THE PERCEPTION OF AUDIO-VISUAL COMPOSITES: ACCENT STRUCTURE ALIGNMENT OF SIMPLE STIMULI

Scott D. Lipscomb
Northwestern University

In contemporary society, the human sensory system is bombarded by sounds and images intended to attract attention, manipulate state of mind, or affect behavior.[1] Patients awaiting a medical or dental appointment are often subjected to the "soothing" sounds of Muzak as they sit in the waiting area. Trend-setting fashions are displayed in mall shops blaring the latest Top 40 selections to attract their specific clientele. Corporate training sessions and management presentations frequently employ not only communication through text and speech, but a variety of multimedia types for the purpose of attracting and maintaining attention, e.g. music, graphs, and animation. Recent versions of word processors allow the embedding of sound files, animations, charts, equations, pictures, and information from multiple applications within a single document. Even while standing in line at an amusement park or ordering a drink at the local pub, the presence of television screens providing aural and visual "companionship" is now ubiquitous. In each of these instances mentioned, music is assumed to be a catalyst for establishing the mood deemed appropriate, generating desired actions, or simply maintaining a high level of interest among participants within a given context.

Musical affect has also been claimed to result in increased labor productivity and reductions in on-the-job accidents when music is piped into the workplace (Hough, 1943; Halpin, 1943-4; Kerr, 1945), though these studies are often far from rigorous in their method and analysis (McGehee & Gardner, 1949; Cardinell & Burris-Meyer, 1949; Uhrbock, 1961). Music therapists claim that music has a beneficial effect in the treatment of some handicapped individuals and/or as a part of physical rehabilitation following traumatic bodily injury (Brusilovsky, 1972; Nordoff & Robbins, 1973; an opposing viewpoint is presented by Madsen & Madsen, 1970). Individuals use music to facilitate either relaxation or stimula-

---

[1] The author would like to acknowledge the assistance of both the University of California, Los Angeles and The University of Texas at San Antonio. In addition, the support of Northwestern University has been integral to continuing research efforts. Without the use of the research facilities provided, funding for necessary equipment, and colleagues with whom the results could be discussed, completion of this project would not have been possible. I would especially like to thank Dr. Roger A. Kendall for his invaluable assistance.

tion in leisure activities. With the increase in leisure time during the 1980s (Morris, 1988), many entertainment-related products began to utilize music to great effect in augmenting the aesthetic affect of these experiences. Executives of advertising agencies have realized the impact music has on attracting a desired audience, as evidenced recently by the use of classic rock songs to call baby-boomers to attention or excerpts from the Western art music repertoire to attract a more "sophisticated" audience.

One of the most effective uses of music specifically intended to manipulate perceptual response to a visual stimulus is found in motion pictures and animation. The present study investigated the relationship of events perceived as salient (i.e., accented), both aurally and visually. As a result, this study focused on an aspect of the motion picture experience that had never before been addressed explicitly in music perception literature. Many studies had examined associational and referential aspects of both sound and vision. Some investigations had even examined explicitly the relationship of music to visual images in the context of the motion picture experience. However, none have proposed an explicit model based on stratification of accent structures or set out to test the audio-visual relationship on the basis of accent structure alignment.

Before considering the specific interrelationship between the aural and visual components of animated sequences, several issues were carefully examined. First, what are the determinants of "accent" (i.e. points of emphasis) in the visual and auditory fields?; and second, is it *necessary* for accents in the musical soundtrack to line up precisely with points of emphasis in the visual modality in order for the combination to be considered effective? The ultimate goal of this line of research is to determine the *fundamental principles governing interaction* between the auditory and visual components in the motion picture experience.

RELATED LITERATURE

To the present, there has been little empirical work specifically directed at studying the symbiotic relationship between the two primary perceptual modalities normally used in viewing films (Lipscomb, 1990; Lipscomb & Kendall, 1996). In the field of perceptual psychology, interaction between the aural and visual sensory modalities is well-documented (see, for example, Radeau & Bertelson, 1974; Staal & Donderi, 1983; Bermant & Welch, 1976; Ruff & Perret, 1976; Massaro & Warner, 1977; Regan & Spekreijse, 1977; and Mershon, Desaulniers, Amerson, and Kiever, 1980). For a detailed discussion of film music research (e.g., Tannenbaum, 1956; Thayer & Levenson, 1984; and Marshall & Cohen, 1988), see Lipscomb (1995) and Lipscomb & Kendall (1996). The latter paper was included in a special issue of *Psychomusicology* (vol. 13) devoted to the topic

the topic of Film Music research, including investigations by a wide array of scholars (Thompson, Russo, & Sinclair, 1996; Bolivar, Cohen, & Fentress, 1996; Lipscomb & Kendall, 1996; Bullerjahn & Güldenring, 1996; Sirius & Clarke, 1996; Iwamiya, 1996; and Rosar, 1996). Though the list is not long, there have been many approaches to the study of combined sound and image. Marilyn Boltz and her colleagues have investigated the relationship between the presence of musical sound and memory for filmed events and their duration (e.g., Boltz, 2001; Boltz, 1992; and Boltz, Schulkind, & Kantra, 1991). Krumhansl & Schenck (1997) investigated the relationship between dance choreography by Balanchine and the music by which it was inspired, Mozart's *Divertimento No. 15*. In a study by Vitouch (2001), after seeing a brief film excerpt with one of two contrasting musical soundtracks, subjects provided a written prediction of how the plot would continue, revealing that anticipations of future events are "systematically influenced" by the accompanying musical sound (p. 70). None of these investigations, however, addressed the synchronization between the musical and visual components of the motion picture experience.

Proposed Model and Its Foundation

What is the purpose of a musical soundtrack? An effective film score, in its interactive association with the visual element, need not attract the audience member's attention to the music itself. In fact, the most successful film composers have made a fine art of manipulating audience perception and emphasizing important events in the dramatic action without causing a conscious attentional shift. When watching a film, a typical audience member's perception of the musical component often remains at a subconscious level (Lipscomb, 1989).

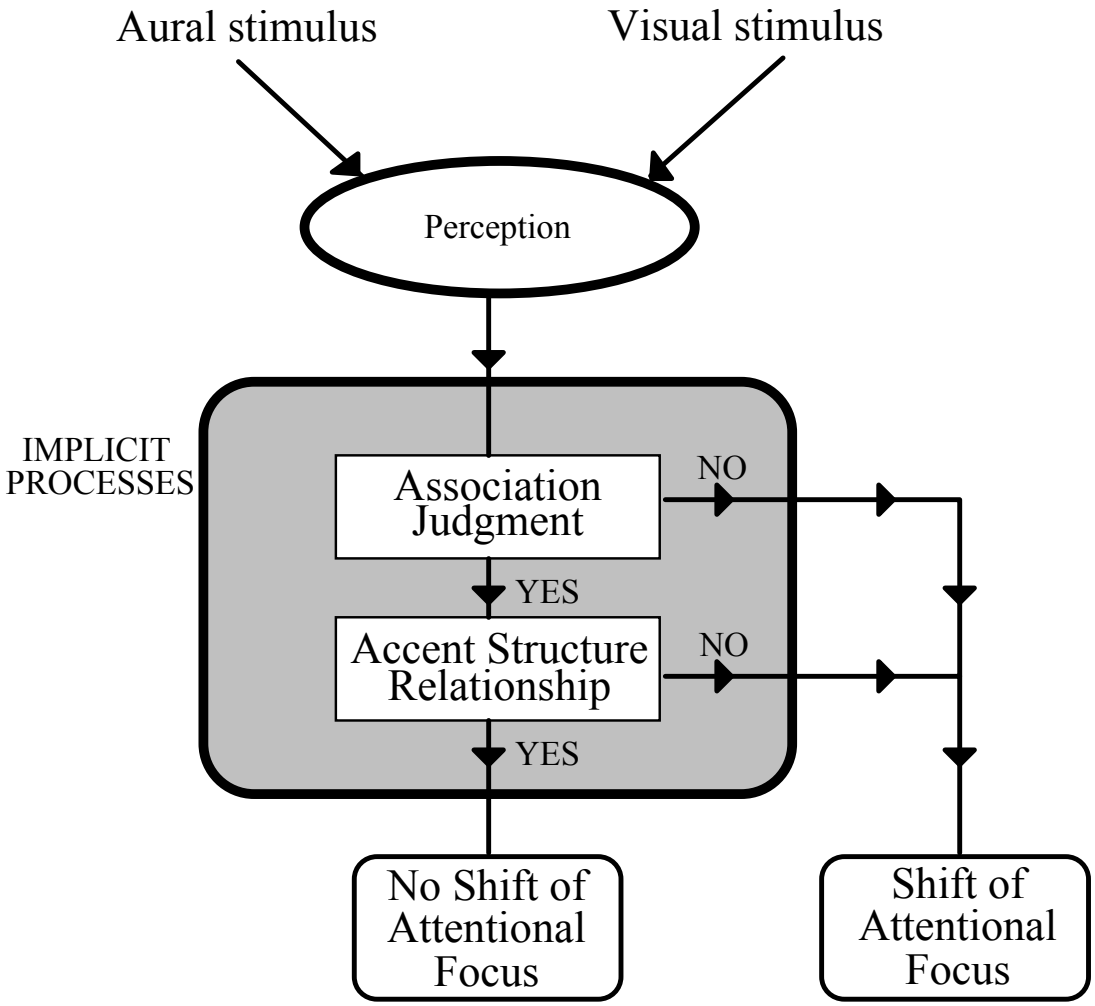
Marshall & Cohen (1988) provided a paradigm to explain the interaction of musical sound and geometric shapes in motion entitled the "Congruence-Associationist model." They assumed that, in the perception of a composite AV presentation, separate judgments were made on each of three semantic dimensions (i.e. Evaluative, Potency, and Activity; see Osgood, Suci, & Tannenbaum, 1957) for the music and the film, suggesting that these evaluations were then compared for congruence at a higher level of processing.

A model proposed by Lipscomb & Kendall (1996) suggests that there are two implicit judgments made during the perceptual processing of the motion picture experience: an association judgment and a mapping of accent structures (see Figure 1). The association judgment relies on past experience as a basis for determining whether or not the music is appropriate within a given context. For example, a composer may have used legato string lines for "romantic" scenes, brass fanfares for a "majestic" quality, or low frequency synthesizer tones for a sense of "foreboding". The ability of music to convey such a referential "meaning" has been explored in great detail by many investigators, e.g. Heinlein (1928),

Hevner (1935 & 1936), Farnsworth (1954), Meyer (1956), Wedin (1972), Eagle (1973), Crozier (1974), McMullen (1976), Brown (1981), and Asmus (1985).

