggests that there are underlying embodied principles based on which cross-modal associations could be made [8], [12], [13], [17], [26]. Although we are still at the early stages of exploration of multimodal perception and cognition, we have adequate evidence to support that cross-modal binding and interactions between the different sensory modalities occur. Moreover we know that humans use common linguistic labels to express in semantic terms perceived properties of stimuli that are received by the different sensory organs [5], [6], [16], [20]. Cross-modal binding and conceptual blending are undoubtedly central and prevalent in perceiving, disambiguating, interpreting and interacting with the physical environment [12], [11], [5]. 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) in terms of time (i.e. synchrony), and (iii) in terms of space (i.e. collocation) [13], [14], [21], [26]. 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) [13], [14]. Plausible common cause is a concept that deserves more consideration, as in our 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 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 modal objects and their attributes.

Research findings show that multimodal associations can occur automatically when we are exposed to sensory stimuli, [1], [6], [8]. 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, [1], [17]. These re-appropriation of past experience is a form of conceptual blending that enable us to interpret and understand phenomena we have not perceived in the past in terms of phenomena that have been previously experienced. Therefore conceptual blending is central to the way humans think and interact and enables the expression of abstract thought [5], [16]. However the underlying principles that mediate cross-modal binding and congruency effects 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 as well as non-linguistic metaphoric congruency and utilise this knowledge to inform digital constructed audio-visual associations for creative sound and music practice.

2. RELATED WORK

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 [4], [9], [10], [11], [18], [23], [24], [25]. These research efforts demonstrate the importance of enacting prior embodied knowledge and deploying this knowledge to represent, organise, and interact with sensory representations and their parameters intuitively, and with minimal training and learning. Previous studies that investigated the perceived correlation between auditory and visual features have shown that there is consistency in the feature correlates which subjects perceive as a good match [3], [7], [15], [19]. This consistency is rather encouraging as it suggests that gathering more empirical evidence could help in the formation of a framework for the association between visual and aural structures. The present study was designed to investigate the perceived similarity of two multidimensional audiovisual mappings. The audio-visual associations in the mappings were based on features associations that were highly rated according to previous research. The present study extends previous research by testing the perceived similarity of audio-visual associations in a multidimensional context and the effects of the audio corpus in the similarity and detection judgments obtained by subjects.

3. THE SYSTEM & THE MAPPINGS

This section presents the system and the mappings used to create the audio stimuli which were used in the present studies. In the context of this research, the inquiry about audio-visual associations and how it could be approached in objective terms began while designing *Morpheme.* M*orpheme* is a multidimensional audio-visual interface that enables interaction with a corpus-based concatenative synthesis [22]. *Morpheme* extracts visual features from a sketch developed by a practitioner and associates these to audio features for the retrieval of audio -units, and the control of the synthesis parameters. This research project considers that it is of paramount importance to achieve an intuitive mapping in order to enable interaction with concatenative synthesis for creative purposes (e.g. sound design, electroacoustic composition). The aim of the interface is to enable synthesis of sound and the expression of compositional intention by providing perceptually meaningful visual description of sound properties.

Two mappings have been developed for associating auditory and visual features. The distinction between the two mappings is that one is achromatic while the other is chromatic. The first mapping is considered achromatic in the sense that the visual features extracted from the sketch are estimated based on volumetric and spatial attributes of the sketch (see table 1). 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 (see table 2). For more detailed information about the visual features extraction please follow the link that is provided in the Appendices section 9.

