# The Effectiveness of Two Audiovisual Mappings to Control a Concatenative Synthesiser

**Augoustinos Tsiros, Grégory Leplâtre**
Centre for Interaction Design
Edinburgh Napier University
10 Colinton Road, EH10 5DT
a.tsiros@napier.ac.uk,g.leplatre@napier.ac.uk

## ABSTRACT

This paper presents the results of an investigation into audio-visual (AV) correspondences conducted as part of the development of Morpheme, a painting interface to control a corpus-based concatenative sound synthesiser. Our goal was to measure the effects of the corpora on the effectiveness of the AV mappings used by Morpheme to synthesise sounds from images. Two mapping strategies and four corpora were empirically evaluated by 110 participants. To test the effectiveness of the mappings, sounds were generated using Morpheme from images designed to exhibit specific properties. The sounds were then played to the participants of the study and for each sound, they were presented with three images: the image used to synthesise the sound and two similar distractor images. Participants were asked to identify the correct source image and rate their level of confidence. The results show that the audio corpus used to synthesise the audio stimuli had a significant effect on the subjects' ability to identify the correct image while the mapping didn't. No effect of musical/sound training was observed.

## 1. INTRODUCTION

In recent years, there has been a growing interest in the study of multisensory associations, in particular between audio and visual stimuli [1]–[7]. Previous research suggests that the mapping between the user sensorimotor inputs 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 [8], [9]. This understanding has led artists and scientists to investigate cross-sensory mapping strategies for musical interaction.

The research presented in this paper is motivated by the development of *Morpheme*, an interactive application that allows users to synthesise sounds through sketching. 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. For Morpheme to work effectively, its design must be based on meaningful relationships between what the user paints and what they hear. Sounds are generated in Morpheme using corpus-based concatenative synthesis [10].

Previous studies that investigated correspondences between the different senses have shown that there is consistency in the feature correlates which subjects perceive as a good match. For instance, listeners have consistently reported high levels of perceived correspondence between auditory pitch and light intensity (higher auditory pitches for higher luminosity) [11]–[16], visual size and pitch (lower pitches for bigger visual shapes), visual size to loudness (louder sounds for bigger visual shapes) [17]–[22], spatial height to pitch (higher pitches for higher spatial location) [12], [14], [16], [23]–[30]. Some studies have reported a high correlation between texture granularity and sound compactness (high visual granularity for dense audio spectrum) [31], [32] and visual repetitiveness and sound dissonance (harmonic sound for repetitive textures) [19], [31], [32]. While these results can be applied to design of an empirically validated mapping to convert the user's sketches into sound, some important questions remain regarding the effectiveness of such AV mappings in a real-world application.

Recent findings suggest that interactions between different dimensions of auditory and visual objects in multidimensional contexts can occur. For example, an increase in tempo rate is often correlated to an increase in visual motion speed, however when an increase in auditory tempo is accompanied by a decreasing loudness, the correspondence between tempo and speed is weakened [33]. These types of interactions between auditory and visual parameters could potentially compromise the direct application of AV associations that have only been tested independently. Mappings between users' sensorimotor input and sound and musical parameters commonly consist of more than one dimension. For instance, *Morpheme*'s mapping associates five auditory and visual features. A question then arises: are there interactions between the AV associations that the mapping consists of?

In addition, the strength of an association also depends on the dynamism of the parameters involved [33], [34]. For instance, higher pitch sounds are usually perceived as congruent with smaller visual size. However, ascending pitch is congruent with a growing size and descending pitch with

shrinking size [34]. This suggests that the congruence strength can be affected and even reversed depending on whether the stimuli are dynamic or static.

Furthermore, many descriptors deriving from various types of audio analysis are available for use in a potential AV mapping, likewise for the descriptors of an image. This raises the question of which descriptors will be most effective in AV mappings. Finally, the auditory stimuli in most of the studies commonly undertaken tend to be limited to simple synthetic tones. Considering that most natural and musical sounds are more complex than pure sine tones whether these results still apply to complex real-world sounds is uncertain. These are important questions to answer for the development of Morpheme as it uses a multi-dimensional mapping and concatenative sound synthesis uses pre-recorded audio material to synthesise sound that can have complex timbre characteristics.

