ON THE SUITABILITY OF MAPPINGS FROM MUSICAL TO VISUAL FEATURES

Regina Collecchia

CCRMA
Stanford, CA, USA
colleccr@ccrma.stanford.edu

ABSTRACT

A study of audio to visual mappings is executed. The audio envelopes of FM synthesis are the independent variables, and the visual axes of intensity, radius, color, and roughness are the dependent variables. Responses to a two-interval forced choice experiment that asked for a preference between two different mappings of the same sound were collected. Results were inconclusive and for the most part statistically insignificant.

1. INTRODUCTION

1.1. Background

Generating cohesive visual images to a given sound has been an important part of multimedia applications such as video gaming, algorithmic visual accompaniment for musical performance, and speech-driven facial animation [1].

Judging by the usefulness of spectrograms in visually identifying musical events, timbres, and patterns, it seems reasonable that compelling visualizations of musical data in psychological and emotional domains could very well exist. Even the visual responses we have to spectrograms about music have been studied [2], [3]. Still, research on audio-visual relationships is widely experimental.

In the present study, I seek to further our understanding of how we abstract musical sounds as visual objects, and the sort of mappings we can use to successfully encode them. This is largely motivated by algorithmic generation of visuals for live musical performance, which is attracting more attention than ever before with acts like Amon Tobin's ISAM show (projection mapping) and Squarepusher's Ufabulum tour (large LED arrays with variable intensity). Audiences walk away from these examples clearly impressed much more by the visual component than the musical. This especially applies to ISAM, for which the visuals took a year to create and millions of dollars [4].

Evans and Treisman [5] mapped visual height (and in another experiment, radius) to pitch and visual orientation (angle) to instrument (timbre). They also tried mapping contrast (which looks a lot like opacity) AND height to pitch, with significant results from their two-interval forced choice (2IFC) experiments.

Drawing from these experiments and ideas, I design my own investigation into the visualization of music.

1.2. Experiment design

This experiment has a 2IFC design in which subjects are presented with 2 stimuli, one presented after a second with a black fullscreen after each video, and then they are asked to make a preference.

Because of the enormous number of (simple) variables in the visual world compared to the musical world, the visual axes I chose were inspired by the aural qualities of the audio stimuli. Therefore, the visual stimuli need to be generated by the sound. Frequency-modulation (FM) synthesis is used to generate the audio stimuli. In FM synthesis, there are no less than four numerical parameters used to define sound: (1) frequency, (2) harmonicity, (3) modulation index, and (4) amplitude [6]. Each of these can have a time-domain envelope, and each more or less influence the resulting sound to an equal degree. Furthermore, the features are linearly independent of one another, and musical qualities like timbre and melody can be described as linear combinations of them.

I presuppose that there are natural ways to connect some of the musical to the visual features, such as musical intensity to visual brightness and/or size, and musical complexity/noisiness to visual chaos. Timing is another obvious connection between music and

animation. Less obvious, something like the foreground & background voices in music could map to focus & blur in a picture. I emphasized salience in selecting the visual features. I think of this as "roughness" (and later, *concavity*). I used these sorts of intuitions to guide my experiment design.

Finally, there exist $\sum n$ possible mappings with which to generate audio-visual pairs, where n the order of both the visual and audio space.[1] Therefore, a smaller feature space is desirable.

1.3. Hypothesis

I hypothesize that the radius, transparency, and concavity of the visual object will be compelling features in the visual domain. I suspect that radius and transparency will be more natural to amplitude, and that the concavity of the edges of the shape will be a good mapping of the spectral content, where more complex timbres will map well to more extreme concavities.

On the contrary, I predict that color will not bear a strong relationship to any of the audio features, except for subjects with musical synesthesia. They typically experience a correlation of pitch to frequency and amplitude to brightness during auditory phenomenon [7].

It is hard to say whether radius or opacity will be a better mapping to loudness. Perhaps some combination of both would be most suitable, but that is outside of the scope of this experiment.

Finally, I predict that (consistent) preferences of mappings might vary subject-to-subject. Perhaps this will be correlated with musical experience.