In the current implementation of *Morpheme* we can distinguish two mapping layers. The first layer consists of a mapping between visual and auditory features for the retrieval of audio units, shown in table 1 and 2. The second layer consists of associations for mapping the distances between target and selected feature vectors to the synthesis parameters. Table 3 shows how the distances between visual and their respective audio features shown in tables 1 and 2

Table 1. Achromatic mapping

Visual Features	**Audio Feature**
Size/Thickness	Loudness
Vertical position	Pitch
Texture granularity	Dissonance
Horizontal length	Duration

Table 2. Chromatic mapping

Visual Features	**Audio Feature**
Opacity & Size	Loudness
Color hue	Spectral centroid
Color variance	Dissonance
Horizontal length	Duration

Table 3. Mapping the distances between audio-visual feature vectors and synthesis parameters

Audio Feature	**Synthesis parameters**
Loudness	Amplitude
Spectral centroid, Pitch	Transposition
Dissonance	Transposition and audio-unit selection standard deviation

are associated with the synthesis parameters. For example, if the feature of size in the visual domain is mapped to loudness, and the target requests a very loud sound but only quiet sounds are stored in the database, then a quiet sound will be retrieved. In this case the distances between the target values of loudness (defined by the visual feature of the size) and the values of loudness of the retrieved audio-units are mapped to control the amplitude parameter of the sound synthesis. The aim is to match the requested feature values as closely as possible.

4. STUDY 1

4.1 Research Questions & Hypothesis

Three primary research questions were investigated. First, will participant responses reveal a consistent preference to any of the two mappings (i.e. chromatic, achromatic)? Based on the findings of previous research, which reveal that pitch to vertical position is one of the strongest audio-visual correlates, we predicted that the achromatic mapping will be preferred. Second, are some audio-visual features associations consistently perceived as a better match? If so which audio-visual associations are the best correlates? Based on previous research, we predicted that some associations will be preferred over others. Third, does the source audio that the corpus consists of have a significant effect on the perceived similarity of the mapping and/or the audio-visual feature associations involved in each mapping? We predicted that the typology of the corpus will have a significant effect on subjects' judgements and similarity ratings because the source audio in the corpus can affect the saliency of the audio-visual feature associations.

4.2 Procedures

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 the apparatus used for viewing the stimuli and for capturing and storing their responses. Before beginning the main experiment, each participant did a warm-up trial consisting of four 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 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. labeled *Not Similar – 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. 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 effects on their responses due to biases.

4.3 Subject

Fifteen non-musician subjects took part in the study. Eight male and seven female. Participants' age ranged from 24 to 56 years old. None of the participants reported having any known auditory and/or visual impairments which could affect the validity of their responses.

4.4 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. All stimuli cited in this paper are available via the URL which can be found in the appendix, please see section 9. Below follows a more detailed description of the source sounds in each corpus that were used to generate the audio stimuli for this experiment.

4.4.1 First 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.

4.4.2 Second 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.

4.4.3 Third 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.

4.5 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 figure 1 below. However complete isolation of feature dimension was not always possible. For example in the right visual stimulus shown in figure 1, 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. The visual stimuli were designed using a third party graphics software.

Figure 1. The three visual stimuli used in the first study for testing the achromatic mapping.

4.6 Results

This section presents the results obtained from the first experiment. Figure 2 and 3 display the results obtained for all the audio-visual stimuli tested. More specifically figure 2 shows the statistics from the subjects' binary responses (i.e. not similar / similar) for each audio-visual stimuli. While figure 3 shows the statistics from the subjects' similarity ratings (i.e. 0 - 100) for each audio-visual stimuli. The Audio-Visual Stimuli Indexes shown along the x axis in the figures 2 and 3 are explained in table 4. Also note that the audio-visual indexes in table 4 are organized in groups of three as each audio-visual association was tested using the three corpora (i.e. Impacts, Strings and Wind). For instance indexes 1, 2 and 3 correspond to the first audiovisual association size-loudness, indexes 4, 5 and 6 correspond to the relationship between vertical position– pitch.

