Animating Timbre - A User Study

Sean Soraghan
ROLI
Centre for Digital Entertainment
sean@roli.com

ABSTRACT

The visualisation of musical timbre requires an effective mapping strategy. Auditory-visual perceptual correlates can be exploited to design appropriate mapping strategies. Various *acoustic* descriptors and *verbal* descriptors of timbre have been identified in the psychoacoustic literature. The studies suggest that the verbal descriptors of timbre usually refer to material properties of physical objects. Thus, a study was conducted to investigate the visualisation of acoustic timbre features using various visual features of a 3D rendered object. Participants were given coupled auditory-visual stimulations and asked to indicate their preferences. The first experiment involved participants rating audio-visual mappings in isolation. The second experiment involved participants observing multiple parameters at once and choosing an 'optimal' mapping strategy. The results of the first experiment suggest agreement on preferred mappings in the isolated case. The results of the second experiment suggest both that individual preferences change when multiple parameters are varied, and that there is no general consensus on preferred mappings in the multivariate case.

1. INTRODUCTION

Timbre is a complex and multi-dimensional attribute of audio. It has been defined as the perceptual attribute of audio by which two sounds with identical pitch, loudness and duration can be discriminated [1]. Before the introduction and popularisation of the computer, the easiest way to produce differences in timbre was through varying instrumentation or articulation. Musical scores therefore elicit changes in timbre by using various articulation indicators (e.g. legato). Computers have introduced the possibility to produce widely varying timbres, in real-time, through the exploration of complex parameter spaces. These parameter spaces have been referred to as 'timbre spaces' [2, 3]. On a traditional musical instrument, timbre manipulation is directly related to articulation. With timbre spaces, however, any form of control interface can be designed since the sound is produced digitally [4].

In modern audio production software environments and graphical user interfaces (GUIs), control interfaces for the exploration of timbre spaces invariably take the form of arrays of sliders and rotary knobs. This form of interaction is sub-optimal and comes from a tendency towards skeuomorphism in interface design. 3D software environments offer the opportunity to present timbre as a complex 3D object, each of its visual features (e.g. brightness, texture) representing a particular parameter of the timbre [5]. This would facilitate intuitive exploration of the timbre space, as the overall timbre would be represented visually as one global object. Such 3D control environments would require the design of a *mapping strategy* such that timbre features are effectively and intuitively visualised to the user.

The aim of this study has therefore been to explore user preferences for timbre-feature to visual-feature mappings. Existing research into both *acoustic* descriptors and *verbal* descriptors of timbre has been drawn upon in order to identify timbre-feature and visual-feature groups and explore user preferences for mappings between the two. As will be explored in the next section, existing research into audio-visual mappings has mainly focussed on static, 2D visual stimuli and rarely concentrates on timbre. This study explores mappings in 3D visual space and is focussed on visual representations of timbre features.

2. RELATED WORK

Most of the previous research into audio-visual mappings has found that users tend to pair colour and position with pitch and volume, and pair timbre features with features of texture/shape [6, 7, 8, 9].

Lipscomb and Kim conducted a user study that investigated the relationship between auditory and visual features of randomised audio-visual stimuli. As audio features they used pitch, loudness, timbre and duration. The visual features they used were colour, vertical location, size and shape [9].

Giannakis and Smith have carried out a number of studies looking at auditory-visual perceptual correlates [10, 7, 11]. Most related to this study is their investigation into sound synthesis based on auditory-visual associations [11]. In that particular study they present a number of corresponding perceptual dimensions of musical timbre and visual texture. Their study focusses on texture alone, however it has been suggested that visual texture qualities are only one type of semantic descriptor used to identify timbre [12]. The present study therefore explores entire 3D structures and includes material properties such as reflectance and transparency. These properties have been chosen in accordance with salient semantic timbre descriptors that have

been identified in existing research (e.g. *volume/fullness* [13, 14, 15, 16, 17], *vision/clearness* [15, 16], *brightness* [13, 18, 16, 17]).