2. METHOD

Trials consist of 2 videos of audio-visual pairs which subjects are asked to compare. I define 4 features of FM and 4 features of animations, and choose 4 mappings out of the possible mappings for each trial. Map 1 differs from Map 2 in two dimensions and is identical in the other 2. The same applies to Map 3 versus Map 4. Therefore, Map 1 is only compared to Map 2, while Map 3 is only compared to Map 4 to try to isolate parts of the mappings. In order for all features to be contrasted, three experiments would need to be done (Map 1 versus 3 and Map 2 versus 4, Map 1 versus 4 and Map 2 versus 3). In this paper, only the first experiment was performed.

[1] In a forced choice framework, n must be even.

2.1. Subjects

Nine subjects from Stanford University were issued the experiment, 2 females and 7 males. From the survey they were given, 3 of the subjects had some form of synesthesia (the author among them, with color-number synesthesia), 1 had perfect pitch, the average age was 24.33 years, and the average musical experience was 11.33 years. Subjects were administered the experiment in quiet conditions with Audio Technica MTH-50 headphones.

2.2. Stimuli

The audio stimuli were all created by FM synthesis in Max 6. The carrier frequency, harmonicity, modulation index, and amplitude envelopes were all carefully recorded for each sample to produce the data for the associated visual mappings. These envelopes for each sound are shown below.

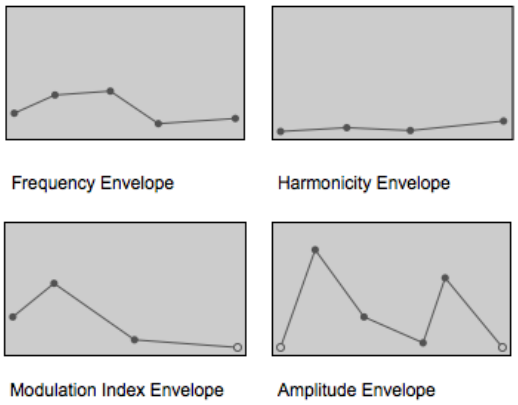

Figure 1. Sound 1 envelopes.

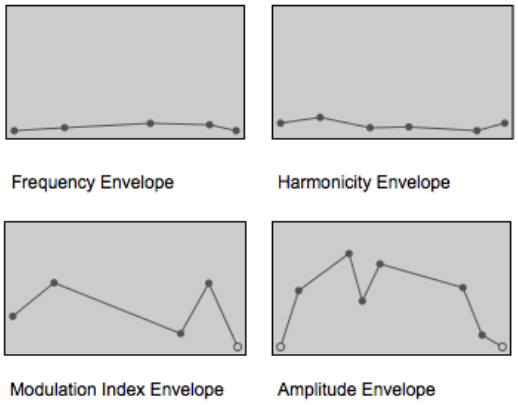

Figure 2. Sound 2 envelopes.

Upon later reflection, I believe that some of the resulting sounds are too similar to one another on the

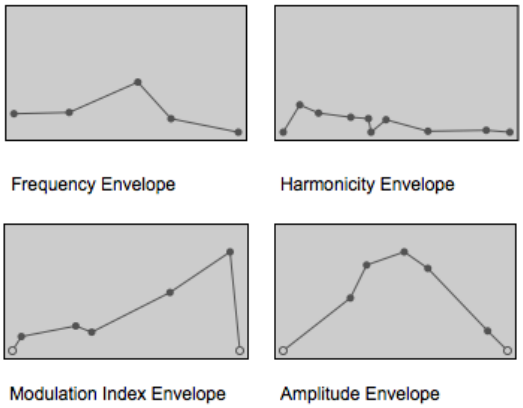

Figure 3. Sound 3 envelopes.

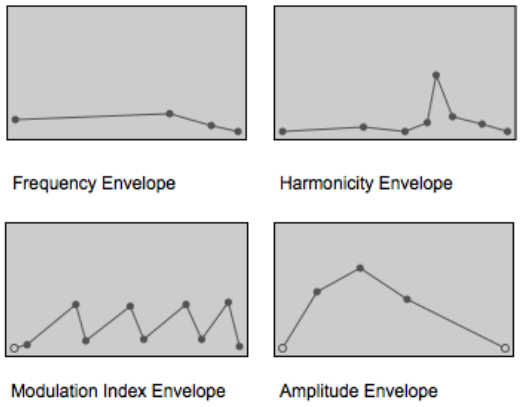

Figure 4. Sound 4 envelopes.

