Michael Quigley
Psychology of Music
5/20/13

Final Report

I. Abstract

Over time instrumentalists come to understand the distinctions between primary physical actions that produce sound and ancillary actions that aid in expressing musical or emotional information, which can be vital to audience understanding. It is possible that the loss of haptic feedback makes the properties of a musical instrument less understandable to both performer and audience. Perhaps then visual feedback can allow an instrumentalist to better understand the limitations of his or her instrument when no such haptic feedback exists, as is the case with many musical systems that incorporate live video interaction. This experiment tested the hypothesis that, in a video camera controlled software instrument, the presentation of visual feedback will aid in the users' understanding of musical mappings to visual space. Participants were familiarized with such a software instrument and asked to recreate recordings of changes in pitch and timbre within timed trials. A visual feedback display was designed and presented only to an experimental group. For three separate recordings, each changing in one parameter (pitch or timbre), participants of the experimental group, on average, performed faster than those of the control group. These findings imply that visual feedback does allow an instrumentalist to better understand the limitations of his or her instrument.

II. Introduction

The multimodal characteristics of music have been a topic of great interest for researchers in music cognition, composers, performers, sound designers, and artists. Live performance has held an important role in many cultures, socially and religiously. Both musician and audience perceive music in a number of sensory modes, simultaneously and fused together to form a cohesive experience. Musicians take in to account the sound, feel, and look of an instrument when deciding its appropriateness for a particular performance. Audience members pay attention to performers' body movements and interactions with their instruments in addition to the sound produced. In any perceptual experience a multitude of sensory interactions occur for the individual to process and integrate in some fashion.

One area of multimodal perception that has had extensive research is the relationship between auditory and visual perception. In order for auditory and visual stimuli to be perceptually integrated as a multimodal event there needs to be some level of synchrony between them. When watching a film, the movement of an actors mouth must match the dialogue heard in order for the viewer to associate each sensory stimuli as a single event: speaking. In order for this sense of matching to occur, whether in a recreation of human experience (dialogue in film) or an artistic work, the auditory and visual stimuli must have complementary attributes, allowing for what Iwamiya (2004) calls "cooperative interaction" in which each sensory mode strengthens the impression of the other. One of the features responsible for a sense of matching between audio and visual stimuli is the perceived synchrony of the two events. Given an audiovisual sequence to be observed, Cook, Van Valkenburg, and Badcock (2011) found that changes in the predictability of pitch and

temporal density changed how subjects perceived the synchrony of the audiovisual events. In reducing subjects' ability to perceive the coming pitch and temporal pattern in a sequence, they were more likely to judge the audiovisual events as synchronous. Extrapolating, this might suggest that making it more difficult to predict musical features causes the observer to broaden the range of what he/she finds to be synchronous. When musical features are predictable, the range of expectation is narrowed, allowing the observer to be more stringent about what is synchronous. This would have influence on how a composer creates an audiovisual work or how an engineer might design a piece of software for an installation.

Studies have also been done to examine the perceptual influences of one modality on another. An influence of auditory information on visual perception is likely familiar with most people. It is widely utilized in film scoring to strengthen the intended emotional content of a scene. Less investigated is the influence of visual information on auditory perception. To explore this influence in one study, researchers paired images of positive or negative content with relatively ambiguous melodies and found that the addition of such images caused the subjects to describe each melody more similarly to whichever quality of image it was presented with than when presented without an image (Boltz, Ebendorf, & Field, 2009). A particularly interesting aspect of these results is that auditory and visual stimuli appear to influence the perception of each other in a somewhat symmetrical way, at least at the level of emotional complexity. Sound can influence intended emotion of an image, and image can influence the intended emotion of sound.

Work has also been done to suggest that relationships exist at the perceptual level for observing physical properties. Bruns and Getzmann (2008) found that sound can influence the temporal perception of motion in a visual field. By presenting a sound between the flashing of two lights, spaced horizontally apart, they were able to elicit from the subject a perception of motion in the light, as opposed to two discrete flashing events. Not only, then, is integrated auditory and visual perception capable of conveying and manipulating higher level features like emotion, it also can affect the perception of physical motion. Research by Effenberg (2005) on the information conveyed through multimodal presentation supports the importance of auditory and visual representation of events. In having subjects evaluate a video of an athletic jump and recreate the jump themselves, participants were more successful in both tests after having viewed an audiovisual representation, in which the athlete's jump was sonified. The results then imply that information about a physical property (height) can be represented in an auditory fashion. Further, humans are capable of extracting that information even when the sound is imposed on the event, as opposed to extracting information from the captured acoustic events.