One issue with the aforementioned studies is that they have focussed on 2D static images. As timbre is a multi-dimensional attribute of sound, it seems reasonable that more efficient mapping strategies could be designed in 3D space. Recent research has explored this idea. For example, the TimbreFields project by Corbett et al. involves the simulation of timbre variations produced by physical objects through physical contact such as touch [19]. It is a Virtual Reality project that simulates widely-varying timbres based on user location relative to the object and point of interaction on the object.

More recently, Berthaut et al. conducted a user study very similar to this one in which they presented participants with various audio-visual stimuli and measured their mapping preferences. The visual stimuli were 3D and included features such as texture and rotation. However, their study included pitch and loudness, whereas this study focusses on timbre. Pitch and loudness have been excluded from this study since their mappings are always 1-to-1 (e.g. pitch-colour, loudness-size). The mapping of timbre is more complex as it is an n-to-m mapping, since timbre is a multi-dimensional attribute of audio. As mentioned earlier, the identification of salient visual mappings for timbre features could support the development of intuitive 3D digital interfaces for timbre manipulation.

3. METHODOLOGY

3.1 Participants

18 participants took part in the study (mean age = 28.8, 9 female). 11 had received at least some formal musical training, and 6 were regular users of audio production/synthesis software.

3.2 Stimuli

3.2.1 Auditory Stimuli

Audio tones were generated by additive synthesis using Supercollider. The fundamental frequency was kept constant for each tone, at 311 Hz (Eb4). The audio parameters used in the study, along with their values, were based on those reported by Caclin et al. in their study on acoustic correlates of timbre space dimensions [20]. In their study, Caclin et al. identified 3 salient acoustic timbre descriptors: *attack time*, *spectral centre of gravity* (SCG) and *even harmonics attenuation* (EHA). The same 3 features were used in this study. The attack time varied logarithmically between 15ms and 200ms, as it has been suggested that listeners use a logarithmic scale when using attack time to discriminate between timbres [21]. Caclin et al. provide methodologies for varying the SCG and EHA. The same methods were used in this study. SCG was manipulated using

$$A_n = k * 1/n^{\alpha} \quad (1)$$

where $A_n =$ the amplitude of the n^{th} harmonic. The value of α determines the value of the instantaneous spectral centre of gravity. SCG varied linearly between 1400 Hz (4.5 in harmonic rank units) and 664 Hz (2.1 in harmonic rank units). This was achieved by varying α between 1.23 and 2.07. EHA was controlled by changing the level of the even harmonics relative to the odd harmonics using

$$EH_n = OH_n * 10^{\beta/20} \quad (2)$$

where $EH_n =$ the amplitude of the n^{th} even harmonic and $OH_n =$ the amplitude of the n^{th} odd harmonic, and $\beta =$ the relative change in volume (in dB). During experimentation, β ranged from -8 to 0.

3.2.2 Visual Stimuli

In their investigation into semantic descriptors of timbre, Zacharakis et al. observe that it seems reasonable to identify musical timbre using verbal associations to physical objects‘ properties [13]. As mentioned previously, existing research has identified properties of texture and shape as salient visual correlates of timbre features. Colour and position have been identified mainly as correlates of pitch and loudness. For this reason, no colour was used in the animations and the position remained constant.

Each animation consisted of one 3D rendered polyhedron. The polyhedron was modelled using geodesic subdivision, with an icosahedron as the seed shape [22]. The subdivision depth (spherical resolution) was one of the parameters controlled during animations, and this ranged from 0 to 6. The polyhedron was modelled within a unit sphere, and triangular pyramid ‘spikes’ protruded from each surface face. The length of the spikes was controlled during animations, and ranged from 0 to 1. The two other visual parameters that were varied during animations were brightness and opacity. The visual parameters used in this study (spherical resolution, spike length, brightness and opacity) were based on the 3 factors identified by Zacharakis et al., namely *volume/wealth* (spherical resolution), *brightness and density* (brightness/opacity), and *texture and temperature* [13]. Various surface textures were possible on the polyhedron through a combination of different spherical resolution and spike length values, as demonstrated in figure 1. The animations were implemented using c++ and openGL.