The second implicit judgment (i.e. mapping of accent structures) consists of matching emphasized points in one perceptual modality with those in another. Lipscomb & Kendall (1996) proposed that, if the associations identified with the musical style were judged appropriate and the relationship of the aural and visual accent structures were perceived as consonant, attentional focus would be maintained on the symbiotic composite, rather than on either modality in isolation.

Figure 1. Lipscomb & Kendall's (1996) model of Film Music Perception. Reprinted with permission of *Psychomusicology*.



Musical and Visual Periodicity. In the repertoire of mainstream motion pictures, one can find many examples that illustrate the film composer's use of periodicity in the musical structure as a means of heightening the effect of recurrent motion in the visual image. The galley rowing scene from Miklos Rosza's score composed for Ben Hur (1959) is an excellent example of the mapping of accent structures, both in pitch and tempo of the musical score. As the slaves pull up on their oars, the pitch of the musical motif ascends. As they lean forward to prepare for the next thrust, the motif descends. Concurrently, as the Centurion orders them to row faster and faster, the tempo of the music picks up accordingly, synchronizing with the accent structure of the visual scene. A second illustration may be found in John Williams' musical soundtrack composed for ET: The Extraterrestrial (1982). The bicycle chase scene score is replete with examples of successful musical emulation of the dramatic action on-screen. Synchronization of the music with the visual scene is achieved by inserting 3/8 patterns at appropriate points so that accents of the metrical structure remain aligned with the pedaling motion.

In the process of perception, the perceptual system seeks out such periodicities in order to facilitate data reduction. Filtering out unnecessary details in order to retain the essential elements is required because of the enormous amount of information arriving at the body's sensory receptors at every instant of time. "Chunking" of specific sensations into prescribed categories allows the individual to successfully store essential information for future retrieval (Bruner, Goodnow, & Austin, 1958).

Therefore, in the context of the decision-making process proposed by Lipscomb & Kendall (1996), the music and visual images do not necessarily have to be in perfect synchronization for the composite to be considered appropriately aligned. As the Gestalt psychologists found, humans seek organization, imposing order upon situations that are open to interpretation according to the principles of good continuation, closure, similarity, proximity, and common fate (von Ehrenfels, 1890; Wertheimer, 1925; Köhler, 1929; and Koffka, 1935). In the scenes described above, the fact that *every* rowing or pedaling motion was not perfectly aligned with the musical score is probably not perceived by the average member of the audience … even if attention were somehow drawn to the musical score. Herbert Zettl (1990) suggests the following simple experiment:

> *To witness the structural power of music, take any video sequence you have at hand and run some arbitrarily selected music with it. You will be amazed how frequently the video and audio seem to match structurally. You simply expect the visual and aural beats to coincide. If they do not, you apply psychological closure and make them fit. Only if the video and audio beats are, or drift, too*

> *far apart, do we concede to a structural mismatch—but then only temporarily. (p. 380)*[1]

The degree to which the two strata must be aligned before perceived synchronicity breaks down has not yet been determined. The present experimental investigation manipulated the relationship of music and image by using discrete levels of synchronization. If successful in confirming a perceived difference between these levels, future research will be necessary to determine the tolerance for misalignment.

Accent Structure Alignment

Two issues had to be addressed before it was possible to consider accent structure synchronization. First, what constitutes an "accent" in both the visual and auditory domains? Second, which specific parameters of any given visual or musical object have the capability of resulting in perceived accent?

The term “accent” will be used to describe points of emphasis (i.e., salient moments) in both the musical sound and visual images. David Huron (1994) defined “accent” as “an increased prominence, noticeability, or salience ascribed to a given sound event.” When generalized to visual images as well, it is possible to describe an A-V composite in terms of accent strata and their relationships one to another.

Determinants of Accent. In the search for determinants of accent, potential variables were established by considering the various aspects of visual objects and musical phrases that constituted perceived boundaries. Fraisse (1982, p. 157) suggested that grouping of constituent elements results "as soon as a difference is introduced into an isochronous sequence...." Similarly, in a discussion of Gestalt principles and their relation to Lerdahl & Jackendoff's (1983) generative theory of tonal music, Deliege stated that “... in perceiving a difference in the field of sounds, one experiences a sensation of accent" (1987, p. 326). Boltz & Jones (1986) propose that “accents can arise from any deviation in pattern context” (p. 428).

Following an extensive review of the literature relating to the perception of accent in both the aural and visual modalities, a limited number of potential variables were utilized in creating a musical stimulus set and a visual stimulus set that—considering each modality in isolation—resulted in a reliably consistent perception of the intended accent points. Accents were hypothesized to occur at moments in which a change occurs in any of these auditory or visual aspects of the stimulus. This change may happen in one of two ways. First, a value that remains consistent for a period of time can be given a new value (e.g. a series of soft tones may be followed suddenly by a loud tone or a blue object may suddenly turn red). Second, change in the direction of a motion vector will cause a per-

ceived accent (e.g. melodic contour may change from ascending to descending or the direction of an objects motion may change from horizontal left to vertical up). The variables selected for use in the following experiments are listed in Table 1, along with proposed values for the direction and magnitude characteristics.

Table 1. Proposed variables to be utilized in the initial exploratory study labeled with direction.

Variables	**Vectors**	
	Direction	**Magnitude of change**
Musical		
Pitch	up/unchanging/down	none/small/large
Loudness	louder/unchanging/softer	none/small/large
Timbre	simple/unchanging/complex	none/small/large
Visual		
Location	left/unchanging/right up/unchanging/down	none/small/large
Shape	simpler/same/more complex	none/small/large
Color		
hue	red-orange-yellow-green-blue-indigo-violet	none/small/large
saturation	purer/unchanging/more impure	"
brightness	brighter/unchanging/darker	"

METHOD

This study was a quasi-experimental investigation, consisting of a post-test only, repeated measures factorial design. The experiment was preceded by a series of exploratory studies that assisted in selecting stimulus materials. The main experiment incorporated two independent methods of data collection, verbal ratings and similarity judgments.

Subject Selection

Every participant was required to have seen at least four mainstream, American movies during each of the past ten years, ensuring at least a moderate level of "enculturation" with this genre of synchronized audio-visual media. Musical training was the single between-subjects grouping variable considered, using the following three levels: untrained (less than two years of formal music training), moderate (2 to 7 years of formal music training), and highly trained (more than 7 years of formal study).

Stimulus Materials

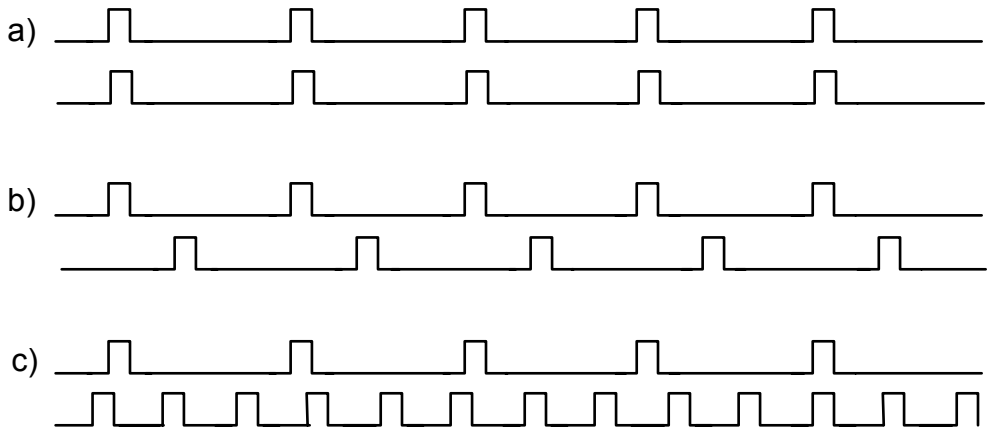
Prior to the main experiment, a series of exploratory studies was run to determine auditory and visual stimuli that are consistently interpreted by subjects as generating an intended accent point. The sources of musical and visual accent delineated in Table 1 were used as a theoretical basis for creating MIDI file and generating computer animations for use as stimuli in this experiment. Both the sound files and the animations were limited to approximately five seconds in length, so that a paired comparisons task could be completed by subjects within a reasonable period of time, as discussed below.

The points of accent were periodically spaced within each musical and visual example. Fraisse (1982, p. 156) identified temporal limits for the perceptual grouping of sound events. The lower limit (approximately 120 ms apart) corresponded closely to the separation at which psychophysiological conditions no longer allowed the two events to be perceived as distinct. The upper limit (between 1500 and 2000 ms) represented the temporal separation at which two groups of stimuli are no longer perceptually linked (Bolton, 1894; MacDougall, 1903). Fraisse suggested a value of 600ms as the optimum for both perceptual organization and precision. Therefore, the first independent variable utilized in the present experimental procedure, i.e. variance of the temporal interval between accent points, consisted of values representing a median range between the limits explicated by Fraisse. This variable had three discrete levels: 500ms, 800ms, and 1000ms. The first and last temporal values allowed the possibility of considering the nesting of accents (e.g. within every 1000ms interval two accents 500ms apart may occur). The 800ms value was chosen because it allowed precise synchronization with the visual stimulus at the rate of 20 frames per second (fps),[2] yet it aligned with the other accent periodicities only once every 4 seconds … beyond Fraisse's (1982) upper limit for the perceptual linking of stimuli). Seven musical patterns and seven animation sequences utilizing each temporal interval were generated, from which the actual stimuli were selected in a second exploratory study.

The manner in which audio and visual stimuli were combined served as the independent variable manipulated by the investigator. Three possible levels of juxtaposition were utilized: consonant, out-of-phase, and dissonant (Yeston, 1976; Monahan, Kendall, & Carterette, 1987; Lipscomb & Kendall, 1996). Figure 2 presents an idealized visual representation of these three relationships. In each pair of accent strata (one depicting the visual component, the other the audio component), points of emphasis are represented by pulses [⎍] in the figure. *Consonant* relationships (Figure 2a) may be exemplified by accent structures that are perfectly synchronized. Accent structures that are *out-of-phase* (Figure 2b) share a common temporal interval between consecutive points of emphasis, but the strata are offset such that they are perceived as out of synchronization. Juxta-

position of the 500ms periodic accent structure and the 800ms periodic accent structure mentioned in the previous paragraph would result in a *dissonant* relationship (Figure 2c).[3] Because of the possibility of nesting the 500ms stimulus within the 1000ms stimulus, it was necessary to distinguish between identical consonance (e.g. synchronization of a 500ms temporal interval in both the audio and visual modalities) and nested consonance (e.g. synchronization of a 500ms temporal interval in one modality and a 1000ms temporal interval in the other). The same distinction was considered in the out-of-phase relationship between the 500ms and the 1000ms periodicities.