Acoustic events themselves carry information regarding other modal properties. Similarly to Effenberg, Grassi (2005) found that subjects were able to extract information regarding the size of two physical objects from the sound produced in a collision between the two, the important distinction being that in Grassi's study the audio presented was an actual acoustic event, and in Effenberg's study the audio was a sonification of the athlete's jump. Humans then appear to posses some level of perceptual abstraction that allows sensory information to be understood and represented amodally, under certain conditions. As in the understanding of spatial properties through sonification in Effenberg's study, there would appear to be some human capacity for sensory substitution, in which sensory information can be obtained through the input of a different sensory modality (Bach-y-Rita, 1969 and Bach-y-Rita, 1972 as cited in Kim and Zatorre, 2010). Kim and Zatorre (2010) combined crossmodal learning and sensory substitution in a study where subjects were able to convey information about physical shape through auditory stimuli, provided spatial

relation was preserved in a systematic manner. In these three studies (Effenberg, Grassi, and Kim & Zatorre), humans were able to draw information about spatial properties from auditory stimuli.

As was the case with Boltz, Ebendorf, and Fields' image/melody pairings, humans are capable of drawing information from visual stimuli that influences auditory perception, even of higher level complexities like those found in music. Emotions, attitudes, and musical attributes can be displayed in the physical interaction of the musician and the instrument. For instance in an instrumental performance the musician expresses cues for musical features in his/her body movements and facial expressions. In a study by Schutz and Lipscomb (2007), the perceived length of notes played on a marimba were effectively altered by the performer's musical gesture. When audio of a short percussive strike was spliced with the video of a longer striking gesture, subjects judged the sound accompanying the longer gesture to be longer, despite that not actually being the case. Furthermore, Schutz and Lipscomb's work involved musically trained and non-musically trained participants, suggesting that visual influence over audition is a "real-life cross-modal interaction."

Other performance characteristics are conveyed through holistic body motion, with the listener actively observing the multi-sensory experience but not attending to any one feature in particular. Some musical properties can be conveyed through the ancillary body movement of the musician, like phrasing and expressiveness (Nusseck and Wanderley, 2009). Similarly Vines, Krumhansl, Wanderley, & Levitin (2006) investigated the role of a performer's movement on the perception of musical performance, particularly by trained musicians. They found that visual stimuli of the performer helped to increase and decrease perceived tension at different points in the piece, as well as increase the clarity of phrasing throughout. These movements are paramount to the performance, as they are often what helps to convey the musical features that may not be written in a score and are brought forth in the characteristics of the individual performer.

In trying to understand the representation of music in visual space, Evans and Treisman (2010) conducted a study to determine if relationships exist between auditory and visual features at the perceptual level as opposed to arbitrarily decided mappings. They found evidence for perceptual level associations of pitch and visual position, pitch and spatial frequency, and pitch and size. Eitan and Granot (2006) examined listener perception of the relationship between musical parameters and motion, concluding that perceptual mapping of music in space and motion is multifaceted, with musical features often corresponding to more than one field of motion and there existing an asymmetry in many musical-spatial relationships. Causes for the particular perceptual mappings and associations discussed above are examined by Walker (1987), who studied the role of enculturation and training on crossmodal analogies and found that common associations are only specific to Western cultures, suggesting the possibility of alternate analogies. This may appear contradictory to the results of Schutz and Lipscomb, which suggested an audiovisual link between gesture and temporal perception regardless of musical training. However, the participants in that study were all students from Northwestern University and could presumably have had extensive experience with Western musical performance regardless of individual training.

With widespread interest in audiovisual relationships, there has been a considerable amount of experimental and artistic work done using systems that incorporate video cameras as input parameters and controllers. With each system comes the need to define how and why a visual feature will be mapped to a musical parameter, or vice versa. Within that framework there is also the distinction of gesture within the visual field and how any given

gesture should be interpreted. Schacher (2012) presents a discussion of gesture in electronic music performance and the distinction between motion and movement, in which he addresses that gesture cannot be measured but must be the interpretation of "measured movements and actions with a set of mental and cultural criteria." This set of clearly defined criteria is necessary for building any system that effectively integrates both auditory and visual modalities. A general survey of contemporary video based computer music systems would give a picture of many loosely defined systems designed for the flowing, broad movements of dancers and full stage performers. However, there appears to be a lack of systems designed for the instrumentalist setting in which a performer might interact with a video controlled synthesis engine, for example. The loss of haptic feedback when using a video source for control of a system, something that would surely affect a performer's interaction with his/her instrument, appears to be a major obstacle for designing intuitive computer music instruments. Brent (2011) states, "Whether one chooses to highlight, de-emphasize, or ignore it, the issue of physicality is of central importance to the performance of computer-based music." Perhaps the human capability for sensory substitution and our strongly integrated relationship between auditory and visual perception would allow for a system that incorporates visual feedback for the user in place of haptic feedback.