## 2. MORPHEME

Morpheme is an application that allows sound synthesis to be controlled by sketching. The underlying synthesis technique used is concatenative synthesis. For more information about the graphical user interface of the system, please see [35]. Morpheme extracts visual features from the analysis of the sketch drawn and uses these features to query audio units from Catart's database [10]. During playback, windowed analysis is performed on the pixels of the sketch. 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 analysed. The window dimensions can be determined by the user, however the default size of the analysis window is 9 pixel wide by 240 pixel height. For more detailed information about the implementation, see [36].

### 2.1 The AV Mapping

Some of the most effective empirically validated audio-visual associations discussed in the introduction were selected to develop two mappings that enable the selection of audio units from the corpus. One of the main distinctions between the mappings is that one is *achromatic* while the other is *chromatic*. The first mapping is considered achromatic because the visual descriptors extracted from *Morpheme's* sketch do not use any colour information, whereas the second mapping does. **Table 1** and **Table 2** show the details of the two mappings.

**Table 1.** Achromatic mapping associations between audio and visual descriptors for the retrieval of audio.

| Visual Features | Audio Features |
|---|---|
| Texture granularity | Spectral flatness |
| Vertical position | Pitch |
| Texture repetitiveness | Periodicity |
| Size | Loudness |
| Horizontal length | Duration |

**Table 2.** Chromatic mapping associations between audio and visual descriptors for the retrieval of audio.

| Visual Features | Audio Features |
|---|---|
| Color variance | Spectral flatness |
| Color brightness | Spectral centroid |
| Brightness variance | Periodicity |
| Size | Loudness |
| Horizontal length | Duration |

## 3. METHOD

An experiment was conducted to answer the following questions

- Which overall mapping (chromatic/achromatic) is the most effective?
- Do the sounds used as the source audio (corpus) by the concatenative synthesiser have an effect the effectiveness of the mappings?
- Is the effectiveness of the mappings affected by the experience of the respondents with sound/music?

### 3.1 Subjects

A total of 110 volunteers took part in the study. The study was conducted in two conditions, in what will be referred to as the *supervised* and *unsupervised* conditions. The supervised study took place in Edinburgh Napier University's Auralization suite. Participants used Beyer Dynamics DT 770 Pro monitoring headphones to listen to the audio stimuli. A 17.3 inch laptop screen was used for viewing the visual stimuli. SurveyGismo was used for viewing the stimuli and for recording the participants' responses. The unsupervised study ran as an online survey. In the supervised study a total of 44 participants volunteered to take part. Sixteen were sound practitioners (referred to as experts), and twenty-eight non-sound practitioners (referred to as non-experts). All of the expert participants except two played a musical instrument and the self-reported levels of expertise were 4 basic, 6 intermediate and 4 advanced.

In the unsupervised study, participants were similarly divided into two groups. There were a total of 66 participants recruited via a post on audio and music related mailing lists. 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 levels of expertise were 11 basic, 14 intermediate and 14 advanced. Twenty eight of the expert 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.

### 3.2 Stimuli

#### *3.2.1 Audio Stimuli*

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 four corpora (i.e. String, Wind, Impacts, Birds). The resulting sounds used as stimuli had a duration of 8 seconds each. The string and the birds' corpora consisted of mainly harmonic and periodic sounds, while 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 were 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).

#### *3.2.2 Visual Stimuli*

Six visual stimuli were designed for this experiment. The aim was to test all the associations used in the mapping and to manipulate multiple variable simultaneously. To further validate the effectiveness of the mapping the three images presented for each audio stimuli have been designed to be similar and vary only in some respects (see **Figure 1**). For instance, the first and second images in **Figure 1** are more similar in terms of pitch contour, while the first and third images are similar in terms of texture granularity. This decision aimed to increase the level of difficulty of the task by introducing a confusion variable.

### 3.3 Procedure and Study Design

Participants were presented with a sequence of short sounds synthesised as described in the previous section. They were given a selection of three images with each sound. Participants were told during the training session that only one of the three images displayed on the screen had been used to generate each sound they were presented with. By clicking on an on-screen button using a computer mouse, subjects could playback the audio files as many times as they liked. After listening to each individual sound, participants could use an on-screen radio button to pick one of the three images they thought had been used to generate the sound using Morpheme. After each choice, they were asked to rate their level of confidence in their judgement. The series of tasks took approximately five minutes. The order of presentation of the sounds was randomized.