3.3 Experimental Procedure

Participants were asked to complete two separate tasks, both of which involved giving indications of their preference for different audio-visual mapping strategies. For each task, participants sat in front of a 15“ laptop display and were equipped with headphones. Participants used simple graphical interface panels (developed in Supercollider) on the right of the screen in order to listen to different audio tones and cycle though different mapping strategies. The resulting visulisations were displayed in a large window on the left of the screen.

3.3.1 Task 1: Individual Parameter Mapping

Objectives

Task 1 was designed in order to introduce participants to the different audio and visual parameters, and to record

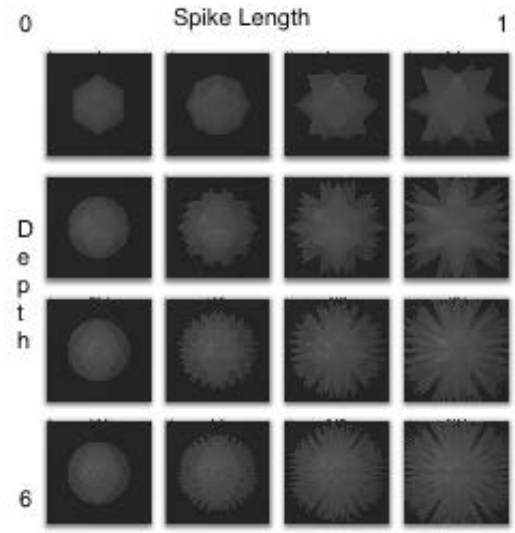

Figure 1. Rendered polyhedra with varying spherical resolution and spike length values.

their favourite to least favourite visual mappings for each audio parameter.

Procedure

During task 1, a single audio parameter changed while the others remained constant. For each audio parameter, participants were presented with three consecutive tones. The selected audio parameter was increased between each tone. Parameter values increased between the values reported previously in section 3.2.1. Each tone produced a resulting visualisation in which one of the visual features changed along with the audio feature, according to the selected mapping. Participants could observe an audio-visual stimulation by pressing the 'play' button in the control interface. There were also buttons to change which audio feature was being varied, and which visual feature the audio feature was mapped to.

For each audio feature, participants were asked to cycle through the different visual mappings and rank them from 3 (favourite) to 0 (least favourite). Thus, they constructed a *preference table* for the various mapping possibilities. An example of a participant's preference table is given in table 1. The participants filled in their preference table as they progressed through the task. They were able to observe stimuli as many times as they needed.

	Res	Spike	Bright	Opacity
Attack	3	2	0	1
SCG	1	3	2	0
EHA	1	2	3	0

Table 1. An example preference table for a participant (3 = favourite, 0 = least favourite).

Results

The 'Borda count' can be used to analyse the results of a preference vote. It is basically a weighted count of votes, weighted on the preference of each vote. In this case, for each audio parameter, every visual mapping is given a Borda count. Every time a participant gives a preference rating to a visual mapping, the value of that preference rating is added to the overall Borda count of the visual mapping.

Figure 2 shows the Borda scores for all visual mappings, for attack (blue), SCG (red) and EHA (green). Similarly,

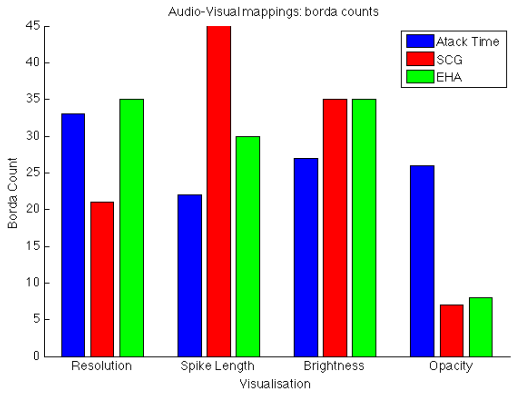

Figure 2. Borda counts for each visual mapping, for all audio features.

figure 3 shows the Borda counts as a scatter plot, where the point size represents the overall popularity of each audio-to-visual mapping, compared to the other options.