III. Methodology

Participants

Participants were 15 adults between 23 and 47 years old, with varying degrees of musical training. Musical training was determined by taking the Ollen Musical Sophistication Index (OMSI) of each subject. Participants were separated into an experimental group and a control group. The experimental group had a mean age of 29.8 years old and a mean OMSI of 876.1. The control group had a mean age of 27.1 years old and a mean OMSI of 342.7. All of the participants reported not to have Absolute Pitch. One participant in the control group reported having hearing impairments, but did not elaborate on the degree of severity. On a 7-point scale, the experimental group had a mean familiarity with Human Computer Interaction of 6, while the control group had that of 4.

Materials

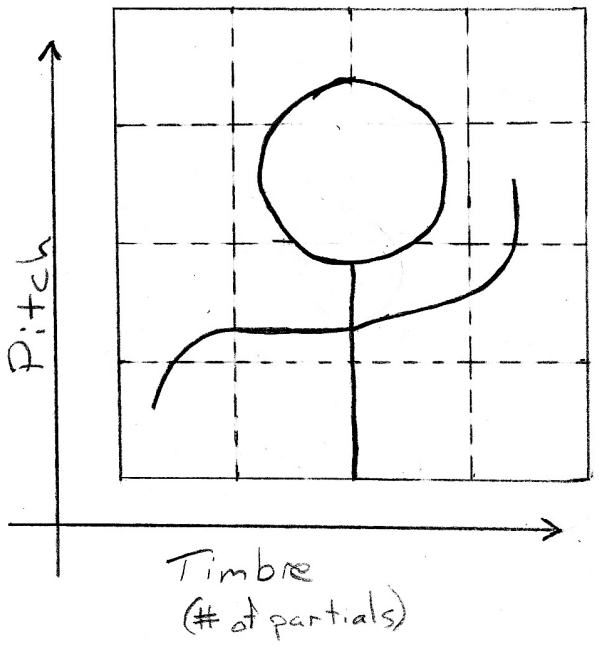
Stimuli for the study included three audio recordings of musical gestures created using the software instrument. Each audio recording was of a constant pitch with 4 changes in timbre, or of a constant timbre with 4 changes in pitch. Stimulus 1 was a constant A440 that changed from a sine wave to a square wave in 4 equal steps. Stimulus 2 was a constant sine wave that increased in pitch from A440 to C#554 to E659 to A880. Stimulus 3 was a constant square wave that increased in pitch from A440 to C#554 to E659 to A880. Participants were also exposed to the audio synthesized by the subject's own motion within the field captured by a video camera and processed in Max/MSP/Jitter in real time. The device produced a constant, monophonic tone whose pitch and timbre would change when motion was detected in a particular region of the camera field. Those mappings are explained below:

Vertical Mappings:

top – A 880Hz
mid high – E 659Hz
mid low – C# 554Hz
bottom – A 440Hz

Horizontal Mappings:

left – sine wave
mid left – 2/3 sine wave+ 1/3 square wave
mid right – 1/3 sine wave + 2/3 square wave
right – square wave



In this case, motion in the lower left of the grid would cause the tone produced to be of a sine wave timbre at 440Hz (A4). Motion in the upper right of the grid would cause the tone produced to be of a square wave timbre at 880Hz (A5).

Additionally, the experimental group was exposed to a mirror image of the video input of the built-in laptop video camera, presented on a laptop screen, that was dynamically altered by the subject's particular interactions with the instrument. As the synthesized tone increased in pitch, the screen would be increasingly tinted red. As the synthesized tone increased in timbral complexity, from sine wave to square wave, the screen would increase in brightness. This is the visual feedback component that is used to test the hypothesis against the control group.

