PERCEIVED MATCH BETWEEN VISUAL PARAMETERS AND AUDITORY CORRELATES: AN EXPERIMENTAL MULTIMEDIA INVESTIGATION

Scott D. Lipscomb & Eugene M. Kim

Northwestern University School of Music

ABSTRACT

This paper investigates the relationship between the auditory and visual components of an audio-visual (A-V) composite. Participants (N=28) were asked to rate the perceived degree of "match" between A-V components in a series of randomly presented composites. Manipulated audio parameters included pitch, loudness, timbre, and duration, while visual parameters included color, vertical location, shape, and size. A-V composites were created by combining all possible pairs of audio and visual stimuli using three different magnitudes of change, resulting in 48 A-V stimuli, i.e., 4 x 4 x 3. Data analysis revealed no statistically significant between-subjects differences as a result of either level of musical training or level of visual training. There was, however, a significant within-subjects difference in responses to the audio-visual pairs. Subject mean response ratings suggest the following primary relationships: pitch with vertical location, loudness with size, and timbre with shape. Duration did not pair as "best match" with any visual parameter. Results also revealed that color is equally matched with both pitch and loudness. The long-term goal of this project is to utilize the results of the music perception study reported as a basis for developing an algorithm to transform musical sound into visual animation, based on perceptually salient properties.

1. BACKGROUND

Many multimedia software programs are currently available that provide listeners with a visual analog to musical sound (e.g., Windows Media Player, iTunes, WinAmp, Sonique, etc.). However, the specific manner in which – as well as the basis for – transformation of the sonic component into a visual analog is rarely given much consideration. Though the most frequently used algorithm involves calculation of a fast Fourier transform of the complex audio signal and the mapping of these data onto screen coordinates, email communication with developers who created many of these programs reveals that the selection of visual parameters – color, shape, and motion – altered in association with various aspects of the musical sound – frequency, amplitude, and timbre – tends to be arbitrary. The present investigation sought to identify the degree of match between specific aspects of musical sound and those of animated visual images. The long-term goal of this project is to utilize the results of the music perception study reported herein as a basis for developing an algorithm to transform musical sound into visual animation that is based on perceptually salient properties.

Certain relationships are often assumed to exist. For example, in A-V contexts, the primary dimension of the pitch parameter is typically assumed to be pitch height (i.e., "high" and "low"). As a case in point, consider any "Mickey-moused" orchestral soundtrack to an animated cartoon (e.g., Wyle E. Coyote in *Road Runner* or Baby Herman in the opening sequence of *Who Framed Roger Rabbit?*). These vertical dimension terms are used to refer to pitch sequences that "ascend" or "descend," though there is no inherent justification for using these terms instead of others that could be just as readily apparent, e.g., horizontal dimension terms like "left" and "right." After all, this latter set of terms reflects more accurately, for instance, how the pitches lie across a piano keyboard.

2. SIGNIFICANCE OF THE STUDY

Historically, music has served to supplement or enhance visual perception of an image, as revealed by a careful examination of Hollywood motion pictures, television commercials, and music videos. The importance of musical soundtracks or background music in such contexts has been well documented (Bolivar, Cohen, & Fentress, 1994: Bullerjahn, Güldenring, & Hildesheim, 1994; Iwamiya, 1994; Lipscomb, in press; Lipscomb & Kendall, 1994; Marshall & Cohen, 1988; Thompson, Russo, & Sinclair, 1994). However, the relationship between specific musical features and their visual correlates has rarely been studied. One important exception to this is an investigation conducted by Robert Walker (1987) in which he attempted to determine "visual metaphors" for auditory stimuli.

Another significant aspect of the present study is its focus. From the outset, the intended outcome of this empirical investigation was to determine perceptual judgments of audio-visual match, the results of which will be used as a basis for creating an interface that allows audio-to-visual transformation based on these experimental results. This approach is dramatically different from – and, we would argue, significantly more valid than – that used by other audiovisual computer software applications that focus on the manipulation of computationally convenient quantities instead of perceptually meaningful phenomena.