In contrary to our hypothesis, the effect of the corpus on the perceived similarity of the audio-visual association was not strong, nor do subjects' responses vary significant as a result of the mapping. The two factor analysis of variance revealed no significant differences due to the corpus ($F_{(2, 264)} = 1.92$; $p = .1486$); or the mapping ($F_{(1, 264)} = .62$; $p = .4328$), and no significant interactions were revealed between the variables mapping and corpus ($F_{(2, 264)} = 1.44$; $p = .2387$). However highly significant differences were revealed between the different audio-visual associations ($F_{(5, 252)} = 10.87$; $p < .0001$), while no significant interactions between either the

A/V associations and the corpora ($F_{(10, 252)} = 1.67$; $p < .0873$), or between A/V associations and the mapping ($F_{(2, 252)} = 2.03$; $p = .0306$). Figure 3 shows the results obtained from the second question where subjects rated the degree of perceived similarity between the audio-visual pairs. Similar to the results of the binary responses, data analysis of the ratings revealed no significant differences as a result of either the corpus ($F_{(2, 252)} = 2.16$; $p= .1177$), or the two mappings ($F_{(1, 252)} = 1.17$; $p = .2794$), while no significant interactions were revealed between the variables mapping and corpus ($F_{(2, 252)} = .65$; $p = .5218$). Significant differences were revealed between the individual A/V feature associations ($F_{(5, 252)} = 9.64$; $p < .0001$), while no significant effect was observed due to the corpus ($F_{(2, 252)} = 2.55$; $p = .0805$), nor strong interaction between the A/V associations and the corpus ($F_{(10, 252)} = 1.43$; $p = .1686$).

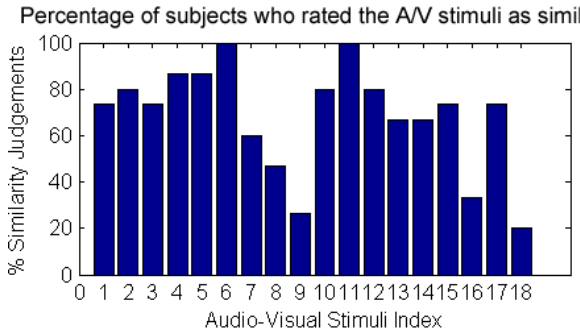

Figure 2. Binary responses for each audiovisual stimuli.

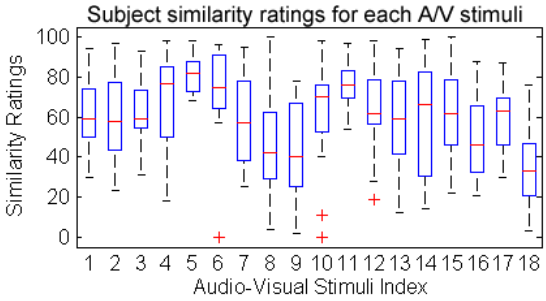

Figure 3. The plot depicts the statistics of the similarity ratings for each one of the audio-visual stimuli used for testing.

Table 4. shows which Audio-Stimuli Index number corresponds to which feature association. Moreover it provides information about the corpora tested.

Achromatic Mapping			
Index	***Audio feature***	***Visual Feature***	***Audio Corpus***
1, 2, 3	Loudness	Size	Impacts, String, Wind
4, 5, 6	Pitch	Vertical position	Impacts, String, Wind
7, 8, 9	Dissonance	Texture granularity	Impacts, String, Wind
Chromatic Mapping			
Index	***Audio feature***	***Visual Feature***	***Audio Corpus***
10, 11, 12	Loudness	Size & Opacity	Impacts, String, Wind
13, 14, 15	Spectral centroid	Hue	Impacts, String, Wind
16, 17, 18	Dissonance	Color variance	Impacts, String, Wind

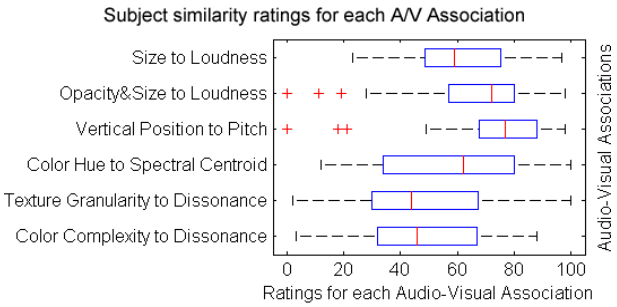

Figure 4. The plot depicts the statistics of the participants' similarity ratings organized by audio-visual association.