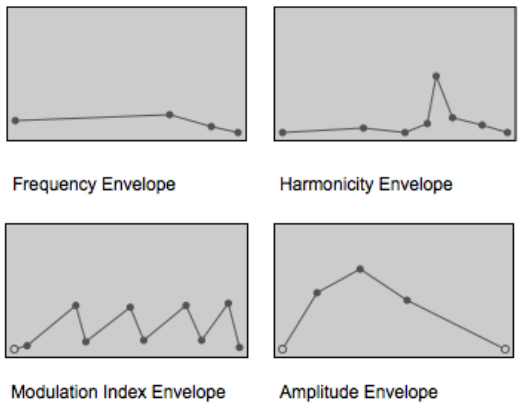

Figure 5. Sound 5 envelopes.

whole. Creating new sounds will be the first step in further experimentation.

The visual stimuli are 6-sided "starburst" shapes, synthesized in Processing 2.0b9 using the MaxLink library to receive messages from Max. By pressing spacebar in the Max window, images were generated in Processing at a frame rate of 8 fps with resolution

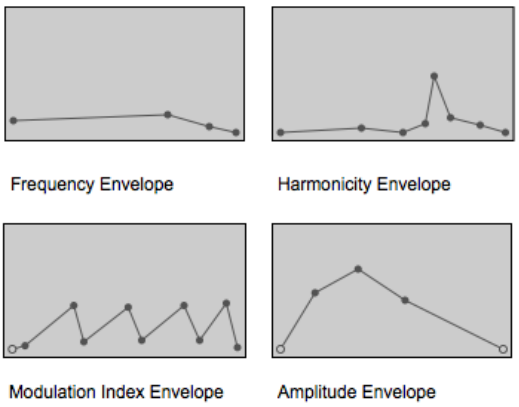

Figure 6. Sound 6 envelopes.

480×320. Quantities larger than these caused Processing to lag at the onset on my machine; perhaps some code can be added to allow the Max envelopes to be buffered.

The shapes rotate constantly with time at a rate of 2.4 degrees per frame. I did this to add interest to the images. I also wished to add a "trail" of the shape that would sustain previous frames with decreasing opacity, but ran out of time as I was learning how to use Processing for the purpose of this class.

2.3. Maps

For the sake of brevity, we use the following names for our variables.

- $x_f \sim$ (audio) frequency envelope
- $x_h \sim$ (audio) harmonicity envelope
- $x_m \sim$ (audio) modulation index envelope
- $x_a \sim$ (audio) intensity/amplitude envelope
- $y_c(red, blue, green) \sim$ (visual) color
- $y_e \sim$ (visual) concavity of the edge/contour
- $y_\alpha \sim$ (visual) intensity, i.e. transparency
- $y_r \sim$ (visual) radius

We map audio features isomorphically to visual features. In order for the experiment to be of a reasonably short length, only 4 mappings will be tested, and they will generate visuals from only 6 sounds.

The algorithms for each mapping stay generally constant, differing only in the values of the coefficients and constants in order to (1) produce significant changes to the visual object given the range

of the audio envelopes, and (2) keep the visualization within the 480×320 frame of the video. The envelopes are normalized, where

$$\hat{A}_f(t) = \frac{x_f(t) - f_{min}}{f_{max}} \quad (1)$$

$$x_{harm}(t) = \frac{x_h(t) - h_{min}}{h_{max}} \quad (2)$$

$$x_{mod}(t) = \frac{x_m(t)}{m_{max}} \quad (3)$$

$$x_{amp}(t) = \frac{x_a(t)}{a_{max}} \quad (4)$$

To avoid repetition, I will give the functions in a general form, with $x_i(t)$ meaning some input envelope, always different from $x_j(t)$ with $i \neq j$ and $t = [0 : 0.125 : 5]$ seconds. I will also specify constants in this way, with a_i and b_i. The width $w = 480$ and height $h = 320$ are also considered occasionally.