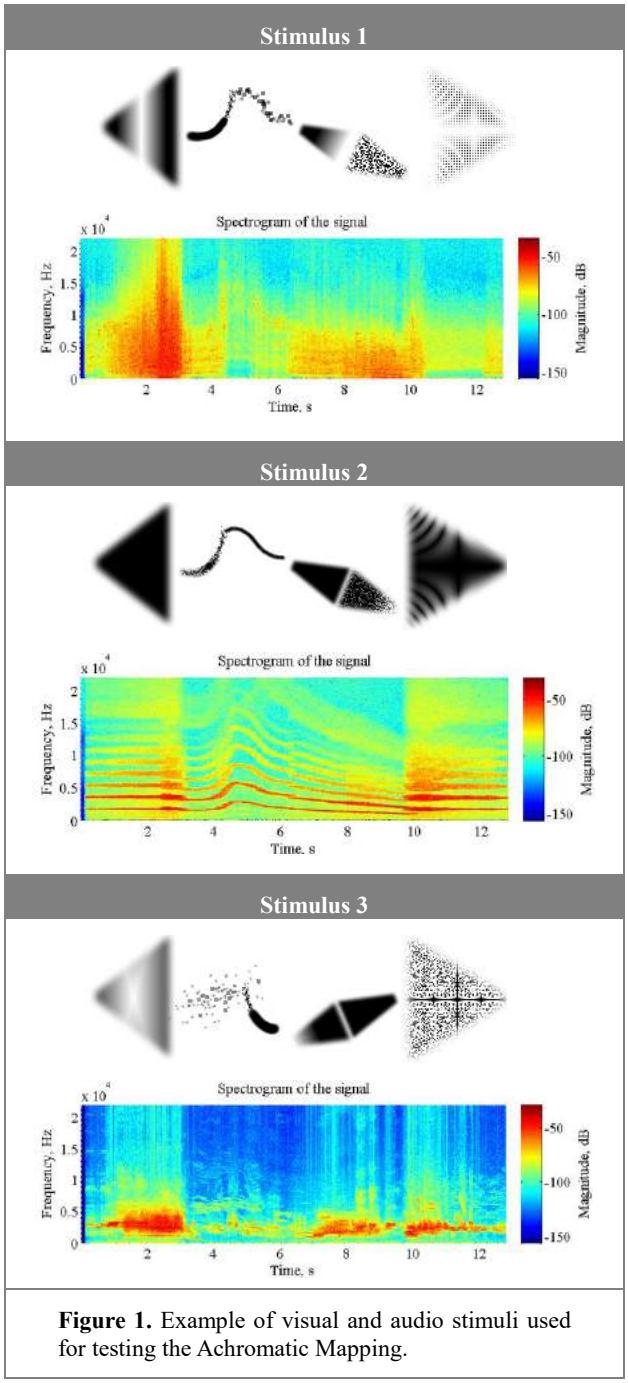


**Figure 1.** Example of visual and audio stimuli used for testing the Achromatic Mapping.

## 4. RESULTS

### 4.1 Correct Image Selection Rate

This section presents the results of the image selection tasks. **Figure 2** and **Figure 3** show the percentage of correct image selection made. Pearson's Chi-square tests of independence were performed to test the null hypothesis of no association between the factors (i) subjects' correct image selection rate and the audio corpora, and (ii) subjects' correct image selection rate and AV mappings.

A strong relationship between correct selection rate and the audio corpora $X^2(3, N = 224) = 19.22$, $p = .0$ was ob-

served. No significant relationship was noted between correct selection rate and the mapping $X^2(1, N = 224) = 1.143$, $p = .28$.

For the expert group in the supervised condition, data analysis shows a strong relationship between the correct selection rate and the audio corpora $X^2(3, N = 128) = 9.86$, $p = .02$. No significant relationship was noted between correct selection rate and the mapping $X^2(1, N = 128) = 3.33$, $p = .068$.

For the non-expert group in the unsupervised condition, data analysis shows that there is a strong relationship between correct selection rate and the audio corpora $X^2(3, N = 216) = 11.37$, $p = .010$. No significant relationship was revealed between correct selection rate and the mappings $X^2(1, N = 216) = .70$, $p = .40$.