5. STUDY 2

5.1 Research Questions & Hypothesis

Two primary research questions were investigated. First, does the mapping affect subjects' ability to detect the correct image used to generate each audio stimulus? We predicted that the mapping will have an effect on subjects' success rate, because of asymmetries in the efficacy of the individual audio-visual associations involved in each mapping. Second, does the source audio in the sound corpus has a significant effect on subjects' ability to detect the correct image used to generate each audio stimulus? We predicted that the typology of the audio-units that the corpus consist of will have a significant effect on subjects' success rate, because the sound properties of the source audio can affect the salience of the individual audio-visual associations. Reduction in the salience of the association was expected to increase the level of difficulty of the task and as a result the likelihood of failing to respond correctly is higher.

5.2 Procedures

Subject responses were collected independently. This experiment was performed almost immediately after the first experiment which was described in section 4. 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. 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 the apparatus used for viewing the stimuli, and for capturing and storing their responses. 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, subjects were asked to watch and listen to a series of auditory and visual stimuli. 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. 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.

5.3 Audio Stimuli

Six audio stimuli were synthesised for this study. Three audio stimuli were tested for each mapping, i.e. three sounds per mapping,

resulting in a total of six audio stimuli. Each sound was synthesized using one of the three corpus which were described in detail in the section 4.4 above. The resulting sounds used as stimuli have a duration of 8 seconds.

5.4 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. Therefore these were the primary criteria used to guide the design of the visual stimuli. 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 5. This decision aimed at increasing the level of difficulty, in order to challenge the participants and put the mappings' efficacy to the test.

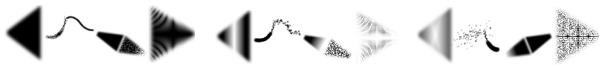

Figure 5. The three visual stimuli used in the second study for testing the achromatic mapping.

5.6 Results

The data analysis (i.e. 2-way anova) revealed statistically significant differences in the ability of the participants to detect the correct image depending on which corpus was used ($F_{(2, 84)} = 11.45$; $p < .0001$); Figure 6 shows the effect of the typology of the audio corpus on participants' detection success. Less significant was the effect of the mapping in the success rate of the participants' responses ($F_{(1, 84)} = 3.8$; $p = .0547$), and no significant interactions were revealed between the variables mapping and corpus ($F_{(2, 84)} = 3.97$; $p = .0224$). Figure 7 shows the distribution of the responses obtained by the subjects for each audio stimulus. The Audio-Visual Stimuli Indexes shown in the figures 7 are explained in table 5. Table 5. shows which Audio Stimuli index from figure 7 corresponds to which audio corpus, and it indicates which the correct answer was for each audio stimuli.

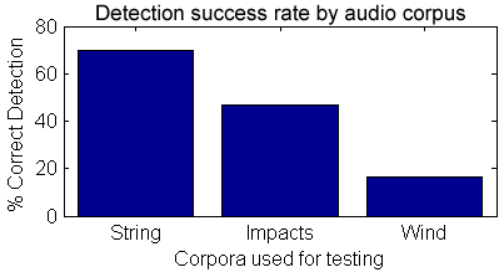

Figure 6. The plot depicts the effect of audio corpus in the ability of the participants to detect the correct image.

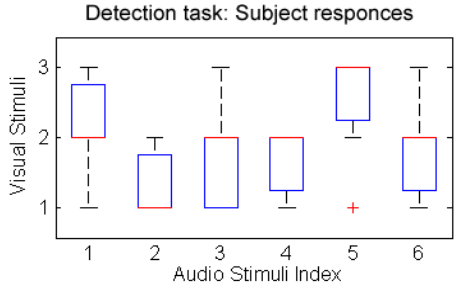

Figure 7. The plot depicts the statistics of the participats' responses in relation to the image that was selected for each audio-stimuli

Table 5. shows which corpora was tested for each audio stimulus across the two mappings. The column labeled *correct answer* indicates which out of the three images was the image used to generate each audio stimulus.