I made the functions all directly proportional to their independent variables, for the sake of time. In future work, other relationships should be studied.

The mapping for **radius** $r(t)$ is given by

$$r_i(t) = (ah)x_i(t) + bh \quad (5)$$

The constants were in the range $0.25 \leq a \leq 0.5$, and $b = 0.125$.

Color[2] $red(t), green(t), blue(t)$ was translated by the equation

$$red_i(t) = 2 - 2x_i(t) \quad (6)$$

$$green_i(t) = 1 - x_i(t) \quad (7)$$

$$blue_i(t) = x_i(t) \quad (8)$$

The **concavity** $c(t)$ of the curve between the 6 vertices was determined by

$$c_i(t) = r_i(t)ax_i(t) + b \quad (9)$$

so it was dependent on the radius. The scalar was in the range $1 \leq a \leq 5$ and was different for each mapping. The constant $b = 10$ only for Map 1; otherwise it was 0.

Finally, **opacity** $\alpha(t)$ (RGBA, in the range 0 to 1) was given by

$$\alpha_i(t) = ax_i(t) + b \quad (10)$$

where $0.9 \leq a \leq 6$ and $0.1 \leq b \leq 0.2$.[3]

[2] This is in the RGBA format for color, which is normalized so that all values are between 0 and 1. So $y_c = (red, green, blue, \alpha)$.

[3] The large a maximum here for example indicates to me that the functions should be normalized on a sound-to-sound basis.

The explicit mappings are given below, and some of the images resulting from Map 2. I attempt to illustrate the extremes of the parameters.

Audio feature(s)	Visual feature(s)
Frequency envelope x_f	Color y_c
Harmonicity envelope x_h	Intensity y_α
Modulation index envelope x_m	Concavity y_e
Amplitude envelope x_a	Radius y_r

Table 1. Map 1 description

Audio feature(s)	Visual feature(s)
Frequency envelope x_f	Color y_c
Harmonicity envelope x_h	Radius y_r
Modulation index envelope x_m	Concavity y_e
Amplitude envelope x_a	Intensity y_α

Table 2. Map 2 description

Audio feature(s)	Visual feature(s)
Frequency envelope x_f	Concavity y_e
Harmonicity envelope x_h	Intensity y_α
Modulation index envelope x_m	Color y_c
Amplitude envelope x_a	Radius y_r

Table 3. Map 3 description

Audio feature(s)	Visual feature(s)
Frequency envelope x_f	Concavity y_e
Harmonicity envelope x_h	Radius y_r
Modulation index envelope x_m	Color y_c
Amplitude envelope x_a	Intensity y_α

Table 4. Map 4 description

Figure 7. Shape generated from Map 2, Sound 4.

2.4. Procedure

The presentation of stimuli in this experiment is as follows:

1. Fullscreen Sound 1 under Map 1, 5000 ms.
2. Black fullscreen for 2000 ms.

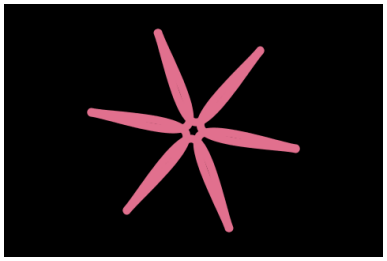

Figure 8. Shape generated from Map 2, Sound 4.

Figure 9. Shape generated from Map 2, Sound 4.

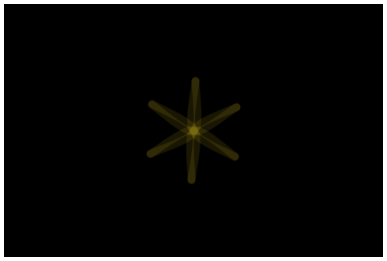

Figure 10. A typical starting (and ending) point of the shape. They began their rotation at $\theta = 0$.

Figure 11. Shape generated from Map 2, Sound 5.