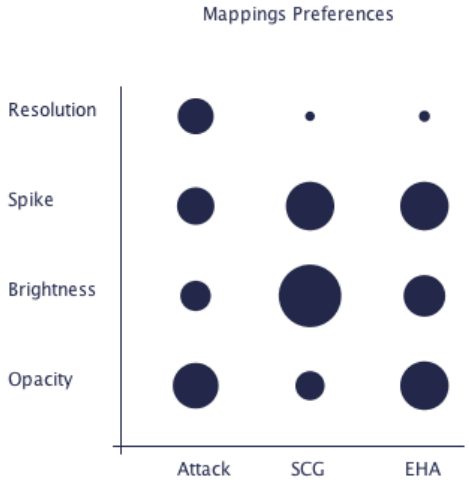

Figure 3. Borda counts as a weighted scatter plot. Point size = relative popularity of an audio-to-visual mapping in comparison to the other options.

Another way to analyse the results of a preference vote is to use a Condorcet method. This pits every candidate (e.g. mapping) against one another in pair-wise hypothetical contests. If one candidate wins every contest, they are considered the Condorcet winner. In this case, a certain visual mapping is the Condorcet winner for an audio feature if it has a higher (or equal) Borda count than all other possible visual mappings for that audio feature.

Each audio feature had a visual mapping that emerged

as the Condorcet winner when all visual mappings were compared using a Condorcet method, as shown in table 2.

	Visual Mapping
Attack	Resolution
SCG	Spike Length
EHA	Brightness

Table 2. Condorcet winner visual mappings for each audio feature.

3.3.2 Task 2: Multiple Parameter Mapping

Objectives

Task 2 involved all of the audio and visual mappings changing at once. The task was designed to encourage the participants to explore different global mapping strategies and to record their recommended optimal global mapping strategy. A key objective was to ascertain to what extent mapping preferences change (if at all) when multiple parameters are varied together.

Procedure

During task 2, participants listened to short audio tones in which each of the audio features were randomised. The SCG and the EHA were also varied *during* the audio tones, using randomised linear envelopes. Such audio tones produced visual animations where the visual features of the polyhedron varied smoothly and in direct response to the audio, according to the mapping configuration. Participants could observe a randomised audio-visual stimulation at any point, by pressing 'play' in the control interface. There were also four buttons for each visual parameter, which allowed the participants to change which audio feature was controlling that visual parameter.

Using the preference tables from task 1 (table 1), suggested optimal mapping strategies were constructed. Table 3 gives an example. These were used to construct the initial mapping strategies in task 2. Participants were then encouraged to explore different mapping configurations and evaluate them by observing some randomised audio-visual stimulations. Ultimately participants indicated what they thought was the *optimal* mapping configuration by filling in an optimal mapping configuration table (see table 4 for an example). The mapping configuration consisted of each visual feature being controlled by a single audio parameter and multiple visual features could be controlled by the same audio feature.

	Suggested Visual Mapping
Resolution	Attack
Spike	SCG
Brightness	EHA
Opacity	Attack

Table 3. An example suggested optimal mapping strategy for a participant (using the results from task 1).

	Optimal Visual Mapping
Resolution	EHA
Spike	SCG
Brightness	EHA
Opacity	SCG

Table 4. An example optimal mapping strategy for a participant (from task 2).

Results

Participants' *suggested* mapping configurations from task 1 were compared to their *'optimal'* mapping strategy from task 2. Comparing the two tables, a measure of the difference between the preference table and the optimal strategy can be evaluated. This difference is calculated as the total number of mappings in the 'optimal' strategy that differ from the suggested mappings from the preference table. For example, the difference between the suggested mapping strategy in table 3 and the optimal mapping strategy in table 4 would be 2. This value gives an indication as to what extent a participant's preferences from task 1 varied, after exploring global strategies during task 2.

In total, 14 (78%) of the participants' optimal strategies differed from their suggested strategies. 8 of these changed by 1 mapping, 2 changed by 2 mappings, and 4 changed by 3 mappings. Figure 4 shows how often each visual attribute changed, between suggested and optimal mapping strategies.