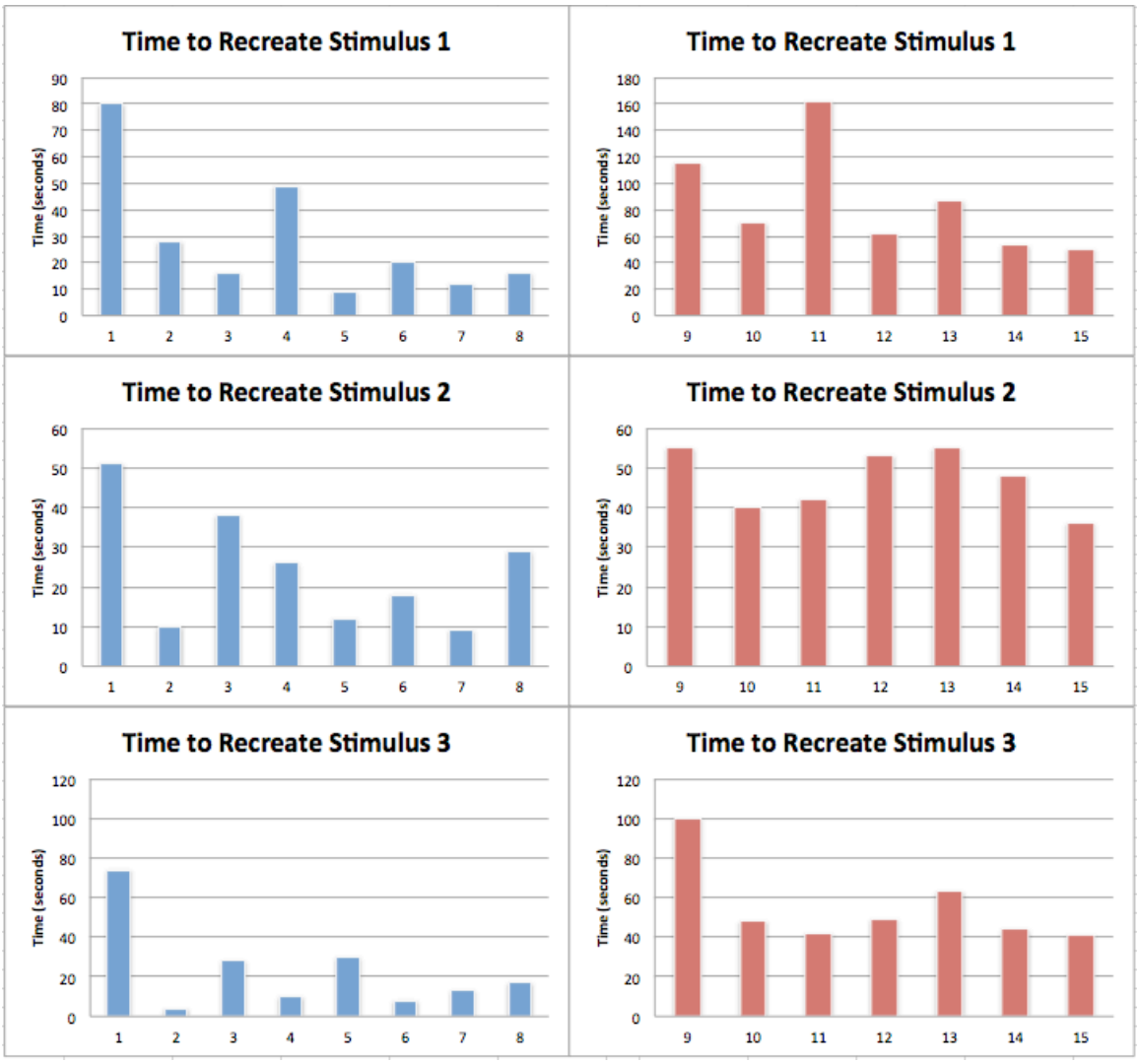
Procedure

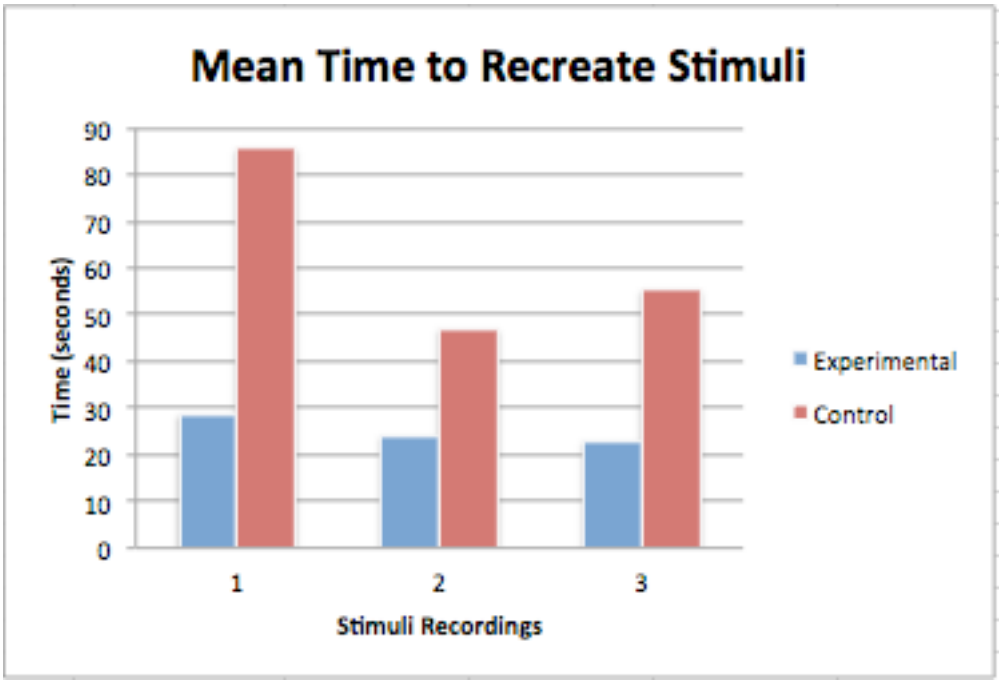
The experiment was conducted individually for each participant. Participants sat at a table with a MacBook Pro laptop hosting the Max/MSP/Jitter software instrument and video camera, and a pair of over-ear headphones. A pre-test survey was given to gather information about each participant. The experimenter then explained that the device would produce a constant tone whose parameters would be altered by the participant's motion. Participants were then given a 1-minute familiarization period to freely interact with the device. Those in the experimental group were exposed to the visual feedback component of the laptop screen; those in the control group were not. After familiarization, participants were played a one of the stimuli recordings. The order of stimuli recordings presented was cycled throughout each participant. The participant was free to listen to the stimulus several times until comfortable in their understanding of the changing musical parameters, at which point he or she would attempt to recreate those changes in the order they occur in the recording, while disregarding tempo. Tempo was not a factor because the test was only interested in participants' understanding of pitch and timbre. The experimenter recorded the amount of time is took until the participant successfully produced the sequential changes in musical parameters that corresponded to the stimulus recording. This process was repeated for all 3 stimuli recordings (exposure to stimulus, recreation by participant). After the three timed trials the participant was given a post-test survey to assess his or her understanding of the relationships between position of motion and musical parameters of the audio produced. Participants in the experimental group were given additional questions regarding the visual feedback presented to them.