For the expert group in the unsupervised condition, data analysis shows that there is a strong relationship between the ability of the subjects to identify the correct image and the audio corpora $X^2(3, N = 312) = 23.15$, $p < .001$. Again, no significant relationship was observed between the correct selection rate and the mappings $X^2(1, N = 312) = .3.07$, $p = .079$.

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 contributed to the statistical significance that was observed. The results of the post-hoc test are presented in **Table 3**. For more information about the approach followed to perform post-hoc testing refer to [37] and [38].

### 4.2 Comparison between expert & non-expert groups

This section presents the results obtained from the comparison between the expert and the non-expert groups in the supervised and unsupervised condition. A Pearson's chi-square test of independence was performed on three between subject variables to test the null hypothesis of no association between: (i) the subjects' skills and correct image selection rate, (ii) subjects' skills and correct image selection rates difference between the two mappings, and (iii) subjects' skills and correct image selection rates difference 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 correct selection rate $X^2(1, N = 352) = 4.43$, $p = .035$.

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 selection rate 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 $X^2(1, N = 176) = .88$, $p = .347$.

The comparison between the expert and the non-expert groups in the unsupervised condition show no association between the variables subjects skills and correct selection rate $X^2(1, N = 528) = 0.10$, $p = .747$. 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 selection rate for either of the mappings: achromatic $X^2(1, N = 264) = .011$, $p = .917$; chromatic mapping $X^2(1, N = 264) = .332$, $p = .564$. Post-hoc testing showed that the relationship between correct detection and subjects' skills in the different experiment conditions was not significant, see **Table 4.**

### 4.3 Confidence Ratings

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.

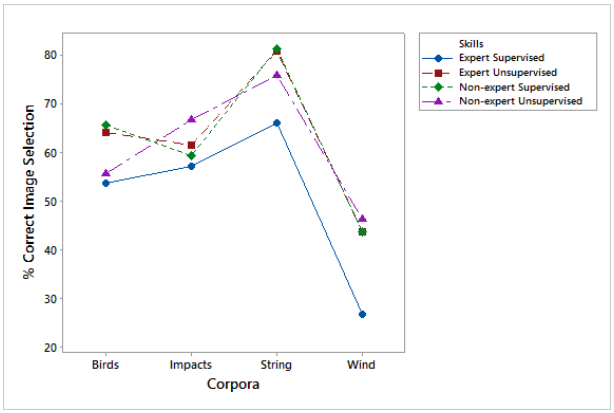


**Figure 2.** Overall effect of the corpus on the subject's ability to identify the correct image.

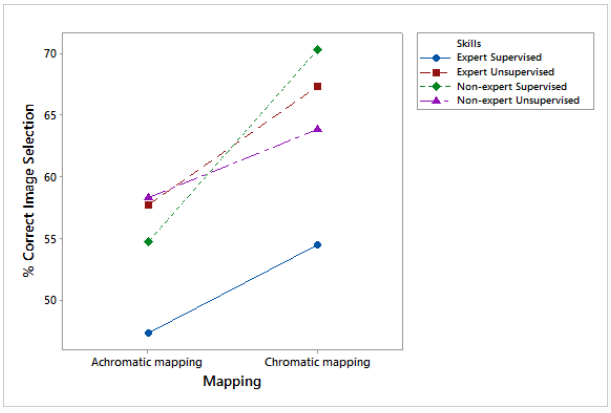


**Figure 3.** Overall effect of the mapping on the subject's ability to detect the correct stimuli.

**Table 3.** 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.

| | **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 0.0062

**Table 4**. Post-hoc testing of correct detection chi-square results for subjects' skills factor.

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

s=supervised/u=unsupervised
Benferroni adjusted p value 0.00625

Analysis of the subjects' confidence ratings revealed no significant differences due to the corpus with an exception for the experts in the unsupervised condition, see **Table 5**. Significant differences were noted between each mapping for the means of the expert group. No interactions between the factors mapping and audio corpus were revealed. The results of the analyses are shown in **Figure 4**. An additional ANOVA model was computed to examine the relationship between reported confidence and correct image identification which was found to be significant for all the groups with only the exception of the expert group in the supervised condition. The results of this analysis are shown in **Table 6**.

### 4.4 Summary of Results

The answer to the first question of this study is that the choice of mapping did not affect participants' image selection success. A trend that can be observed in the data indicates that participants selected the correct images well above chance levels, regardless of the mapping.