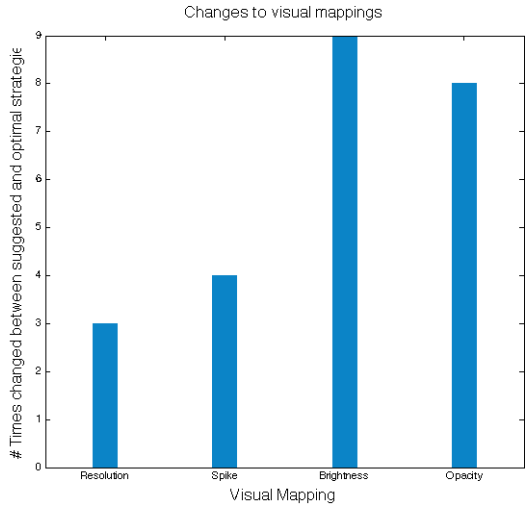

Figure 4. Number of times each visual mapping changed between suggested and optimal mapping strategies.

In total there were 12 unique optimal mapping strategies that emerged from the study, the most popular being common to only 3 participants (shown in table 5). 4 optimal strategies were common to 2 participants, and the other 7 were unique to individual participants.

	Optimal Visual Mapping
Resolution	EHA
Spike	SCG
Brightness	SCG
Opacity	Attack

Table 5. 'Most popular' optimal mapping strategy (common to 3 participants).

4. DISCUSSION

Task 1 identified preferred visual mappings for attack time, SCG and EHA. The existence of a Condorcet winner for each suggests that there was general agreement on which visual features best represented which audio features in isolation, namely attack-spherical resolution, SCG-spike length, and EHA-brightness. However, figure 3 identifies a very wide spread between preferred mappings, especially for attack time and EHA. Attack time, SCG and EHA have been reported as salient axes for timbre discrimination [20]. Volume, texture and brightness have been reported as salient verbal descriptors of timbre variance by [13]. The results from task 1 can possibly indicate which verbal descriptors relate to which acoustic features (where spherical resolution relates to volume and spike length relates to large-scale texture). Larger studies are required to confirm or refute these findings. Some of the participants commented that the mapping from SCG and EHA to opacity should have been inverted. This may have had an effect on the popularity of opacity as a visual mapping.

Task 2 identified variation between many individual participants' isolated mapping preferences and their global optimal mapping strategy. This suggests that mapping preferences change when multiple parameters are in flux. Again, larger studies are required in order to further evaluate this suggestion. The variance in individual preferences could be due to the psychoacoustic perception of timbre, as it has been suggested that the salience of acoustic timbre features depends on the number of features in flux [20]. It is interesting to examine the total number of times (across all participants) each individual visual mapping was changed between suggested and optimal mapping strategies. Figure 4 indicates that there was more disagreement on the use of brightness and opacity as visual mappings than there was for resolution and spike length.

Task 2 was designed to encourage exploration in the participants, such that their preferences were their own, rather than one of a limited number of options presented to them. To facilitate this, the number of possible mapping configurations was left intentionally large. However, this resulted in a large cognitive load for the participants. Thus, despite the objective of avoiding 'right or wrong answers,' it is possible the large cognitive load resulted in the task feeling 'too difficult' for some participants. The measure of 'difference' between mapping strategies, as defined in section 3.3.2 can only be used as a very general indication of difference, since the differences being measured are perceptual and their magnitudes vary.

There was a large degree of variance between the different participants' suggested global optimal mapping strategies. This supports the idea that mapping preferences change as the number of mappings increases, and suggests ambiguity in preferred visual mapping from listener to listener.

5. CONCLUSION

The aim of this study was to combine findings about verbal timbre descriptors and acoustic timbre features and explore preferred mappings between the two. In the case of isolated mappings there seem to be clear audio-to-visual mapping preferences. When multiple mappings are considered, no clear preference emerges, and preferences sometimes change from the isolated case.

These findings suggest that any graphical applications exploiting perceived associations between auditory timbre and visual form may benefit from customisable mapping strategies. However, the participants involved in this study were not widely representative of prospective users of such systems (with only 6 being regular users of audio production software). Future studies should therefore possibly consider more homogeneous participant groups.