IV. Results

Participants of the control group had a mean OMSI of 470.7 (*SD* = 419.4), mean age of 27.1 (*SD* = 5.3), and familiarity with Human Computer Interaction (HCI) of 5 (*SD* = 2.3) on a 7-point rating scale, while the experimental group had a mean OMSI of 876.1 (*SD* = 139.4), mean age of 29.8 (*SD* = 8.2), and familiarity with HCI of 5 (*SD* = 2.4).

For all 3 stimuli tests, participants in the experimental group had lower mean recreation times than those of the control group. For Stimulus 1, the control group had a mean of 86 seconds with *SD* = 40.5, while the experimental group had a mean of 29 seconds with *SD* = 24.2. For Stimulus 2, the control group had a mean of 47 seconds with *SD* = 7.7, while the experimental group had a mean of 24 seconds with *SD* = 14.9. For Stimulus 3, the control group had a mean of 55 seconds with *SD* = 21.1, while the experimental group had a mean of 23 seconds with *SD* = 22.6.





All participants in both groups correctly identified the spatial axes along which changes in pitch and timbre separately occurred. Participants of the experimental group had higher intervallic ratings of their understanding of the relationship between pitch and its spatial equivalent (M = 6.1, SD = 0.8) than those in the control group (M = 3.7, SD = 1.4). Participants of the experimental group also had higher intervallic ratings of their understanding of the relationship between timbre and its spatial equivalent (M = 5.4, SD = 1.9) than those in the control group (M = 2.9, SD = 0.9). Participants of the control group rated the appropriateness of the audio-spatial relationships for a musical context higher (M = 5.6, SD = 1.3) than those of the control group (M = 5, SD = 1.2).