The answer to the second question of the study is that the audio corpus used to synthesise the audio stimuli from the images had a significant effect on the subjects' ability to identify the correct image. The image selection 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 for the wind corpus. The analysis showed that the corpora that had the strongest effect on the subjects' successful selection rate were the string and the wind.

In response to the third question, there was no statistically significant difference between the results of the expert and non-expert groups.

**Table 5.** General ANOVA model computed to investigate the effect of the AV 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 6.** General ANOVA model computed to investigate the effect of correct detection on the reported confidence level. An asterisk (*) indicates statistical significance.

| | **Correct selection** | | |
|---|---|---|---|
| ***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

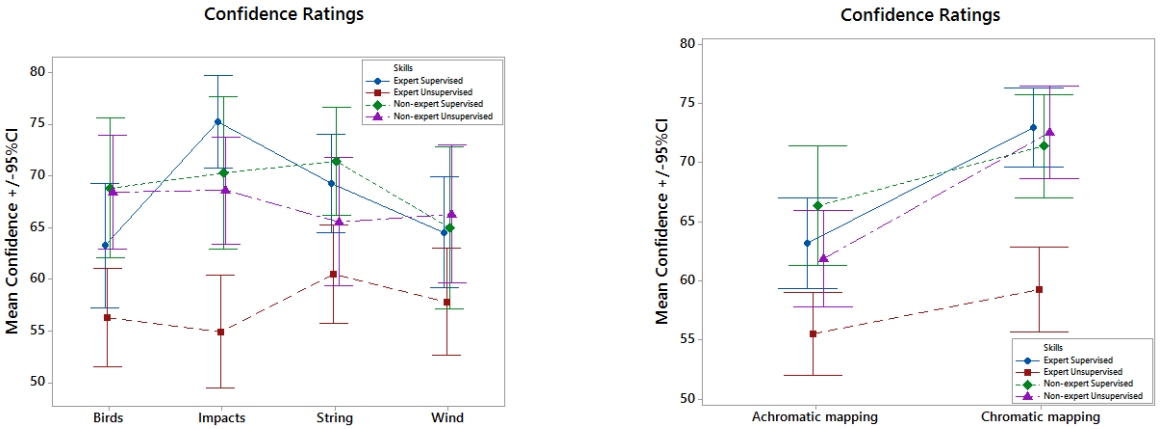


**Figure 4.** Data means and confidence intervals of participants' confidence ratings for the mapping and corpus factors for all of the groups. If intervals do not overlap the corresponding means are statistically significant.

## 5. DISCUSSION

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. experts, non-experts) suggests that musical/sound training was not important. Furthermore consistencies in the subjects' responses between the two experimental conditions (controlled environment and online) suggest that both approaches to gathering data are equally well suited to this type of experiment.

The first result of the study is that the choice of audio corpus had an effect on participants' image selection success rate and therefore affects the effectiveness of the mappings. A first interpretation 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.

As stated in Section 4.4, there was no statistically significant differences between expert and non-expert subjects. The fact that musical/sound training was not a very significant factor suggests that the correspondences might be underpinned by either similarities between auditory and visual perception and/or by cultural conventions, but not specifically by the acquisition of advanced musical and listening skills. These findings are in agreement with [18] and oppose the findings of [17], [20], [39].

Finally, the participants' confidence ratings revealed no correlation between confidence levels and correct image selection. Although a trend can be observed between correct image selection and confidence levels reported by subjects (i.e. confidence levels on average are higher when participants have responded correctly rather than incorrectly), 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 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. CONCLUSIONS

We reported on the evaluation of Morpheme's Audio-Visual mappings used to control a corpus-based concatenative synthesizer through painting. One of the main motivations of the present investigation was to explore how to best utilise empirically validated audio-visual associations in the design of sound synthesis user interfaces. Many studies have shown that people often map stimulus properties from different sensory modalities onto each other in a manner that is surprisingly consistent. Furthermore, previous research suggests that correspondences of stimulus properties may be innate or learned at very early stages of a person's life, [40]–[42]. This research project is motivated by the idea that mappings which are consistent with users' prior perceptual knowledge and the subtleties of multisensory aspects of perception, can better accommodate their sensorimotor skills and support the comprehension of information and interaction with computer systems.