This study used only monophonic, synthesised audio tones. Future work should include natural/acoustic and/or polyphonic audio stimuli. Larger studies are required to confirm or refute the findings reported here. Studies with larger participant numbers could also help identify whether there are different categories of preference (e.g. whether certain mapping combinations usually go together). The acoustic and visual features used in this study were based on findings reported elsewhere, but future studies may benefit from using larger parameter sets.

Acknowledgments

The author would like to thank all of the participants for volunteering to take part in the study.

6. REFERENCES

[1] P. I. Terminology, "American national standard ansi," *ISA S51*, vol. 1, 1979.

[2] D. L. Wessel, "Timbre space as a musical control structure," *Computer music journal*, pp. 45–52, 1979.

[3] C. Nicol, S. A. Brewster, and P. D. Gray, "Designing sound: Towards a system for designing audio interfaces using timbre spaces." in *ICAD*, 2004.

[4] A. Hunt, M. M. Wanderley, and M. Paradis, "The importance of parameter mapping in electronic instrument design," *Journal of New Music Research*, vol. 32, no. 4, pp. 429–440, 2003.

[5] F. Berthaut, M. Desainte-Catherine, M. Hachet *et al.*, "Combining audiovisual mappings for 3d musical interaction," in *Proceedings of International Computer Music Conference*, 2010.

[6] X. Wu and Z.-N. Li, "A study of image-based music composition," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 1345–1348.

[7] K. Giannakis and M. Smith, "Imaging soundscapes: Identifying cognitive associations between auditory and visual dimensions," *Musical Imagery. Swets & Zeitlinger*, pp. 161–179, 2001.

[8] R. Walker, "The effects of culture, environment, age, and musical training on choices of visual metaphors for sound," *Perception & Psychophysics*, vol. 42, no. 5, pp. 491–502, 1987.

[9] S. D. Lipscomb and E. M. Kim, "Perceived match between visual parameters and auditory correlates: An experimental multimedia investigation," in *Proceedings of the 8th International Conference on Music Perception and Cognition*, 2004, pp. 72–75.

[10] K. Giannakis and M. Smith, "Auditory-visual associations for music compositional processes: A survey," in *Proceedings of International Computer Music Conference ICMC2000, Berlin, Germany*. Citeseer, 2000.

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

[12] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustical correlates of polyphonic timbre," 2010.

[13] A. Zacharakis, K. Pastiadis, G. Papadelis, and J. D. Reiss, "An investigation of musical timbre: Uncovering salient semantic descriptors and perceptual dimensions." in *ISMIR*, 2011, pp. 807–812.

[14] G. von Bismarck, "Timbre of steady sounds: A factorial investigation of its verbal attributes," *Acta Acustica united with Acustica*, vol. 30, no. 3, pp. 146–159, 1974.

[15] J. Stepánek, "Musical sound timbre: Verbal description and dimensions," in *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, 2006, pp. 121–126.

[16] O. Moravec and J. Štěpánek, "Verbal description of musical sound timbre in czech language," *organ*, vol. 4, pp. 3–7, 2003.

[17] D. M. Howard and A. M. Tyrrell, "Psychoacoustically informed spectrography and timbre," *Organised Sound*, vol. 2, no. 2, pp. 65–76, 1997.

[18] R. Pratt and P. Doak, "A subjective rating scale for timbre," *Journal of Sound and Vibration*, vol. 45, no. 3, pp. 317–328, 1976.

[19] R. Corbett, K. Van Den Doel, J. E. Lloyd, and W. Heidrich, "Timbrefields: 3d interactive sound models for real-time audio," *Presence: Teleoperators and Virtual Environments*, vol. 16, no. 6, pp. 643–654, 2007.

[20] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tonesa)," *The Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 471–482, 2005.

[21] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychological research*, vol. 58, no. 3, pp. 177–192, 1995.

[22] M. Lounsbery, T. D. DeRose, and J. Warren, "Multiresolution analysis for surfaces of arbitrary topological type," *ACM Transactions on Graphics (TOG)*, vol. 16, no. 1, pp. 34–73, 1997.