Figure 2. Visual representations of relationships between sources of accent.



Exploratory Studies

A series of exploratory studies was run in order to select auditory and visual stimuli that illustrate, as clearly as possible, the presence of accent structures in both perceptual modalities, so that subjects were capable of performing tasks based on the alignment of these two strata. For all experimental procedures, Roger Kendall's *Music Experiment Development System* (MEDS, version 3.1e) was utilized to play the auditory and visual examples and collect subject responses. The author programmed a module for incorporation into MEDS that allowed quantification and storage of temporal intervals between consecutive keypresses on the computer keyboard at a resolution well below .01ms. This facility allowed the subjects to register their perceived pulse simply by tapping along on the spacebar.[4]

Subjects were asked to tap along with the perceived pulse created by the stimulus while either viewing the animation sequences or listening to the tonal sequences. In the exploratory study, stimuli were continuously looped for a period of about 30-seconds so that subjects had an adequate period of time to determine accent periodicities. It was hypothesized that the position of these

perceived pulses coincided with points in time when significant changes in the motion vector (i.e., magnitude or direction) of the stimulus occurred. The purpose of the exploratory studies was to confirm this hypothesis and to determine the audio and visual stimuli that produced the most reliably consistent sense of accent structure.

Main Experiment

There are two methodological innovations incorporated into this study that warrant brief discussion. First, a system of "convergent methods" was utilized to answer the research questions. Kendall & Carterette (1992a) proposed this alternative to the single-method approach used in most music perception and cognition research. The basic technique is to "converge on the answer to experimental questions by applying multiple methods, in essence, simultaneously investigating the central research question as well as ancillary questions of method" (p. 116). In addition, if the answer to a research question is the same, regardless of the method utilized, much greater confidence may be attributed to the outcome. The present investigation incorporated a verbal scaling procedure and a similarity judgment task.

Second, rather than using semantic differential bipolar opposites in the verbal scaling task (Osgood et al., 1957), verbal attribute magnitude estimation (VAME) was utilized (Kendall & Carterette, 1992b & 1993). In contrast to semantic differential scales, VAME provides a means of assigning a specific amount of a given attribute within a verbal scaling framework (e.g., good–not good, instead of good–bad).

Since two convergent methods were utilized, two independent groups of subjects were required for this experiment. Group One was asked to watch every audio-visual composite in a randomly-generated presentation order and provide a VAME response, according to a consistent set of instructions (see Lipscomb, 1995). When the OK button was pressed after a response, location of each button on its respective scroll bar was quantified using a scale from 0 to 100 and stored for later analysis. A repeated measures analysis of variance (ANOVA) was used as the method for determining whether or not there was a significant within-subjects difference between the responses as a function of accent structure alignment and the between-subjects variable, level of musical training.

In a paired-comparison task, Group Two was asked to provide ratings of "similarity" on a continuum from "not same" to "same", according to a consistent set of instructions (see Lipscomb, 1995). The quantified subject responses were submitted for multidimensional scaling (MDS) in which distances were calculated between objects—in this case, AV composites—for placement within a multidimensional space (Kruskal, 1964a, 1964b, & 1978). The resulting points were plotted and analyzed in an attempt to determine sources of commonality and

differentiation. The results were confirmed by submitting the same data set for cluster analysis in order to identify natural groupings in the data.

Alternative Hypotheses

It was hypothesized that Group One would give the highest verbal ratings of synchronization and effectiveness to the consonant alignment condition (i.e., composites in which the periodic pulses identified in the exploratory studies were perfectly aligned). It was also hypothesized that the lowest scores would be given in response to the out-of-phase condition (i.e., combinations made up of identical temporal intervals that are offset), while intermediate ratings would be related to composites exemplifying a *dissonant* relationship. In the latter case, the musical and visual vectors may be perceived as more synchronized because of the process of closure described by Zettl (1990, p. 380).

It was hypothesized that similarity ratings provided by Group Two would result in a multi-dimensional space consisting of at least three dimensions, including musical stimulus, visual stimulus, and accent alignment.

EXPERIMENTAL PROCEDURE

Auditory examples for this experiment consisted of isochronous pitch sequences and visual images were computer-generated animations of a single object (a circle) moving on-screen. Since the stimuli for this experiment were created by the author, a great degree of care was taken in the exploratory study portion to ensure reliability in responses to the selected stimuli.[5] As a result of these carefully controlled preliminary procedures, from the seven audio examples and seven visual examples created, two audio and two visual stimuli were selected for use in the main experiment (Figures 3 & 4).

Figure 3. Audio stimuli selected for use in the Main Experiment. A1 exhibits accent due to interval and direction change, while A2 exhibits accent resulting from dynamic accent and direction change.

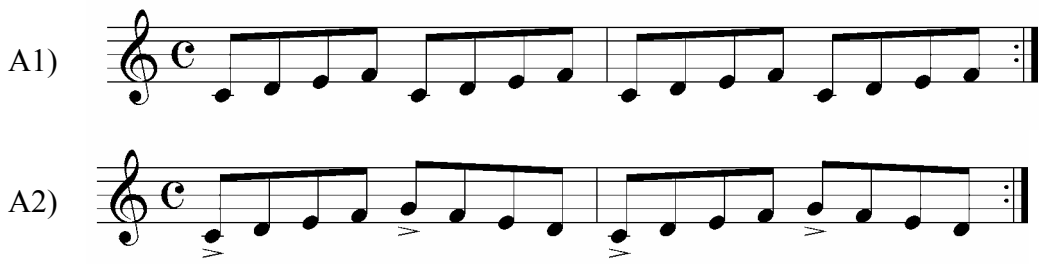
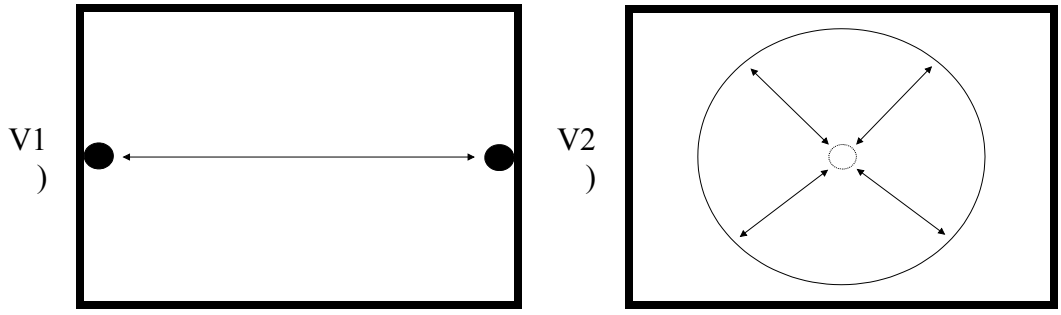


Figure 4. Visual stimuli selected for use in the Main Experiment. V1 exemplifies side-to-side continuous motion, while V2 illustrates apparent near-to-far-to-near continuous motion.



MAIN EXPERIMENT

Subjects

Subjects for this experiment were 40 UCLA students (ages 19 to 31) enrolled in general education classes in the Music Department … either Psychology of Music (Lipscomb; Fall 1994) or American Popular Music (Keeling; Summer I 1994).[6] The 40 subjects were randomly assigned to two groups before performing the experimental tasks. Group One (n = 20) responded using the VAME verbal rating scale and Group Two (n = 20) provided similarity judgments between pairs of stimuli. For each group of subjects, the number of subjects falling into each level of musical training is provided in Table 2.

Table 2. Number of subjects falling into each cell of the between-subjects design (Experiment One).

	Musical Training		
Exp. Task	*Untrained*	*Moderate*	*Trained*
VAME	10	7	3
Similarity	10	8	2

Stimulus Materials

AV composites utilized in the main experiment were created by combining the two audio and the two visual stimuli selected in the exploratory study into all possible pairs ($n_{AV} = 4$). For ease of discussion, these stimuli will heretofore be referenced using the following abbreviations: A1 (Audio 1) consists of a repeated ascending melodic contour), A2 (Audio 2) consists of an undulating melodic contour, V1 (Visual 1) represents left to right apparent motion (i.e.,

translation in the plane along the x-axis), and V2 (Visual 2) represents front to back apparent motion (i.e., *translation in depth* along an apparent z-axis).

In addition to these various AV composites, the method of audio-visual alignment was systematically altered. As explained previously, three alignment conditions were utilized: consonant, out-of-phase, and dissonant. It was important to create composites in which the AV alignment was out-of-phase by an amount that was easily perceivable by the subjects. Friberg & Sundberg (1992) determined the amount by which the duration of a tone presented in a metrical sequence must be varied before it is perceived as different from surrounding tones. This amount is 10ms for tones shorter than 240ms or approximately 5% of the duration for longer tones (p. 107). The out-of-sync versions in this study were offset by 225ms—a value well beyond the just-noticeable difference (JND) for temporal differentiation and also a value that does not nest within or subdivide any of the three IOIs used in this study (500ms, 800ms, and 1000ms).

Stratification of accent structures. In the exploratory study, both the audio and visual stimuli were shown to create a perceived periodic accent where certain moments in the stimulus stream were considered more salient than others. It is possible—using all combinations of synchronization (consonant, out-of-phase, and dissonant) and IOI interval (500ms, 800ms, and 1000ms)—to generate 14 different alignment conditions for each AV composite (Table 5). Notice that there are two distinct types of consonant and out-of-phase composites. The first is an *identical consonance*, e.g., a 1000ms IOI in the audio component perfectly aligned with a 1000ms IOI in the visual component. The second type is referred to as a *nested consonance*, e.g., a 500ms IOI in the audio component that is perfectly aligned with—but subdivides—a 1000ms IOI in the visual component (or vice versa). The corresponding pair of out-of-phase composites is referred to as *out-of-phase (identical)* and *out-of-phase (nested)*. Therefore, the total stimulus set consisted of 56 AV composites (4 AV combinations x 14 alignment conditions). Each composite was repeated for a period of approximately 15 seconds, before requiring a subject response. The order of stimulus presentation was randomized for every subject.