3. RESEARCH QUESTIONS & HYPOTHESES

Two primary research questions were addressed. First, when changes in auditory and visual parameters are synchronized, are some pairings consistently perceived as a better match than other combinations? If so, which specific visual feature is "best match" for each of the selected auditory parameters? Based on previous research (Lipscomb, in press; Walker, 1987), we predicted that there would, in fact, be preferred pairings. Specifically, based on our own experience, we believed that frequency would be matched with

vertical placement, changes in amplitude would be matched with changes in size, and waveform would be matched both with color and pattern. It was predicted that duration will not be matched consistently with any single visual parameter.

Second, is the selection of visual correlates for auditory parameters influenced by either visual training or musical training? We predicted that both visual training and musical training would have a significant effect on subject ratings.

4. Method

The experimental investigation utilized a post-test only, repeated measures quasi-experimental design. Participants were 28 undergraduate and graduate students (ages 20-35) enrolled at Northwestern University.

Two between-subjects variables were considered: musical training and visual training. *Musical training* was operationally defined as "more than three years of formal musical instruction." *Visual training* was operationally defined as "one semester or more of training in drawing, computer graphics, animation, or film." According to these criteria, our sample of 28 participants were categorized as follows: 17 were musically trained (mean years of training = 14.12) and 11 musically untrained (mean years of training = .54); 14 were visually trained (mean years of training = 4.61) and 14 visually untrained (all with no formal training).

Three within-subjects variables were of interest: auditory parameters, visual parameters, and magnitude of change. Based on previous research (Lipscomb, 1995; Walker, 1987), specific attributes of the auditory and visual components were selected. Auditory parameters included pitch, timbre, loudness, and duration. Visual parameters included color, size, shape, and location. The amount of change in any of these auditory or visual attributes could be small, moderate, or large, as described below.

4.1 Stimuli

All stimuli were generated in a manner that allowed the investigators to control all aspects of the sound or image other than that being intentionally manipulated.

Auditory stimuli. The pitch, timbral, loudness, and durational attributes of the sonic materials were manipulated in the following ways, as represented in Table 1, found at the end of this paper.

*Pitch***.** Each tone lasted 950 ms, separated by 50 ms of silence. All sounds in a stimulus were generated with the same synthesized orchestral instrument timbre at the same loudness level, so the only difference between consecutive sounds was in pitch.

*Timbre***.** The specific choice of instrument timbres was based on the results of Kendall, Carterette, & Hajda (1999). To ensure generalizeability of the results from this previous study, all sounds for timbre change were sampled using the same factory presets on a Yamaha DX7 synthesizer. Each stimulus was comprised of three discrete sounds played using three distinct instrument timbres. Based on Kendall, et al.'s (1999) multidimensional scaling solution, the small change in timbre consisted of A3 played with flute, French horn, and Trumpet timbres at the same loudness and duration. Each note lasted 950 ms, separated by 50 ms of silence.

*Loudness***.** These stimuli consisted of three discrete sounds on the same pitch, increasing in energy by a factor of 15%, 30%, or 50%. All notes were played with the same instrument timbre and, similar to the other stimuli, each note lasted 950 ms, separated by 50 ms of silence.

*Duration***.** These stimuli consisted of three sounds on the same pitch, timbre, and loudness exhibiting an incremented duration utilizing a additive factor of 150 ms, 300 ms, or 400 ms.

Visual stimuli. Likewise, similar manipulations were utilized in the visual domain to create small, moderate, and large magnitudes of change. Variations were based upon light wavelength (color), pixel height and width (size), formal elements (shape), and y-coordinate on screen (location). Details related to the visual stimuli can be found in Table 2.

Audio-visual composites. A total of 48 A-V composites were created using Macromedia's Flash. The stimuli were distinguished one from another by three magnitudes of change (small, moderate, and large) within every composite, resulting in three versions of each sound-image pair: 3 pitch-color composites, 3 pitch-size composites, 3 pitch-shape composites, 3 pitch-location composites, 3 timbre-color composites, 3 timbre-size composites, etc.