Additionally, 62.5% of participants in the experimental group correctly identified the visual feedback component that corresponded to change in pitch (red tint). 62.5% of participants in the experimental group correctly identified the visual feedback component that corresponded to change in timbre (brightness). 50% of participants in the experimental group correctly identified both visual feedback components, while 25% identified one but not the other, and 25% incorrectly identified both visual feedback components. Experimental participants also rated how related the visual display was to the audio output (M = 4.4, SD = 1.4); how well he or she understood the pitch-visual relationship (M = 5.1, SD = 1.4); how well he or she understood the timbre-visual relationship (M = 5.1, SD = 1.1); the helpfulness of the visual display in understanding how audio parameters were mapped to participants' interaction (M = 3.8, SD = 2.3); and the appropriateness of the relationship between audio output and visual feedback for a musical setting (M = 4.4, SD = 1.5);

V. Discussion

On average, those participants given access of the visual display performed better than those that were not. Additionally, they gave higher ratings for their understandings of audio-spatial relationships. These results infer the hypothesis that the presentation of visual feedback in a video camera controlled software instrument does aid in the users' understanding of musical mappings to visual space. However, it is important to note the differences in Ollen Musical Sophistication Indexes across groups. The experimental group had both a higher Median (916.5) and Mean (876.1) than the control group Median (222) and Mean (470.7). It is possible that the musical sophistication of each participant affected his or her ability to recreate changes in audio parameters. Thus, it is possible that higher average musical sophistication accounts, to some degree, for the faster recreation times of the experimental group. The researcher feels that the results of this study are still highly valuable in understanding the role of visual feedback in video camera controlled software instruments. Experimental participants provided additional ratings of the visual components with 50% accurately describing the visual equivalents of pitch and timbre, and 62.5% getting at least one component. Further research can be done to determine the usefulness of the particular mappings utilized in this study and suggest more intuitive and/or musical mappings. An improved version of this study could also be conducted with more evenly distributed musical sophistication ratings of participants across the control and experimental groups. Additionally, separate studies could be conducted with only participants above or below a certain musical sophistication index.

VI. Annotated Bibliography

Bach-y-Rita P (1972) Brain mechanisms in sensory substitution. Academic Press, London

Bach-y-Rita P, Collins CC, Saunders FA, White B, Scadden L (1969) Vision substitution by tactile image projection. *Nature, (221)*, 963–964

Boltz, M. G., Ebendorf, B., & Field, B. (2009). Audiovisual interactions: The impact of visual information on music perception and memory. *Music Perception, 27(1),* 43–59.

Brent, W. (2011, August). *Aspects of gesture in digital musical instrument design.* Proceedings from the International Computer Music Conference. University of Huddersfield, UK.

Bruns, P., & Getzmann, S. (2008). Audiovisual influences on the perception of visual apparent motion: exploring the effect of a single sound. *Acta Psychologica, 129(2),* 273–83.

Cook, L. a, Van Valkenburg, D. L., & Badcock, D. R. (2011). Predictability affects the perception of audiovisual synchrony in complex sequences. *Attention, Perception & Psychophysics, 73(7),* 2286–97.

Effenberg, A. O. (2005). Movement sonification: Effects on perception and action. *IEEE Multimedia, 12(2),* 53–59.

Eitan, Z., & Granot, R. Y. (2006). How music moves. *Music Perception, 23(3),* 221–248.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features, *Journal of Vision, 10,* 1–12.

Grassi, M. (2005). Do we hear size or sound? Balls dropped on plates. *Perception & Psychophysics, 67(2),* 274–84.

Iwamiya, S. (1994). Interactions between auditory and visual processing when listening to music in an audiovisual context: 1. Matching 2. Audio quality. *Psychomusicology: A Journal of Research in Music Cognition, 13(1-2),* 133–153.

Kim, J.-K., & Zatorre, R. J. (2010). Can you hear shapes you touch? *Experimental Brain Research, 202(4),* 747–54.

Nusseck, M., & Wanderley, M. (2009). Music and motion - How music-related ancillary body movements contribute to the experience of music. *Music Perception*, 335–354.

Ollen, Joy. Ollen Musical Sophistication Index. MARCS Auditory Laboratories, n.d. Web. 15 May 2013.

Schacher, J. (2012, July). *The body in electronic music performance*. Proceedings from the 9th Sound and Music Computing Conference. Copenhagen, Denmark.

Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception, 36(6),* 888–897.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross modal interactions in the perception of music performance. *Cognition, 101,* 80–113.

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