Table 3. The 14 alignment conditions for A-V composites in Experiment One.

Music IOI	Visual IOI	Audio-visual alignment
500ms	500ms	consonant
500ms	500ms	out-of-phase
500ms	1000ms	consonant
500ms	1000ms	out-of-phase
500ms	800ms	dissonant
1000ms	1000ms	consonant
1000ms	1000ms	out-of-phase
1000ms	500ms	consonant
1000ms	500ms	out-of-phase
1000ms	800ms	dissonant
800ms	800ms	consonant
800ms	800ms	out-of-phase
800ms	500ms	dissonant
800ms	1000ms	dissonant

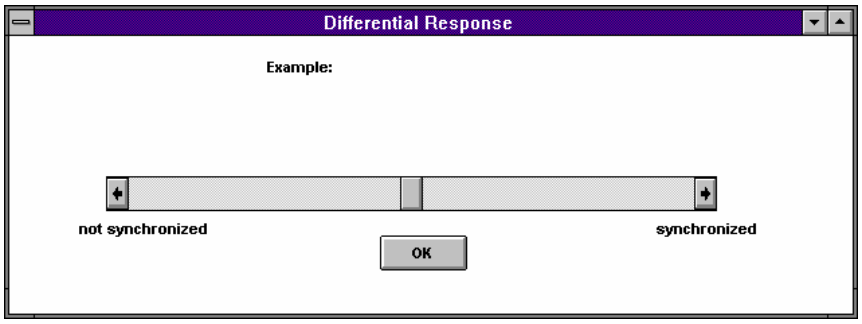
Experimental Tasks

Group One. Each subject in this group was asked to respond to every AV composite on two VAME scales: "not synchronized–synchronized" and "ineffective–effective." After viewing each one of the composites, subjects were given a choice of either providing a response or repeating the stimulus. The response mechanism is shown in Figure 5. The order in which the two VAME scales were presented was also randomized.

Group Two. The similarity scaling task required comparison of all possible pairs of stimuli. Therefore, it was necessary to utilize only a subset of the composites used in the VAME task in order to ensure that the entire procedure could be run within a reasonable time period (i.e., 30 to 45 minutes). Only the 800ms MIDI and animation files were utilized, eliminating nesting and varying temporal interval from consideration. The alignment conditions were simply consonant (800ms IOI MIDI file and 800ms IOI FLI animation, perfectly aligned), out-of-phase (800ms IOI MIDI file and 800ms IOI FLI animation, offset by 225ms), and dissonant (1000ms IOI MIDI file and 800ms IOI FLI animation). The triangular matrix of paired comparisons included the diagonal (identities) as a means of gauging subject performance, i.e., if identical composites are consistently judged to be "different," it is likely that the subject did not understand or was unable to perform the task. Therefore, the total stimulus set consisted of 12 different AV composites (4 AV combinations x 3 alignment conditions), resulting in 78 pairs of stimuli. All paired-comparisons were randomly generated, so that

the subject saw one AV composite and then a second combination prior to providing a similarity judgment.

Figure 5. Scroll bar used to collect Group One subject responses.



Results

Group One Data Analysis and Interpretation. A repeated measures ANOVA was performed on the subject responses to each of the VAME rating scales provided by Group One,[7] considering two within-groups variables (4 AV combinations and 14 alignment conditions) and one between-groups variable (3 levels of musical training). At $\alpha = .025$, neither the synchronization ratings ($F_{(2,17)} = 1.62$, $p < .227$) nor the effectiveness ratings ($F_{(2,17)} = .66$, $p < .528$), exhibited any significant difference across levels of musical training. However, there was a highly significant within-subjects effect of alignment condition for both the synchronization ratings ($F_{\lambda(13,221)} = 88.18$, $p < .0005$) and the effectiveness ratings ($F_{\lambda(13, 221)} = 48.43$, $p < .0005$). The only significant interaction that occurred was an interaction between AV combination and alignment condition for both synchronization ($F_{(39,663)} = 3.05$, $p < .0005$) and effectiveness ($F_{(39,663)} = 2.94$, $p < .0005$). In general, there was a high correlation between subject responses on the synchronization and effectiveness scales ($r = .96$), confirming the strong positive relationship between ratings of synchronization and effectiveness.

Mean subject responses to the VAME scales are represented graphically in Figures 6a to 6d. Each graph represents a different AV combination. There is a striking consistency in response pattern across AV composites, as represented in graphs. This consistency is confirmed by Figure 7, providing a comparison of these same responses across all four AV combinations by superimposing Figures 6a to 6d on top of one another. In the legend to this figure, the labels simply refer to a specific alignment condition of a given AV combination (e.g., V1A2_C refers to the consonant alignment condition of Visual #1 and Audio #2).[8] There is a relatively consistent pattern of responses, based on alignment condition. In general, the consonant combinations receive the highest mean ratings on both verbal

scales. The identical consonant composites (e.g., alignment conditions V5A5_C, V10A10_C, and V8A8_C in Figure 7) are consistently given higher ratings than the nested consonant composites (e.g., alignment conditions V5A10_C and V10A5_C in Figure 7),[9] with the exception of V10A5_C4 which received a mean rating almost equal to that of V10A10_C4.[10]

Figure 6a. Mean subject ratings to the two VAME scales when viewing the combination of Visual #1 and Audio #1 across alignment conditions. For an explanation of x-axis labels, see endnote #8.

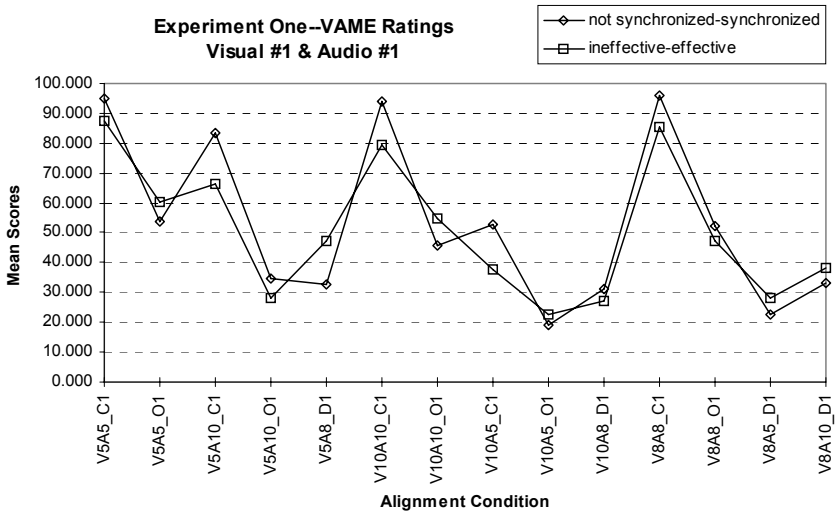


Figure 6b. Mean subject ratings to the two VAME scales when viewing the combination of Visual #1 and Audio #2 across alignment conditions. For an explanation of alignment condition labels, see endnote #8.

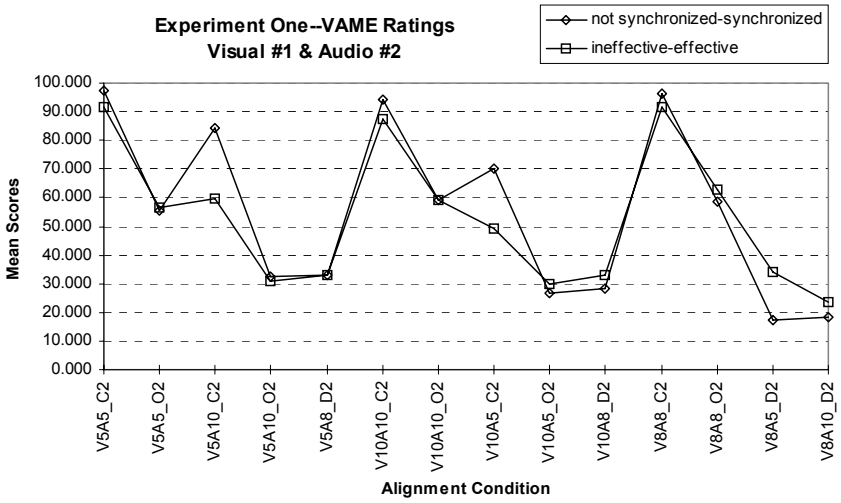


Figure 6c. Mean subject ratings to the two VAME scales when viewing the combination of Visual #2 and Audio #1 across alignment conditions. For an explanation of alignment condition labels, see endnote #8.

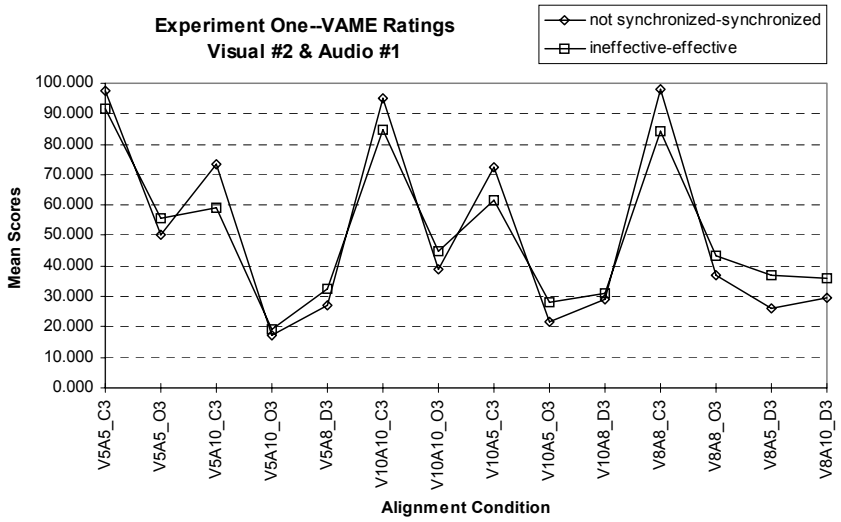


Figure 6d. Mean subject ratings to the two VAME scales when viewing the combination of Visual #2 and Audio #2 across alignment conditions. For an explanation of alignment condition labels, see endnote #8.