Achromatic Mapping			Chromatic Mapping		
Index	***Audio Corpus***	***Correct answer***	***Index***	***Audio Corpus***	***Correct answer***
1	String	2	4	String	2
2	Wind	1	5	Wind	1
3	Impacts	3	6	Impacts	3

6. CONCLUSIONS

The present paper proposes two methods for testing the perceived similarity and the effectiveness of six audio-visual feature associations in two mappings (i.e. 2x3). The primary purpose of the two mappings tested in the present experiments is to enable visual interaction with corpus based concatenative synthesis for creative applications. Experimental results from the first study revealed differences in the degree of similarity reported by the subjects between the individual audio-visual feature associations that were tested. However subjects' responses did not vary significantly as a result of either the mapping or the typological characteristics of the source audio that the corpus consists of. The present study confirms the results revealed by previous studies [7], [15], [19] which found strong relationships between the audio-visual feature associations of size – loudness, pitch – vertical position, and weaker relationships between color – pitch, texture granularity – sound dissonance and color complexity- sound dissonance. The lowest rated audio-visual association was the texture granularity – dissonance and color variance- sound dissonance. These feature associations were rated very low particularly when the wind corpus was used. A first interpretation of the low ratings for these relationships when using the wind corpus, is that the wind sound does not have a single fundamental frequency and even harmonics therefore has little capacity to exhibit dissonance.

The relationship between vertical position - pitch is the most highly rated, while the relationship between color – spectral centroid is weaker. It is worth noting that the relationship between color – spectral centroid is more difficult to define than the more metaphoric relationship of the vertical position – pitch. This is because in the case of color – spectral centroid, it is hard to determine which color scale is the most appropriate and why. For example, Lipscomb et al. tested three different magnitudes of change: small ranging from: dark blue– blue- light blue, moderate: blue – indigo – violet, and large: red – green – violet. They reported no significant effect of the magnitude of change on subjects' responses. In this study, we used a scale with large magnitude of change (i.e. blue- green – yelow) which was different from the color scales that Lipscomb et. al. had tested. However there was no significant difference between the results reported by their study and the results revealed by the present study. Further research will be required to investigate which color scale is the most appropriate for the relationship between color and pitch.

Experimental results from the second study revealed that participants' success rate in detecting the correct image did not vary significantly as a result of the mapping. However data analysis revealed that the typology of the sounds which an audio corpus consists of has significant effect in the subjects' ability to detect the correct image used to synthesise the audio stimuli. Contrary to the results from the first study, in the

context of the second study the typology of the source audio which the audio corpus consists of appears to be important. A first interpretation of the effect of the typology 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 conflicting results between the two studies regarding the influence of the typology 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 it could be argued 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.

The next step in this research will be to use the findings from the present experiments to inform the current mappings and conduct another series of experiments with a larger sample size. Alternative combinations for the audio – visual feature associations that were not highly rated will be tested. 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. 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.

7. ACKNOWLEDGMENTS

The Author would like to acknowledge the IIDI (Institute for Informatics and Digital Innovation) for their financial support and Dr Grégory Leplâtre for his advice and support throughout the course of this research project.

8. REFERENCES

[1] L. W. Barsalou, Perceptual Symbol Systems. *Behavioural and Brain Science,* vol. 22, pp. 577–660, 1999.

[2] P. Bertelson, B. de Gelder, The Psychology of Multimodal Perception. In *Crossmodal space and Crossmodal Attention*, Oxford University Press pp.141 – 178, 2004.