3. Fullscreen Sound 1 under Map 2, 5000 ms.
4. Black fullscreen for 2000 ms.
5. **Forced-choice subject response**.
6. Fullscreen Sound 2 under Map 3, 5000 ms.
7. Black fullscreen for 2000 ms.
8. Fullscreen Sound 2 under Map 4, 5000 ms.
9. Black fullscreen for 2000 ms.
10. **Forced-choice subject response**.
11. Fullscreen Sound 3 under Map 2, 5000 ms.
12. Black fullscreen for 2000 ms.
13. Fullscreen Sound 3 under Map 1, 5000 ms.
14. Black fullscreen for 2000 ms.
15. **Forced-choice subject response**.
16. Fullscreen Sound 4 under Map 4, 5000 ms.
17. Black fullscreen for 2000 ms.
18. Fullscreen Sound 4 under Map 3, 5000 ms.
19. Black fullscreen for 2000 ms.
20. **Forced-choice subject response**.
21. Fullscreen Sound 5 under Map 1, 5000 ms.
22. Black fullscreen for 2000 ms.
23. Fullscreen Sound 5 under Map 2, 5000 ms.
24. Black fullscreen for 2000 ms.
25. **Forced-choice subject response**.
26. Fullscreen Sound 6 under Map 3, 5000 ms.
27. Black fullscreen for 2000 ms.
28. Fullscreen Sound 6 under Map 4, 5000 ms.
29. Black fullscreen for 2000 ms.
30. **Forced-choice subject response**.

and so on, for all 6 sounds and all 4 mappings. Then, the order was reversed, so each unique comparison was made twice. Hence, there were a total of 24 videos and 24 trials (6 sounds $\times$ 4 mappings $=$ 24 videos, 24 videos / 2 possible mappings $\times$ 2 orderings $=$ 24 trials). Subjects were allowed to repeat trials as many times as they wished. There were also two practice trials, about which I verbally informed the second half of subjects, and for the first half simply threw out these trials from analysis. Therefore, the experiment took at least 6 minutes to complete.

2.5. Responses

The experiment's GUI was designed in Max/MSP 6.1. Subjects clicked buttons to answer to the question, "Which animation was a better depiction of the sound you heard? Click one of the following, or play it again." Their response ("1st", "2nd", "played again") was recorded in the Max Window. They could not move on to the next trial until they had picked a preference (forced choice).

I added two checkboxes about halfway through runs that allowed subjects to indicate whether or not they felt both stimuli were bad or good, but they still had to make a preference before they could continue. These results were exported into CSV files, stored in Excel (regrettably), and analyzed in MATLAB.

Subjects were also asked to respond to 5 survey questions before they did the experiment: gender, age, years of musical training (lessons, classes), and whether or not they had synesthesia and perfect pitch.

Each audio-visual pair was shown a total of two times to the subject. Since two videos played per trial, the ordering of these videos alternated in the second half of the experiment.

3. RESULTS

The average responses to each audio-visual pair was as follows. Here, numbers between 0.0 and 1.0 describe the **preference levels**, where 0.0 means that the pair was never preferred to its counterpart, and 1.0 means always preferred.

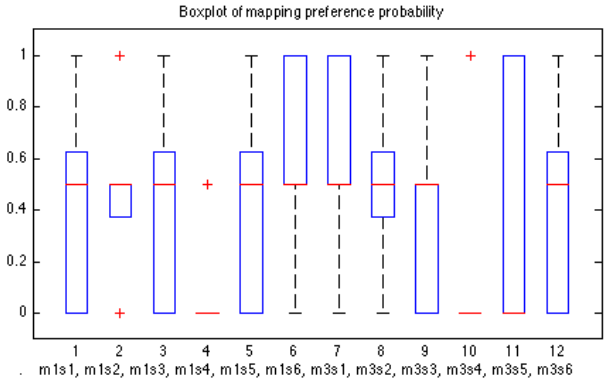

Figure 12. Box plots of the responses from all 9 subjects: 6 stimuli under Map 1 (left) and 6 stimuli under Map 3. Only Map 1 and Map 3 stimuli are shown, as [results(Map 2)]=[1−results(Map 1)], and [results(Map 4)]=[1−results(Map 3)].