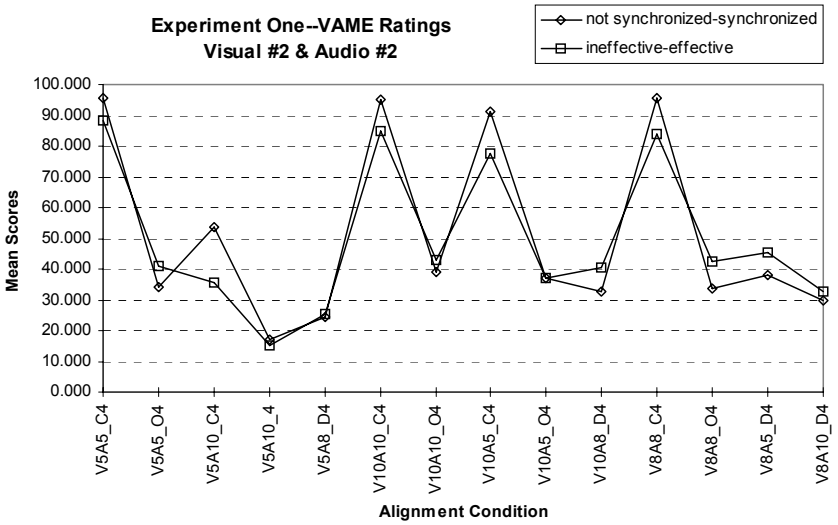
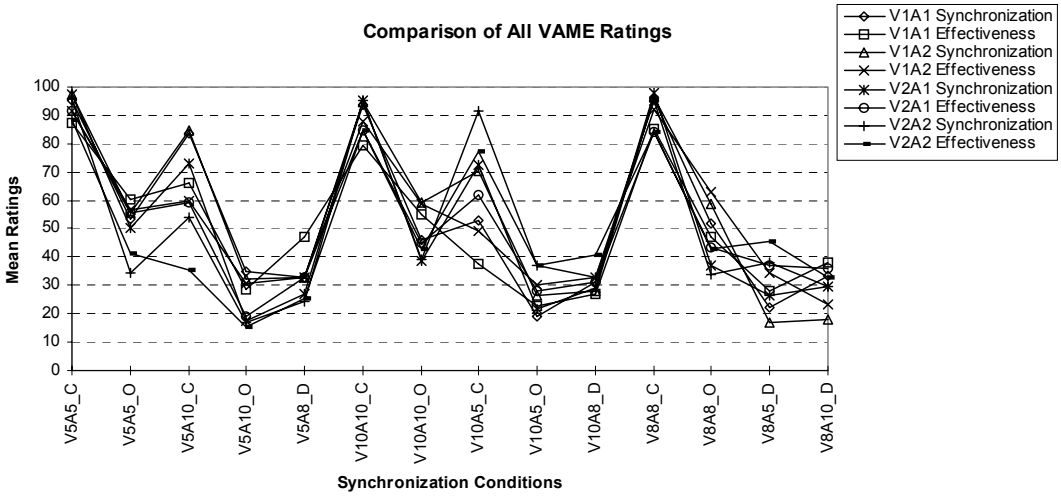


Figure 7. Comparison of all VAME responses across AV composite and alignment conditions. For an explanation of alignment condition labels, see endnote #8.



The second highest mean ratings were given in response to the out-of-phase (identical) composites (e.g., V5A5_O, V10A10_O, and V8A8_O). The lowest mean ratings were always given in response to the out-of-phase (nested) composites (e.g., V5A10_O and V10A5_O) and the dissonant composites (e.g., V5A8_D, V10A8_D, V8A5_D, and V8A10_D) with the former usually being rated slightly

higher than the latter.[11] Therefore, the relationship between subject responses on the two VAME scales and accent structure alignment may be represented as shown in Table 4. This ordering of response means is different from that proposed initially as an alternative hypothesis. Recall that, based upon Zettl's (1990) theory related to closure in the perception of musical and visual vectors, the dissonant conditions were predicted to be perceived as more synchronized and effective than those of the out-of-phase conditions. The responses of Group 1 reveal, however, that higher ratings were given to the out-of-phase conditions than to the dissonant conditions. It is still possible to explain these results in terms of closure, as described by Zettl. However, the process of closure appears to have been applied in a manner different from that predicted at the outset. Subject responses revealed that the out-of-phase conditions were perceived as more synchronized and effective than the dissonant conditions, in contrast to the results predicted. In hindsight, perhaps this result makes more sense than the proposed alternative hypothesis. It appears that, in the process of viewing the present collection of audio-visual composites, subjects sought out recurrent periodicities and considered those that shared the same IOI between accent points to be more synchronized and more effective than those with different IOIs, even if these periodicities were misaligned to a highly perceptible degree. Future research will be required to distinguish between the importance of absolute accent structure alignment and the influence of such matched periodicities.

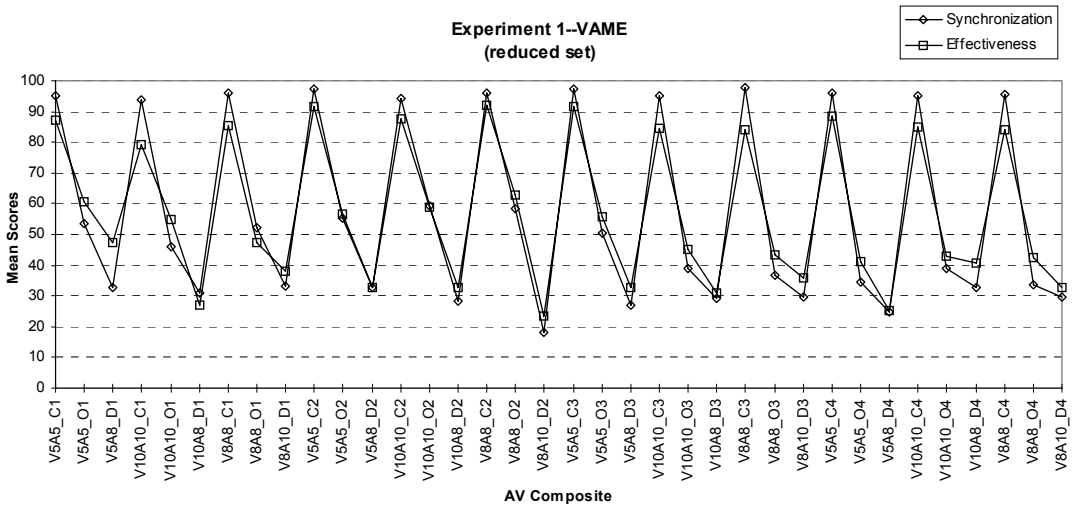
Table 4. AV composites arranged from highest response to lowest on the VAME scales.

Identical Consonant Composites	**Highest**
Nested Consonant Composites	↑
Out-of-Phase (Identical) Composites	|
Out-of-Phase (Nested) & Dissonant Composites	**Lowest**

Collapsing Alignment Conditions Across AV Composites. The subject VAME responses were collapsed across alignment conditions. When compared to a single measurement, such multiple measures of a single condition provide increased reliability (Lord & Novick, 1968). Therefore, the mean of all synchronization ratings given in response to the consonant alignment condition (i.e., V1A1_C, V1A2_C, V2A1_C, and V2A2_C) was calculated and compared to the mean ratings for the out-of-phase and dissonant alignment conditions. The ratings of effectiveness were collapsed as well. An ANOVA on the collapsed data set revealed that the significant interaction between AV composite and alignment condition observed over the complete data set fell to a level not considered statis-

tically significant (synchronization—$F_{\lambda(6,102)} = 1.98959$, $p < .056$; effectiveness—$F_{\lambda(6,102)} = 2.18760$, $p < .117$).[12] Further justification for collapsing the data in this manner can be derived from the VAME data set. Figure 8 represents mean ratings for both VAME scales across all AV combinations at every IOI. Notice the contour similarity across every consonant, out-of-phase, and dissonant combination, i.e., the consonant pairs consistently received the highest rating, the dissonant pairs received the lowest rating, and the out-of-phase pairs received a rating in-between the other two. In addition, the subject responses to the nested conditions exhibited the most influence of specific AV combinations. Therefore, eliminating these conditions further justified collapsing alignment conditions (consonant, out-of-phase, and dissonant) across the various AV combinations. For the remainder of this investigation, only three alignment conditions will be considered: consonant, out-of-phase, and dissonant, eliminating the nesting conditions.

Figure 8. VAME ratings for all consonant, out-of-phase, and dissonant combinations across all AV composites.

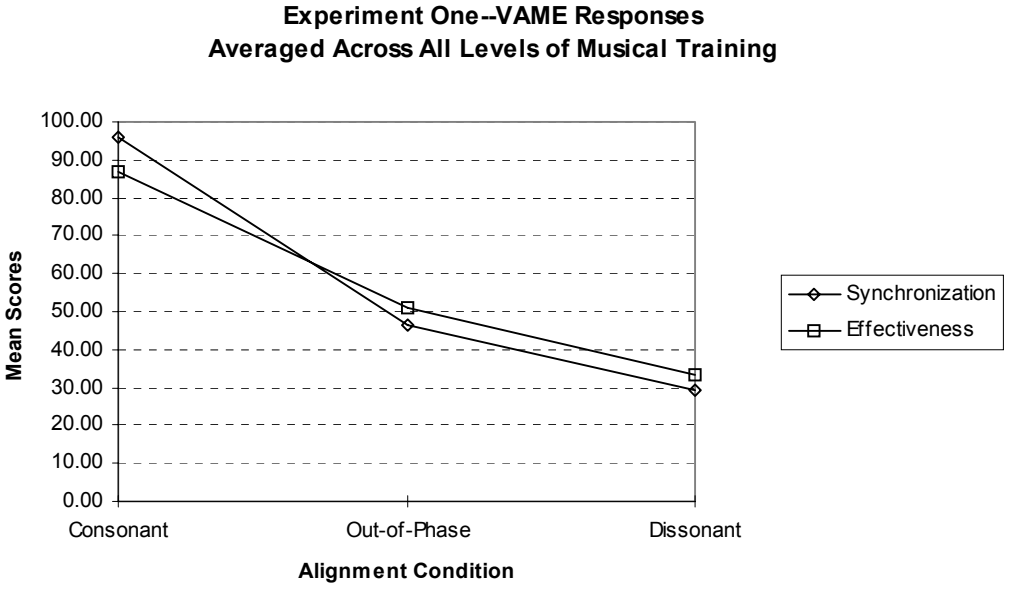


Analysis and Interpretation of Data Collapsed Across Alignment Conditions. An ANOVA across the collapsed data set confirmed that there is no significant difference of the level of musical training for either the synchronization ratings ($F_{(2,17)} = 1.699$, $p < .2125$) or the effectiveness ratings ($F_{(2,17)} = .521$, $p < .603$). Once again, however, analysis reveals a highly significant effect of alignment condition for both the synchronization ratings ($F_{\lambda(2,16)} = 162.274$, $p < .0001$) and the effectiveness ratings ($F_{\lambda(2,16)} = 91.591$, $p < .0001$). The interaction between level of musical training and alignment condition was not significant for

either synchronization ($F_{\lambda(4,32)} = 1.575$, $p < .2048$) or effectiveness ($F_{\lambda(4,32)} = 2.662$, $p < .0504$).