[3] F. Berthaut, M. Desainte-Catherine, Combining Audiovisual Mappings for 3D Musical Interaction, *in proceedings of ICMC, 2010*.

[4] G. Dubus, R. Bresin, A Systematic Review of Mapping Strategies for the Sonification of Physical Quantities. *PLOS one*, Vol. 8, no. 12, 2013.

[5] G. Fauconnier, *Mapping in Thought and Language.* Cambridge University Press, 1997.

[6] D. Gentner, A. B. Markman, Structure mapping in analogy and similarity. *American Psychologist* vol. 52, No.1, pp. 45-56, 1997.

[7] K. Giannakis, A comparative evaluation of auditory-visual mappings for sound visualisation. *Organised Sound Journal,* vol. 11, no. 3, pp. 297–307, 2006.

[8] A. M. Glenberg, What memory is for. *Behavioral and Brain Sciences* vol. 20 pp. 1– 55, 1997.

[9] R. I. Godoy, Knowledge in Music Theory by Shapes of Musical Objects and Sound-Producing Actions. In Music, Gestalt, and Computing. Springer Verlag, Berlin pp. 106-110, 1997.

[10] R. I. Godoy, Gestural-sonorous objects: embodied extension of Schaeffer's conceptual apparatus. Organised Sound Journal, Cambridge University Press, Vol 11. no. 2, pp. 149-157, 2006.

[11] R. I. Godoy, Gestural Affordances of Musical Sound. In *Musical Gestures: Sound, Movement, and Meaning.* Routledge, pp. 103-125, 2010.

[12] S. Handel, Perceptual Coherence: Hearing and Seeing. Oxford University Press, 2006.

[13] M. Kubovy, D. Van Valkenburg, Auditory and visual objects. *Cognition* vol. 80, pp. 97-126, 2001.

[14] M. Kubovy, M. Schutz, Audio-visual objects. *Review of Philosophy and Psychology*, vol. 1, pp. 41–61, 2010.

[15] M. B. Küssner, Music and shape. *Literary and Linguistic Computing,* vol. 15, 2013.

[16] G. Lakoff, M. Johnson, *Metaphors we live by*, University of Chicago Press, 1980.

[17] G. Lakoff, M. Johnson, *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, 1999.

[18] M. Leman, *Embodied Music Cognition and Mediation Technology.* MIT Press Book, 2008.

[19] S. D. Lipscomb, E. M. Kim, Perceived match between visual parameters and auditory correlates: an experimental multimedia investigation. *ICMPC*, 2004.

[20] G. L. Murphy, Conceptual combination. In *The Big Book of Concepts,* Bradford book, MIT press, pp.443-475, 2004.

[21] C. Parise, C. Spence, Audiovisual Cross-modal correspondences in the general population, In J. Simner, E. M. Hubbard, *Oxford Handbook of Synesthesia*. Oxford University Press, pp. 790-815, 2013.

[22] D. Schwarz, G. Beller, B. Verbrugghe, S. Britton, Real Time Corpus Based Concatenative Synthesis with CataRT, *DAFx,* 2006.

[23] D. Smalley, Spectromorphology: Explaining sound shapes. *Organised Sound Journal*, Vol. 2, No. 2. Cambridge University Press: 107-126, 1997.

[24] E. Somers, A Pedagogy of Creative Thinking based on Sonification of Visual Structures and Visualization of Aural Structures, *Proceedings of ICAD*, 1998.

[25] E. Somers, Abstract Sound Objects to Expand the Vocabulary of Sound Design for Visual and Theatrical Media. *Proceedings of ICAD*, 2000.

[26] C. Spence, Audiovisual Multisensory integration. *Journal of Accoustic Science & Technology,* Vol. 28, No. 2, pp 61 -70, 2007.

[27] R. Walker, The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception and Psychophysics,* Vol 42, No. 5, pp. 491–502, 1987.

9. Appendices

All the stimuli cited in this paper, and additional information about the system are available via the following URL: http://avstudy.wordpress.com