The present experiment tested the effects of four concatenative synthesiser corpora on the effectiveness of two AV mappings, measured by the ability of subjects to identify the image used to synthesise a sound they were presented with. The results showed that when multiple parameters vary simultaneously, the nature of the sounds present in the audio corpus has a moderating effect on the mapping effectiveness. For example, when the corpus consisted of harmonic sounds, participants' detection rates significantly increased in comparison to when using inharmonic sounds. Finally, results indicated that sound/musical training had no significant effect on the perceived similarity between AV features or the discrimination ability of the subjects.

Further work includes making improvements to the mapping of Morpheme, for instance to minimise interactions between audio parameters used (e.g. spectral flatness and periodicity). A set of less correlated audio features is likely to lead to more effective mappings. Another direction for

future work involves testing a wider range of audio-visual features.

**Acknowledgments**
The authors would like to acknowledge the Institute for Informatics & Digital Innovation for their financial support and IMTR research centre for sharing the *CataRT* system.

## 2 REFERENCES

[1] B. Caramiaux, J. Francoise, F. Bevilacqua, and N. Schnell., "Mapping Through Listening," in *Computer Music Journal*, vol. 38, no. 2006, 2014, pp. 1–30.

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

[3] M. Leman, *Embodied Music Cognition and Mediation Technology*. MIT press, 2008.

[4] R. I. Godøy, "Images of Sonic Objects," *Organised Sound*, vol. 15, no. 1, p. 54, Mar. 2010.

[5] D. Smalley, "Spectromorphology: explaining sound-shapes," *Organised sound*, vol. 2, pp. 107–126, 1997.

[6] T. Wishart, *On Sonic Art*. Hardwood academic publishers, 1996.

[7] M. Blackburn, "The Visual Sound-Shapes of Spectromorphology: an illustrative guide to composition," *Organised Sound*, vol. 16, no. 1, pp. 5–13, Feb. 2011.

[8] a. Hunt and R. Kirk, "Radical user interfaces for real-time control," *Proc. 25th EUROMICRO Conf. Informatics Theory Pract. New Millenn.*, pp. 6–12 vol.2, 1999.

[9] A. Hunt, M. Wanderley, and R. Kirk, "Towards a Model for Instrumental Mapping in Expert Musical Interaction," in *International Computer Music Conference*, 2000.

[10] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton, "Real-Time Corpus-Based Concatenative Synthesis with CataRT," in *Digital Audio FX*, 2006, pp. 279–282.

[11] L. E. Marks, "On associations of light and sound: The mediation of brightness," *J. Psychol.*, vol. 87, pp. 173–188, 1983.

[12] I. H. Bernstein and B. A. Edelstein, "Effects of some variations in auditory input upon visual choice reaction time.," *J. Exp. Psychol.*, vol. 87, pp. 241–247, 1971.

[13] T. L. Hubbard, "Synesthesia-like mappings of lightness, pitch, and melodic interval," *Am. J. Psychol.*, vol. 109, no. 2, pp. 219–238, 1996.

[14] R. D. Melara and T. P. O'Brien, "Interaction between synesthetically corresponding dimensions.," *J. Exp. Psychol.*, vol. 116, pp. 323–336, 1987.

[15] R. D. Melara, "Dimensional interaction between color and pitch.," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 15, pp. 69–79, 1989.

[16] G. R. Patching and P. T. Quinlan, "Garner and congruence effects in the speeded classification of bimodal signals.," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 28, pp. 755–775, 2002.

[17] M. B. Küssner, "Shape, drawing and gesture: cross-modal mappings of sound and music," King's College London, 2014.

[18] S. Lipscomb and E. Kim, "Perceived match between visual parameters and auditory correlates: An experimental multimedia investigation," in *International Conference on Music Perception and Cognition*, 2004, pp. 72–75.

[19] F. Berthaut, M. Desainte-catherine, and M. Hachet, "Combining audiovisual mappings for 3d musical interaction," *International. Computer Music Conference*, 2010.

[20] R. Walker, "The effects of culture , environment , age , and musical training on choices of visual," vol. 42, no. 5, pp. 491–502, 1987.

[21] L. B. Smith and M. D. Sera, "A Developmental Analysis of the Polar Structure of Dimensions," *Cogn. Psychol.*, vol. 142, pp. 99–142, 1992.

[22] C. Spence, "Crossmodal correspondences: a tutorial review.," *Atten. Percept. Psychophys.*, vol. 73, no. 4, pp. 971–95, May 2011.