Regardless of musical training, subjects are clearly distinguishing between the three alignment conditions on both VAME scales (Figure 9). Consonant combinations were given the highest ratings with a steep decline between consonant and out-of-phase combinations, followed by an even lower rating for the dissonant pairs. Interestingly, the effectiveness ratings were consistently less extreme than the ratings of synchronization. For example, when the mean synchronization rating was extremely high (e.g., the consonant alignment condition), the effectiveness rating was slightly lower. However, when the synchronization ratings were lower (e.g., the out-of-phase and dissonant alignment conditions), the effectiveness ratings were slightly higher. This suggests that, while synchronization ratings varied more consistently according to alignment condition, ratings of effectiveness may have been tempered slightly by other factors inherent in the AV composite.

Figure 9. Mean VAME ratings for Experiment One averaged across all levels of musical training.



Group Two

Data Analysis. A repeated measures ANOVA was also performed on the similarity ratings provided by Group Two, using one within-groups variable (78 paired comparisons) and one between-groups variable (3 levels of musical train-

ing). There was no significant effect of either musical training ($F_{(2,17)} = .40$, $p < .676$) or the interaction between musical training and similarity ratings ($F_{(154, 1309)} = .56$, $p < 1.000$). As one would expect, however, the similarity ratings did vary at a high level of significance ($F_{(77,1309)} = 24.86$, $p < .0005$). Therefore, the null hypothesis is rejected, because subject ratings of similarity between AV composites did, in fact, vary significantly as a function of AV alignment.

Multidimensional Scaling. The triangular mean similarity matrix was submitted for multidimensional scaling (MDS) analysis. Figure 10 provides the MDS solution in three dimensions, accounting for 99.884% of the variance at a stress level of only .01189. The twelve stimuli separated clearly on each dimension. All composites using Audio #1 are on the negative side of the "Audio" dimension (x-axis) and all composites incorporating Audio #2 are on the positive side. Likewise, all composites utilizing Visual #1 are on the negative side of the "Visual" dimension (z-axis) and all composites using Visual #2 are on the positive side. Finally, all of the composites that are considered dissonant, fall within the negative area of the "Sync" dimension (y-axis) and all consonant and out-of-phase composites fall on the positive side, practically on top of one another.

Notice how tightly the stimuli clustered within this 3-dimensional space when viewed from above (i.e., across the Visual and Audio dimensions).[13] To further examine the group membership among the various AV composites, the same triangular matrix was submitted for cluster analysis.

Cluster Analysis. Cluster analysis provides a method for dividing a data set into subgroups without any *a priori* knowledge considering the number of subgroups nor their specific members. The tree diagram presented in Figure 11 graphically illustrates the clustering of AV composites used in the present study.

Figure 10. Multidimensional scaling solution for the similarity judgments in Experiment One.

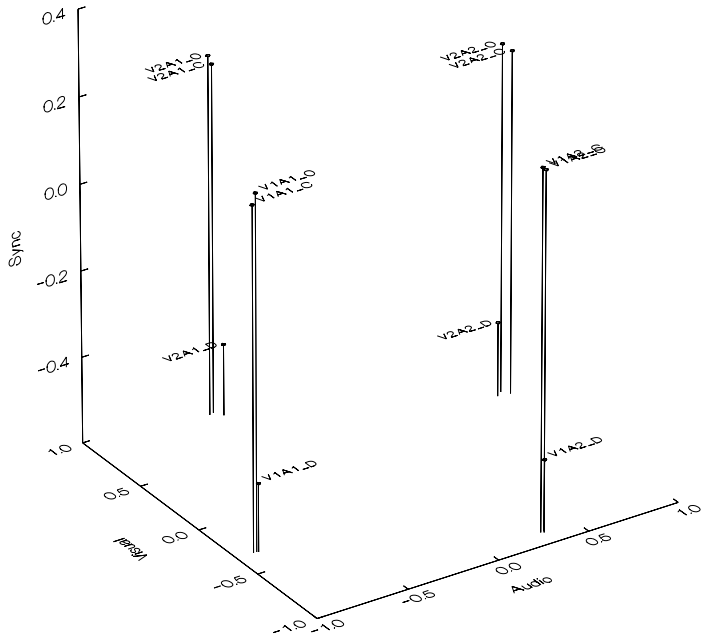
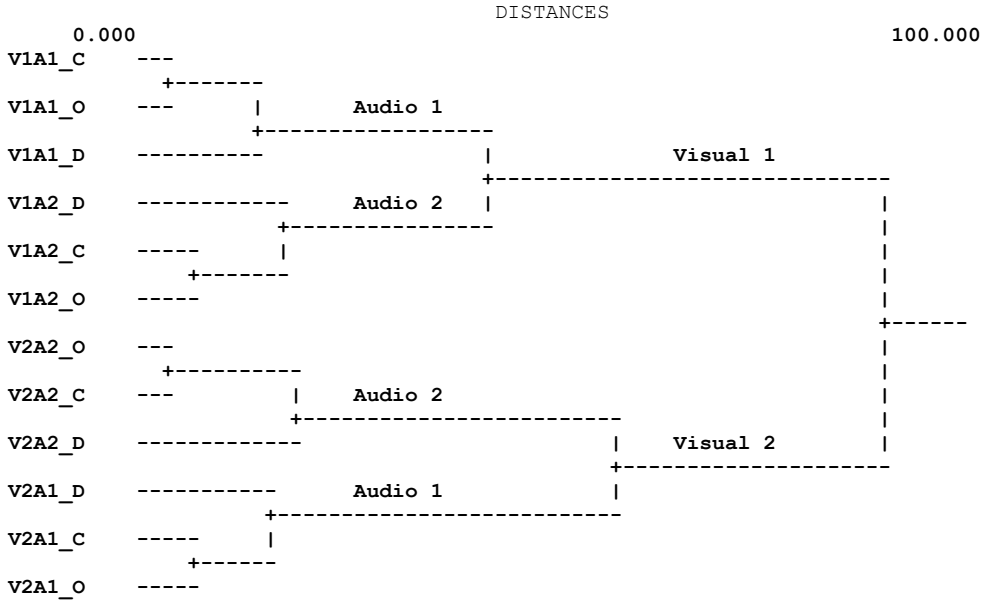


Figure 11. Cluster Analysis tree diagram—complete linkage (farthest neighbor)—for similarity ratings provided by subjects in Experiment One, Group Two.



As is readily apparent when considering this cluster diagram from right to left, the initial branching of AV composites into subgroups clearly separates the composites according to the visual component, i.e., all composites on the upper branch utilize Visual One (V1) and all composites on the lower branch utilize Visual Two (V2). The next subdivision separates the composites according to audio component, as labeled in the diagram. The third subdivision separates the composites with the same IOIs (i.e., consonant and out-of-phase composites) from those composites in which the audio and visual components are of differing IOIs (dissonant composites). Finally, the fourth subdivision divides the consonant composites from the out-of-phase composites. Notice also the mirroring relationship within the lower cluster of six composites (those using V2), based upon alignment condition (see Figure 12a). The closest cross-cluster relationship between those composites incorporating A1 and those using A2 is the dissonant condition, neighbored by the consonant condition, and working outward finally to the out-of-phase condition. A similar mirroring is apparent when comparing the composites that incorporate A2, whether combined with V1 or V2 (Figure 12b). In this case, the alignment condition of the closest pair (V1A2_O and V2A2_O) is out-of-phase, working outward to consonant, then dissonant. It is worthy of notice that, when considering the main (i.e., visual) branches of the cluster solution in Figure 11, the two neighbor composites (V1A2_O and V2A2_O) share the same audio track *and alignment condition*. These relationships within the cluster branching structure further confirmed the role of alignment condition in the subject ratings of similarity.

Figure 12a. Illustration of the mirroring relationship between elements in the upper cluster of composites incorporating Visual One.

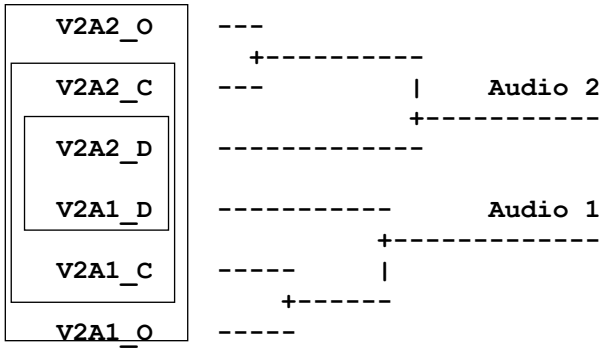
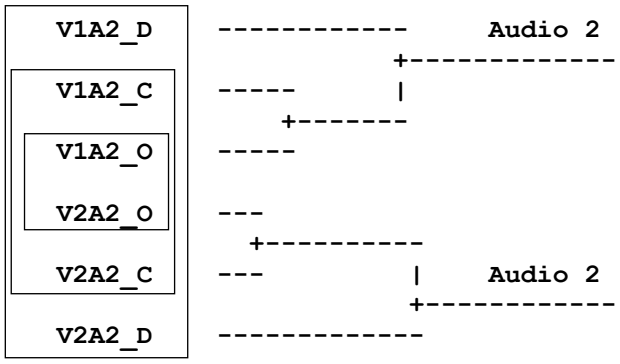


Figure 12b. Illustration of the mirroring relationship between elements in the middle cluster of composites incorporating Audio Two with either Visual One or Visual Two.



Conclusions

Summarizing the results of the main experiment, both of the converging methods (i.e., VAME ratings and similarity judgments) substantiated the fact that alignment condition between audio and visual components of an AV composite were a determining factor in the subject responses. In the VAME scores, verbal ratings of "synchronization" and "effectiveness" did, in fact, vary as a function of AV alignment. In general, the highest ratings on both scales were given in response to the identical consonant composites followed (in order of magnitude) by nested consonant composites, and then out-of-phase identical composites. The lowest ratings were consistently given to either the out-of-phase (nested) or dissonant composites. Collapsing VAME ratings across alignment condition confirmed the relationship between consonant, out-of-phase, and dissonant pairs, revealing that the ratings of effectiveness were consistently less extreme than the synchronization ratings.