4.2 Experimental Procedure

Subject responses were collected independently. In each session, a single participant completed the following series of tasks. First, the subject signed a consent form and provided basic demographic information, including years of musical training and years of visual training. Roger Kendall's *Music Experiment Development System* (MEDS, 2002) was used to present experimental stimuli and collect subject responses. Before beginning the main experiment, each participant read instructions presented on the computer screen and completed a warm-up experiment consisting of eight A-V examples to confirm their understanding of the task and response procedure. Subjects responded by controlling an on-screen scrollbar with the computer mouse, positioning the movable button between two ends labeled "unmatched" and "perfectly matched." The scrollbar was configured as an unmarked 101-point scale from 0 to 100.

5. RESULTS

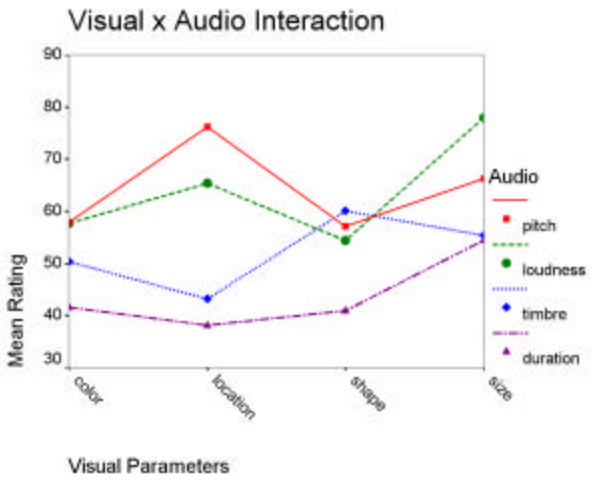
Data analysis revealed no statistically significant differences in the responses between groups [musical training ($F_{(1,24)} = 2.396$; $p = .135$), visual training ($F_{(1,24)} = 1.277$; $p = .270$)], nor in the interaction between musical and visual training ($F_{(1,24)} = .020$; $p = .889$). However, there were highly significant differences revealed by the within-groups comparisons: visual component ($F_{(3,72)} = 6.548$; $p = .001$), audio component ($F_{(3,72)} = 36.142$; $p < .0005$), and the interaction between audio and visual components ($F_{(9,216)} = 8.259$; $p < .0005$). Interestingly, there was no main effect of

magnitude of change, nor was there any statistically significant interaction involving this variable.

6. CONCLUSIONS

Experimental results reveal that there are, in fact, differences in the degree of "match" perceived by subjects, depending on the audio and visual components of an A-V composite (Figure 1), but these ratings do not vary significantly as a result of either musical or visual training. The highest ratings of perceived match suggest the following pairings: pitch-location, loudness-size, and timbre-shape. Also emerging, however, is evidence of non-unitary relationships. For example, the visual attribute of color matched equally well with both pitch and loudness in the auditory domain. It is worthy of note that duration did not pair as "best match" with any visual parameter. Finally, in addition to the relationship of both pitch and loudness to color cited above, there are several instances in which secondary relationships suggest that the primary relationships mentioned previously do not present a singular appropriate "matched" pairing. For example, though the primary pairing receives a higher mean score, secondary relationships (such as loudness-location or pitch-size) may provide a source of variation that can be integrated into the audio-to-visual algorithm. Also, the auditory parameters of timbre, pitch, and loudness may all provide an acceptable matched pairing for the visual attribute of shape.

Figure 1. Graph of subject mean responses.



These results differ somewhat from those obtained by Walker (1987). The present study confirms the relationship between pitch and vertical location, loudness and size, and timbre and shape. In the context of the present study, duration was not a primary match for any of the visual attributes manipulated. However, Walker's subjects mapped changes in duration to horizontal length, an attribute not manipulated in the present study. In the present study, color – an attribute not manipulated by Walker – was equally matched with both pitch and loudness. Also, results of Walker (1987) revealed significant between-groups differences as a result of musical training and age, while cultural and environmental effects were observed in subjects who had less exposure to Western lifestyles. In the present study, neither musical training nor visual training was shown to be a statistically significant factor. It is possible that the difference in method could be responsible for these discrepancies. Walker's subjects listened to a sound and selected a "visual metaphor" in a 4-level forced choice procedure, whereas, in the present study, subjects were exposed to a series of A-V stimuli in a random order and provided a rating for the "degree of match" between the audio and visual components.