[23] E. Ben-Artzi and L. E. Marks, "Visual-auditory interaction in speeded classification: Role of stimulus difference," *Percept. Psychophys.*, vol. 57, no. 8, pp. 1151–1162, Nov. 1995.

[24] Z. Eitan and R. Timmers, "Beethoven's last piano sonata and those who follow crocodiles: cross-domain mappings of auditory pitch in a musical context.," *Cognition*, vol. 114, no. 3, pp. 405–22, 2010.

[25] K. Evans and A. Treisman, "Natural cross-modal mappings between visual and auditory features," *J. Vis.*, vol. 10, pp. 1–12, 2010.

[26] E. Rusconi, B. Kwan, B. L. Giordano, C. Umiltà, and B. Butterworth, "Spatial representation of pitch height: the SMARC effect.," *Cognition*, vol. 99, no. 2, pp. 113–29, 2006.

[27] S. Hidaka, W. Teramoto, M. Keetels, and J. Vroomen, "Effect of pitch-space correspondence on sound-induced visual motion perception.," *Exp. brain Res.*, vol. 231, no. 1, pp. 117–26, Nov. 2013.

[28] L. E. Marks and Z. Eitan, "Garner's paradigm and audiovisual correspondence in dynamic stimuli: Pitch and vertical direction," *Seeing Perceiving*, vol. 25, pp. 70–70, Jan. 2012.

[29] L. E. Marks, "On Cross-Modal Similarity : The

Perceptual Structure of Pitch , Loudness , and Brightness," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 15, no. 3, pp. 586–602, 1989.

[30] R. Chiou and A. N. Rich, "Cross-modality correspondence between pitch and spatial location modulates attentional orienting.," *Perception*, vol. 41, pp. 339–353, 2012.

[31] K. Giannakis, "Sound Mosaics," Middlesex University, 2001.

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

[33] Z. Eitan, "How pitch and loudness shape musical space and motion: new findings and persisting questions," in *The Psychology of Music in Multimedia*, S.-L. Tan, A. J. Cohen, S. D. Lipscomb, and R. A. Kendall, Eds. Oxford University Press, 2013, pp. 161–187.

[34] Z. Eitan, A. Schupak, A. Gotler, and L. E. Marks, "Lower pitch is larger, yet falling pitches shrink: Interaction of pitch change and size change in speeded discrimination," *Proc. Fechner Day*, 2011.

[35] A. Tsiros and G. Leplâtre, "Evaluation of a Sketching Interface to Control a Concatenative Synthesiser," in *42th International Computer Music Conference*, 2016, pp. 1–5.

[36] A. Tsiros, "A Multidimensional Sketching Interface for Visual Interaction with Corpus-Based Concatenative Sound Synthesis," Edinburgh Napier University, 2016.

[37] T. M. Beasley and R. E. Schumacker, "Multiple Regression Approach to Analyzing Contingency Tables: Post Hoc and Planned Comparison Procedures," *Journal of Experimental Education*, vol. 64, no. 1. pp. 79–93, 1995.

[38] M. a. Garcia-perez and V. Nunez-anton, "Cellwise Residual Analysis in Two-Way Contingency Tables," *Educ. Psychol. Meas.*, vol. 63, no. 5, pp. 825–839, 2003.

[39] M. B. Kussner and D. Leech-Wilkinson, "Investigating the influence of musical training on cross-modal correspondences and sensorimotor skills in a real-time drawing paradigm," *Psychol. Music*, vol. 42, no. 3, pp. 448–469, Jul. 2013.

[40] S. Jeschonek, S. Pauen, and L. Babocsai, "Cross-modal mapping of visual and acoustic displays in infants: The effect of dynamic and static components," *J. Dev. Psychol.*, vol. 10, pp. 337–358, 2012.

[41] S. Wagner, E. Winner, D. Cicchetti, and H. Gardner, "Metaphorical" mapping in human infants.," *Child Dev.*, vol. 52, pp. 728–731, 1981.

[42] P. Walker, J. G. Bremner, U. Mason, J. Spring, K. Mattock, A. Slater, and S. P. Johnson, "Preverbal infants' sensitivity to synaesthetic cross-modality correspondences.," *Psychol. Sci.*, vol. 21, no. 1, pp. 21–5, Jan. 2010.