In the similarity judgments, an analysis of variance confirmed that there was a significant difference between ratings given to composites exemplifying the various alignment conditions. MDS revealed three easily interpretable dimensions. Cluster analysis confirmed the three criteria utilized by subjects in the process of determining similarity. In decreasing order of significance, these were the visual component, the audio component, and alignment condition. The fact that the alignment condition plays a significant role in both the multidimensional scaling solution and in the cluster analysis confirms the importance of including "Accent Structure Relationship" as one of the Implicit Judgments in the model of Film Music Perception (Figure 1).

DISCUSSION

Research Questions Answered

The first question posed was "What are the determinants of 'accent?'" A review of related literature revealed numerous potential sources of accent in both the aural and visual domains. Several researchers and theorists proposed that introducing a change into a stimulus stream results in added salience (i.e., accent). The exploratory study confirmed that hypothesized accent points using parameters (both aural and visual) gleaned from this literature review were, in fact, perceived by subjects and reproduced in a tapping procedure. Particularly reliable in producing an event perceived as musically salient were pitch contour direction change, change in interval size, dynamic accent, and timbre change. Likewise, particularly reliable in producing events perceived as salient in the visual domain were translations in the plane (top-to-bottom, side to side, left-bottom-to-right-top, etc.) and translation in depth (back-to-front).

The second research question—and the main source of interest in the present study—concerned whether accent structure alignment between auditory and visual components was a necessary condition for the combination to be considered effective when viewing an AV composite. This question is answered very clearly by the VAME responses of Group One. Calculation of the Pearson correlation coefficient revealed that subject ratings of synchronization and effectiveness shared a strong positive relationship ($r = .96$). Therefore, AV combinations that were rated high in synchronization also tended to be rated high on effectiveness and vice versa. In addition, results of multidimensional scaling and cluster analysis revealed clear influence of the synchronization condition. We may conclude that, when using simple audio and visual stimuli, accent structure alignment does appear to be a necessary condition in order for an AV combination to be considered effective.

As a result, in addition to the overwhelming evidence supporting the important referential aspect of the role played by film music, it is imperative that attention be given to the manner in which the audio and visual components are placed in relation to one another. More specifically, the present study has shown that the manner in which salient moments in the auditory and visual domains are aligned results in a significantly different perceptual response to the resulting composite. Therefore, in light of the findings of the present investigation, it is important to reconsider the results of investigations that simply juxtaposed sound upon image with little attention to the manner in which they were combined (e.g., Tannenbaum, 1956; Thayer & Levinson, 1984; and Marshall & Cohen, 1988). Marshall & Cohen appear to have considered this possibility in their proposed "Congruence-Associationist" model, in which semantic association and temporal

congruence form the basis for judgments concerning the appropriateness of an audio-visual combination. In a following published discussion of these results, Bolivar, Cohen, & Fentress (1996) state that "the greater [the] temporal congruence the greater the focus of visual attention to which the meaning of the music consequently can be ascribed" (p. 32). Further research will be required to determine the relationship between accent structure alignment (temporal congruence) and referential meaning (association).

Suggestions for Further Research

The most important issue to be addressed in a series of future investigations is whether accent structure alignment remains a necessary condition when viewing more complex stimuli. The present author is currently in the process of preparing a paper that reports findings of an experiment utilizing moderately complex stimuli (experimental animations by Norman McLaren) and highly complex stimuli (actual movie excerpts from Brian DePalma's *Obsession*, 1977).

Future research is also needed to determine the relative importance of referential (i.e., associational) and accent structure (i.e., syntactic) aspects within the motion picture or animation experience. These results would help further revise the model of Film Music Perception. Accuracy of the model could also be enhanced by experimental designs incorporating more complex AV interrelationships. For example, instead of simply having consonant, out-of-phase, and dissonant alignment conditions, it would be possible to create a whole series of consonant alignment periodicities using a basic subset of temporal patterns. Monahan & Carterette (1985) performed a study of this kind using pitch patterns with the four rhythmic patterns: iambic, dactylic, trochaic, and anapest. These four rhythmic patterns could provide the basis for creating a series of animations and a series of pitch patterns. The two could then be combined in all possible pairs for use in a similarity scaling procedure to determine what aspects of the AV composite are particularly salient to an observer. An investigator could incorporate this same stimulus set into a tapping procedure to determine whether subjects tap with the audio, the video, some underlying common pulse, or a complex combinatory rhythmic pattern.

A significant limitation of the present study is that only a small number of audio and visual stimuli were used in order to ensure that subjects could complete the experimental tasks within a reasonable amount of time. In future investigations, the use of blocked designs would allow incorporation of a larger number of stimuli and, hence, improve the investigator's ability to generalize results.

Currently, the temporal duration by which visual images and sounds must be offset in order to be perceived as misaligned (i.e., j.n.d. or just-noticeable difference, in psychophysical terminology) remains undefined. In the present study a liberal amount was selected (225ms) in order to ensure that the offset

amounts were well beyond any psychophysiological or perceptual limitations. Friberg & Sundberg (1992) determined that when introducing a temporal duration change into a series of isochronous tones, the variation could be perceived at temporal intervals as small as 10ms. The amount of temporal offset in a cross-modal perception task would likely be significantly longer, but that determination must be made through rigorous scientific investigation. Such an experimental design should incorporate stimuli of varying levels of complexity, in order to determine whether the j.n.d. is a constant or relative value.

Much research is needed to assist in the quantification of various parameters of the audio-visual experience. Reliable metrics are needed to express accent prominence as well as complexity of a musical passage, a visual image, or an AV combination in quantitative terms. Creating a method to quantify the degree of referentiality in a musical or visual excerpt would be helpful in further developing the model of Film Music Perception.

Finally, the present investigation selected one between-groups variable of interest, i.e., musical training. It would also be equally relevant to run a series of similar studies, using visual literacy[14] as a potential grouping variable. In fact, incorporating both musical training and visual literacy would allow consideration of the musical training by visual literacy interaction, which might prove very interesting.

Conclusion

Scientific investigations into the relationship of visual images and musical sound in the context of motion pictures and animation are a relatively new area of study. The artforms themselves have only existed for a century. However, given the sociological significance of the cinematic experience, it is quite surprising that there is still only a small amount of research literature available addressing issues involved in the cognitive processing of ecologically valid audio-visual stimuli. The present series of experiments, along with those proposed above, will provide a framework upon which to build a better understanding of this important, but underrepresented, area of research.

REFERENCES

Asmus, E. (1985). The effect of tune manipulation on affective responses to a musical stimulus. In G.C. Turk (Ed.) Proceedings of the Research Symposium on the Psychology and Acoustics of Music, pp. 97-110. Lawrence: University of Kansas.

Bermant, R.I. & Welch, R.B. (1976). Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition. Perceptual and Motor Skills, 43, 487-493.

Bolivar, V.J., Cohen, A.J., & Fentress, J.C. (1996). Semantic and formal congruency in music and motion pictures: Effects on the interpretation of visual action. *Psychomusicology*, 13, 28-59.

Bolton, T.L. (1894). Rhythm. *American Journal of Psychology*, 6, 145-238.

Boltz, M. (2001). Musical soundtracks as a schematic influence on the cognitive processing of filmed events. *Music Perception*, 18(4), 427-454.

Boltz, M. (l992). Temporal accent structure and the remembering of filmed narratives. *Journal of Experimental Psychology: Human Perception and Performance, 18, 90-105.*

Boltz, M., Schulkind, M. & Kantra, S. (l991). Effects of background music on the remembering of filmed events. *Memory & Cognition, 19, 593-606.*

Boltz, M. & Jones, M.R. (1986). Does rule recursion make melodies easier to reproduce? If not, what does? *Cognitive Psychology*, 18, 389-431.

Brown, R.W. (1981). Music and language. In *Documentary report of the Ann Arbor Symposium*. Reston, VA: pp. 233-265.

Bruner, J., Goodnow, J.J., & Austin, G.A. (1986). A study of thinking 2nd ed. New Brunswick: Transaction Publishers.

Brusilovsky, L.S. (1972). A two year experience with the use of music in the rehabilitative therapy of mental patients. *Soviet Neurology and Psychiatry*, 5(3-4), 100.

Bullerjahn, C. & Güldenring, M. (1996). An empirical investigation of effects of film music using qualitative content analysis. *Psychomusicology*, 13, 99-118.

Cardinell, R.L. & Burris-Meyer, H. (1949). Music in industry today. *Journal of the Acoustical Society of America*, 19, 547-548.

Crozier (1974). Verbal and exploratory responses to sound sequences varying in uncertainty level. In D.E. Berlyne (Ed.) Studies in the new experimental psychology: Steps toward an object psychology of aesthetic appreciation, pp. 27-90. New York: Halsted Press.

Deliege, I. (1987). Grouping conditions in listening to music: An approach to Lerdahl & Jackendoff's Grouping Preference Rules. *Music Perception*, 4(4), 325-360.

Eagle, C.T. (1973). Effects of existing mood and order of presentation of vocal and instrumental music on rated mood response to that music. *Council for Research in Music Education*, no. 32, 55-59.

Farnsworth, P.R. (1954). A study of the Hevner adjective list. *Journal of the Aesthetics of Artistic Criticism*, 13, 97-103.

Fraisse, P. (1982). Rhythm and tempo. In D. Deutsch (Ed.), The psychology of music (pp. 149-180). New York: Academic Press.

Friberg, A. & Sundberg, J. (1992). Perception of just-noticeable displacement of a tone presented in a metrical sequence of different tones. *Speech Transmission Laboratory—Quarterly Progress & Status Report*, 4, 97-108.

Halpin, D.D. (1943-4). Industrial music and morale. *Journal of the Acoustical Society of America*, 15, 116-123.

Heinlein, C.P. (1928). The affective characters of major and minor modes in music. *Journal of Comparative Psychology*, 8, 101-142.

Hevner, K. (1936). Experimental studies of the elements of expression in music. American Journal of Psychology, 48, 246-269.

Hevner, K. (1935). Expression in music: A discussion of experimental studies and theories. Psychological Review, 42(2), 186-204.

Hough, E. (1943). Music as a safety factor. *Journal of the Acoustical Society of America*, 15, 124.

Huron, D. (1994, June). What is melodic accent? A computer-based study of the *Liber Usualis*. Paper presented at the Canadian University Music Society Theory Colloquium (Calgary, Alberta).