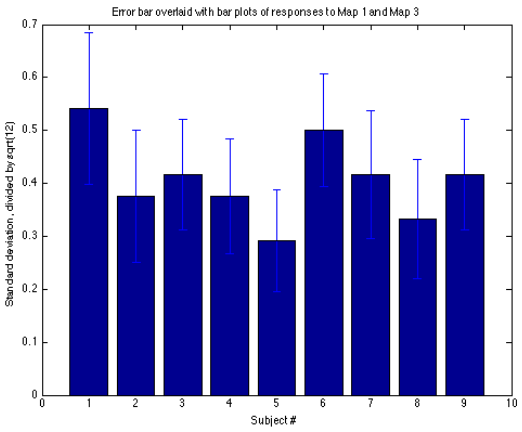

Figure 13. Error bar plot overlaid with the bar graph of the standard deviation of subjects' responses to Map 1 and Map 3.

The function RMAOV2 was also run on the data from Maps 1 and 3. This is what it output.

```
The number of IV1 levels are: 2

The number of IV2 levels are: 6

The number of subjects are: 9

Repeated Measures Two-Way Analysis of Variance Table.
------------------------------------------------------------------
SOV                SS       df        MS          F        P
------------------------------------------------------------------
Subjects         14.667      8       1.833[-inf   1.0000]
IV1               0.000      1       0.000          NaN      NaN
Error(IV1)        0.000      8       0.000
IV2               0.000      5       0.000       -8.000   1.0000
Error(IV2)       -0.000     40      -0.000
IV1xIV2           0.000      5       0.000       -8.000   1.0000
Error(IV1xIV2)   -0.000     40      -0.000
[Error           -0.000     88      -0.000]
Total            14.667    107
------------------------------------------------------------------
```

There were only 2 significant trials in this experiment, and they both pertain to Sound 4. Map 2 of Sound 4 was preferred 13 times over Map 1, while Map 1 was only preferred 1 time ($0.9286, p = 0.000044$). Similarly, Map 4 of Sound 4 was preferred 15 times over Map 3, while Map 3 was only chosen once ($0.9375, p = 0.0081$).

We can also see from Fig. 13 that Sound 2 was involved in both of the least significant stimuli. Referring to Fig. 2, I note that its frequency and harmonicity envelopes change very slowly, which must be why subjects didn't exhibit any preference for a mapping: there were no significant events in the sound to visualize.

I recorded each time a subject would repeat a trial, and have yet to interpret that data. However, this event happened quite infrequently. I also recorded "good ties" and "bad ties", an optional response that the subject could give in addition to a preference if he or she felt that the 2 videos were both poor or both great. These responses were completely noisy, with 24 indications of a "good tie" and 25 indications of a "bad tie," and with respect to trial, ties were again spread evenly, with "difficult" trials being judged as good as often as they were bad.

Each unique comparison of sounds to two different mappings was made twice. Both comparisons were included in the final analysis, but no bias was found for picking the first versus second video.[4]

Overall, Map 2 was preferred 64 out of 98 times to Map 1 (65%), and Map 4 preferred 56 times out of 98 to Map 3 (57%). When we omit Sound 4 from these numbers as an outlier, there is a much less dramatic difference; Map 2 is preferred 51 out of 88 times (58%) and Map 4 preferred 41 out of 82 times– exactly 50%.

Sound 4 has a particularly fast change in the harmonicity envelope around the middle of the clip (see Fig. 4). This produced a very marked change aurally

[4] 46.4% of the time, subjects chose the first video.

that caused Map 2 and Map 4 to dominate over Map 1 and Map 3.

Map 2 and Map 4 both have the harmonicity envelope mapped to radius and the amplitude envelope to intensity. This implies that these mappings are stronger than the opposite, mapping harmonicity to intensity and amplitude to radius as in Maps 1 and 3.

4. DISCUSSION

The experiment design was in many ways flawed. Findings were incomplete for this reason, and also because I only did one of the three possible combinations of mapping comparisons.