In closing, the fact that musical training and visual training did not significantly affect subject ratings in the present study, suggests that there may be an implicit, culturally-determined agreement regarding these audio-visual relationships, but future research will be required to confirm such a claim and to address the matter of disagreement between this study and Walker (1987) in reference to musical training. The next step in this research study will be the development of a transformation algorithm utilizing the results of this perceptual experiment as its basis. A software interface will then be developed, performing perceptually meaningful visual transformations of a selected digital audio input.

7. ADDITIONAL FILES

Four animation files are included on this CD-ROM to illustrate the types of A-V stimuli to which participants were responding. All stimuli cited in this study are available via the following URL: http://faculty-web.at.northwestern.edu/music/lipscomb/stimuli/.

The filenames of the immediately available animations are:

1. **pitch_color_large.mpg**
2. **pitch_location_large.mpg**
3. **timbre_shape_large.mpg**
4. **loudness_size_large.mpg**

8. REFERENCES

Bolivar, V.J., Cohen, A.J, & Fentress, F. C. (1994) Semantic and formal congruency in music and motion pictures: Effects on the interpretation of visual action. *Psychomusicology*, 13, pp. 28-59.

Bullerjahn, C., Güldenring, M., & Hildesheim, U. (1994) An empirical investigation of effects of film music using qualitative content analysis. *Psychomusicology*, 13, pp. 99-118.

Iwamiya, S. (1994) Interactions between auditory and visual processing when listening to music in an audio visual context: 1. matching 2. audio quality. *Psychomusicology*, 13, pp. 133-154.

Kendall, R.A. (2002). Music Experiment Development System. Los Angeles: University of California, Los Angeles. [computer software]

Kendall, R.A., Carterette, E.C., & Hajda, J.M. (1999). Perceptual and acoustical features of natural and synthetic orchestral instrument tones. *Music Perception*, 16(3), pp. 327-364.

Lipscomb, S.D. (in press) The perception of audio-visual composites: Accent structure alignment of simple stimuli. *Selected Reports in Ethnomusicology*, 12.

Lipscomb, S.D. & Kendall, R. A. (1994) Perceptual judgement of the relationship between musical and visual components in film. *Psychomusicology*, 13, pp. 60-98.

Lipscomb, S.D. (1995). Cognition of musical and visual accent structure alignment in film and animation. Unpublished doctoral dissertation. Los Angeles: University of California, Los Angeles.

Marshall, S.K. & Cohen, A.J. (1988). Effects of musical sound-tracks on attitudes toward animated geometric figures. *Music Perception*, 6, 95-112.

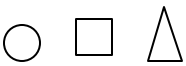
Thompson, W.F., Russo, F.A., & Sinclair D. (1994) Effects of underscoring on the perception of closure in filmed events. *Psychomusicology*, 13(9), pp.9-27.

Walker, Robert. (1987) The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. *Perception & Psychophysics*, 42(5), pp. 491-502.

Table 1. Descriptions of the auditory stimuli. [In the Timbre descriptions, the following abbreviations were used to describe orchestral instrument timbres used: FL= flute, FH=French horn, TR=trumpet, CL=clarinet, VN=violin, and OB=oboe.]

	Small	**Moderate**	**Large**
Pitch	C3-D3-E3-D3-C3	C3-E3-G3-E3-C3	C3-G3-C4-G3-C3
Timbre	FL-FH-TR-FH-FL	FL-CL-TR-CL-FL	FL-VN-OB-VN-FL
Loudness	15% louder than the previous tone	30% louder than the previous tone	50% louder than the previous tone
Duration	350-500-650-500-350ms	200-500-800-500-200ms	100-500-900-500-100ms

Table 2. Descriptions of the visual stimuli.

	Small	**Moderate**	**Large**
Color	Dark blue-blue-light blue	Blue-indigo-violet	Red-green-violet
Size	120% change from preceding	150 % change	200% change
Shape			
Location