Iwamiya, S. (1996). Interactions between auditory and visual processing when listening to music in an audio visual context: 1. Matching 2. Audio Quality. *Psychomusicology*, 13, 133-153.

Kendall, R.A. & Carterette, E.C. (1993). Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck adjectives. *Music Perception*, 10(4), 445-467.

Kendall, R.A. & Carterette, E.C. (1992a). Convergent methods in psychomusical research based on integrated, interactive computer control. *Behavior Research Methods*, 24(2), 116-131.

Kendall, R.A. & Carterette, E.C. (1992b, February). Semantic space of wind instrument dyads as a basis for orchestration. Paper presented at the Second International Conference on Music Perception and Cognition, Los Angeles, CA.

Kerr, W.A. (1945). Effects of music on factory production. *Applied Psychology Monographs*, no. 5. California: Stanford University.

Koffka, K. (1935). Principles of Gestalt psychology. New York: Harcourt, Brace.

Köhler, W. (1929). Gestalt Psychology. New York: Liveright.

Krumhansl, C.L. & Schenck, D.L. (1997). Can dance reflect the structural and expressive qualities of music? A perceptual experiment on Balanchine's choreography of Mozart's *Divertimento No. 15*. *Musicae Scientiae*, 1(1), 63-85.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika*, 29, 1-27.

Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115-129.

Kruskal, J.B. (1978). Multidimensional Scaling. Beverly Hills, CA: Sage Publications.

Lerdahl, F. & Jackendoff, R. (1983). A generative theory of Tonal Music. Cambridge, MA: MIT Press.

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

Lipscomb, S.D. (1990). Perceptual judgment of the symbiosis between musical and visual components in film. Unpublished master's thesis, University of California, Los Angeles.

Lipscomb, S. D. (1989, March). Film music: A sociological investigation of influences on audience awareness. Paper presented at the Meeting of the Society of Ethnomusicology, Southern California Chapter, Los Angeles.

Lipscomb, S.D. & Kendall, R.A. (1996). Perceptual judgment of the relationship between musical and visual components in film. *Psychomusicology*, 13(1), 60-98.

Lord, F.M. & Novick, M.R. (1968). Statistical theories of mental test scores. Menlo Park, CA: Addison-Wesley Publishing Co.

MacDougall, R. (1903). The structure of simple rhythm forms. *Psychological Review, Monograph Supplements*, 4, 309-416.

Madsen, C.K. & Madsen, C.H. (1970). Experimental research in music. New Jersey: Prentice Hall.

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

Massaro, D.W. & Warner, D.S. (1977). Dividing attention between auditory and visual perception. Perception & Psychophysics, 21, 569-574.

McGehee, W. & Gardner, J.E. (1949). Music in a complex industrial job. *Personnel Psychology*, 2, 405-417.

McMullen (1976). Influences of distributional redundancy in rhythmic sequences on judged complexity ratings. *Council for Research on Music Education*, 46, 23-30.

Mershon, D.H., Desaulniers, D.H., Amerson, T.C. (Jr.), & Kiever, S.A. (1980). Visual capture in auditory distance perception: Proximity image effect reconsidered. Journal of Auditory Research, 20, 129-136.

Meyer, L.B. (1956). Emotion and meaning in music. Chicago, IL: University of Chicago Press.

Monahan, C.B. & Carterette, E.C. (1985). Pitch and duration as determinant of musical space. *Music Perception*, 3(1), 1-32.

Monahan, C.B., Kendall, R.A., & Carterette, E.C. (1987). The effect of melodic and temporal contour on recognition memory for pitch change. *Perception & Psychophysics*, 41(6), 576-600.

Morris, Phillip, Companies Inc. (1988). Americans and the arts: V. New York: American Council for the Arts.

MTV (1997, May). The Pink Floyd/Wizard of Oz connection. Retrieved May 13, 2002, from: http://www.mtv.com/news/articles/1433194/19970530/story.jhtml.

Nordoff, P. & Robbins, C. (1973). Therapy in music for handicapped children. London: Gallancz.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). The measurement of meaning. Urbana: University of Illinois Press.

Pink Floyd (1973). *Dark Side of the Moon*. Capitol CDP 7 46001 2.

Radeau, M. & Bertelson, P. (1974). The after-effects of ventriloquism. Quarterly Journal of Experimental Psychology, 26, 63-71.

Regan, D. & Spekreijse, H. (1977). Auditory-visual interactions and the correspondence between perceived auditory space and perceived visual space. Perception, 6, 133-138.

Rosar, W.H. (1996). Film music and Heinz Werner's theory of physiognomic perception. *Psychomusicology*, 13, 154-165.

Ruff, R.M. & Perret, E. (1976). Auditory spatial pattern perception aided by visual choices. Psychological Research, 38, 369-377.

Sirius, G. & Clarke, E.F. (1996). The perception of audiovisual relationships: A preliminary study. *Psychomusicology*, 13, 119-132.

Staal, H.E. & Donderi, D.C. (1983). The effect of sound on visual apparent movement. American Journal of Psychology, 96, 95-105.

Tannenbaum, P. H. (1956). Music background in the judgment of stage and television drama. Audio-Visual Communications Review, 4, 92-101.

Thayer, J.F. & Levenson, R.W. (1984). Effects of music on psychophysiological responses to a stressful film. Psychomusicology, 3, 44-54.

Thompson, W.F., Russo, F.A. & Sinclair, D. (1996). Effects of underscoring on the perception of closure in filmed events. *Psychomusicology*, 13, 9-27.

Uhrbock, R.S. (1961). Music on the job: Its influence on worker morale and production. *Personnel Psychology*, 14, 9-38.

Vitouch, O. (2001). When your ear sets the stage: Musical context effects in film perception. *Psychology of Music*, 29, 70-83.

von Ehrenfels, C. (1890). Über Gestaltqualitäten Vierteljahrschrift für wissenschaftliche Philosophie, 14, 249-292.

Wedin, L. (1972). A multidimensional study of perceptual-emotional qualities in music. *Scandinavian Journal of Psychology*, 13, 1-17.

Wertheimer, M. (1925). Über Gestalttheorie. Erlangen: Weltkreis-Verlag.

Yeston, M. (1976). The stratification of musical rhythm. New Haven, CT.: Yale University Press.

Zettl, H. (1990). Sight, sound, motion: Applied media aesthetics, 2nd ed. Belmont, CA: Wadsworth Publishing Co.

Endnotes

[1] For another interesting experience of this type, view *The Wizard of Oz* while listening to Pink Floyd's *Dark Side of the Moon* (1973) as the musical soundtrack. If you start the music on cue with the third roar of the MGM lion, you will be surprised how well the audio and visual components appear synchronized at certain transitional points in the film. Though songwriter Roger Waters has remained silent on the matter, both drummer Nick Mason and engineer Alan Parsons deny any intended relationship (MTV, 1997).

[2] The specific relationships are: 1000ms = 20 frames; 800ms = 16 frames, and 500ms = 10 frames.

[3] Even when using extreme differences between temporal intervals of periodicity, it is inevitable that, at some point in time, the two strata will align for a simultaneous point of accent. This possibility occurs, as mentioned, every 4 seconds when using the 800ms temporal interval with the 500ms or every 8 seconds when combined with the 1000 ms intervals. The fifth pulse in the upper stratum of Figure 5c illustrates such a coincidental alignment.

[4] More information about MEDS is available from Dr. Roger A. Kendall directly at: UCLA Dept. of Ethno- & Systematic Musicology, Schoenberg Hall, 405 Hilgard Ave., Los Angeles, CA 90024 (kendall@ucla.edu). In addition to the KeyPress module, the author also incorporated commands into MEDS, allowing selection of any of the digital or analog audio tracks available on the laserdisc recording. These capabilities were necessary for later experiments.

[5] For a detailed discussion of the exploratory studies, see Lipscomb, 1995, pp. 53-65.

[6] This subject pool consisted of 24 males and 16 females.

[7] The data set for the main experiment (both synchronization and effectiveness ratings) failed the likelihood-ratio test for compound symmetry, violating one assumption of the ANOVA model. Therefore, when appropriate, transformed F- and *p*-values were provided using Wilks' *lambda* (F_λ), which did not assume compound symmetry.

[8] In Figures 6a to 6d, the x-axis labels consist of acronyms formed to identify the specific combination represented. These acronyms consist of the visual stimulus IOI (5, 8 or 10), the aural stimulus IOI (5, 8 or 10), and the alignment condition (C for consonant, O for out-of-phase, or D for dissonant). In each case, the actual IOI value in milliseconds is divided by 100 to make the labels shorter. For example, the label V5A8_D identifies a composite consisting of a visual stimulus with an IOI of 500 ms and an aural stimulus with an IOI of 800 ms resulting in a dissonant alignment condition. Each graph represents a different combination of visual and aural stimuli.

[9] Remember that there is no nested consonant composite for the 800ms stimuli.

[10] Recall that this is the Visual pattern that, because of the results of the exploratory study, was changed from the originally hypothesized accent periodicity to that perceived by all subjects in the tapping task. Perhaps some of the subjects in Experiment One perceived composite V10A5_C4 as nested and others (sensing an accent point at both the nearest and farthest location of visual apparent motion) considered it an identical consonance.

[11] Notice also that the widest spread of mean responses to any of the AV composites is associated with the nested consonant composites (i.e., V5A10_C and V10A5_C).

[12] Running a second ANOVA on this same data set caused the probability of alpha error (α) to increase. The data from each of three experiments were analyzed independently for significant differences and then one final ANOVA was run across the entire data set, including subject responses from all three experiments. Included in these three experiments are the main experiment reported in this paper and two additional experiments that are presently being prepared for publication. Since the alpha error level was set *a priori* to .025, the resulting level of confidence remained above 95% (i.e., .975 x .975).

The single exception to this rule was the analysis of the data from the main experiment reported herein. One ANOVA was already run on the complete data set. Along with the following ANOVA on the collapsed data set and the final ANOVA across all three experiments, the resulting level of confidence was reduced to about 93% (i.e., .975 x .975 x .975).

[13] They cluster so tightly in fact that, when the similarity matrix was forced into two dimensions, it became immediately apparent that the MDS solution was degenerate. Therefore, results of the MDS solution will be supported by consideration of cluster analyses. For a complete discussion, see Lipscomb (1995).

[14] Visual literacy refers to an individual's capability to process visual sensory input. For instance, individuals trained as artists, animators, or film directors tend to be more aware of elements in their visual environment.