In my design of the visual mappings, it slipped my mind to assume the Doppler effect. Radius and transparency as an indicator of loudness is Doppler (though again, a combination of both makes more sense) and this was accounted for by all mappings, but frequency should also correspond to these features, as when sound objects move closer or farther they undergo shifts in pitch. More application of simple auditory phenomenon like this would serve the experiment design well.

The black fullscreen windows that displayed after each 5-second clip were there to try to isolate each pair/video during trials. Perhaps they were too long, as when I took it, occasionally I would feel as though I had completely forgotten the first mapping. This also makes a case for playing the videos simultaneously (since the sound is the same between them), one on the left and the other on the right, or even somehow overlaying them (and playing twice), decreasing the opacity of the "inactive" image.

As I stated in the Stimuli section, the numerical values of scalars and constants in the functions that translated audio to visual features were occasionally changed between mappings. I now believe that there should be some normalization of these functions to the ranges of the envelopes on a sound-to-sound basis, i.e., Max should send minimum and maximum data about the envelopes. This would force every visual feature to travel from the minimum and maximum of its specified range.

It then seems reasonable to design sounds that change often and to similar degrees, or in other words, to include more salient features within the sounds. One subject reported that she had preferences especially when they could clearly identify the mapping, and responses to Sound 4 verify this. It is possible that FM is not a good choice for this sort of experiment since modulation index and harmonicity contribute somewhat similarly to the spectral envelope, but it was quite easy to encapsulate its sound generation, and allowed explicit mappings with no spectral processing.

5. CONCLUSIONS & FUTURE WORK

The results[5] were statistically insignificant, so the experiment at large is inconclusive. Subjects showed a slight preference for Map 2 over Map 1 (57%), but since the difference between Map 3 and Map 4 was the same as the difference between Map 1 and Map 2, and Map 3 was preferred exactly the same number of times as Map 4, it is not possible to say that there exists the types of audio-visual connections I tried to make.

In spite of this (and considering the reactions to Sound 4), I do believe there is something here. Tweaking the audio stimuli, normalizing (and otherwise modifying) the mappings, generating trials that compare the other mappings to each other, and running more subjects should heed better results from this experiment, and I plan to host the experiment online in the near future so that I can potentially collect a lot of data.

6. REFERENCES

[1] S. Fu, "Audio/Visual Map With Cross-Modal Hidden Markov Models," *IEEE Transactions on Multimedia*, **7**(2), 2005, pp. 243–252.

[2] H. Deshpande, R. Singh, U. Nam, "Classification of Music Signals in the Visual Domain," *Proceedings of the COST G-6 Conference on Digital Audio Effects*, **1**, 2001, from http://www.csis.ul.ie/dafx01/proceedings/papers/unjung.pdf. Accessed 5 May 2013.

[3] G. Yu, J. Slotine, "Audio Classification from Time-Frequency Texture," from http://arxiv.org/pdf/0809.4501.pdf, 25 September 2008. Accessed 5 May 2013.

[4] Ninja Tune Record Label, "Amon Tobin 'ISAM' Live (Extended Trailer)," http://www.youtube.com/watch?v=WLrt7-kIgIM, May 31, 2011. Accessed 12 June 2013.

[5] K. Evans, A. Treisman, "Natural cross-modal mappings between visual and auditory features," *Journal of Vision*, **10**(1):6, 2010, pp. 1–12.

[5] In this conclusion, results from the outlier (Sound 4) are not considered.

[6] J. Chowning, "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation," *Journal of the Audio Engineering Society*,**21**(7), 1973, pp. 526–534.

[7] J. Ward, B. Huckstep, and E. Tsakanikos, "Sound-Colour Synaesthesia: to What Extent Does it Use Cross-Modal Mechanisms Common to us All?," *Cortex*, **42**(2), 2006, pp. 264–280.

[8] J. Ward, "Emotionally Mediated Synaesthesia," *Cognitive Neuropsychology*, **21**(7), 2004, pp. 761–772.

[9] M. Rizzo, P. J. Eslinger, "Colored hearing synesthesia: An investigation of neural factors," *Neurology*, **39**(6), pp. 781–784.