

The Perception of Audio-visual Composites: Accent Structure Alignment of Simple Stimuli

SCOTT D. LIPSCOMB
Northwestern University

This investigation examines the relationship between musical sound and visual images when they are paired in simple animated sequences. Based on a model of film music proposed in 1996 by Lipscomb and Kendall, the study focuses specifically on the relationship of points perceived as accented musically and visually. The following research questions were answered: (1) What are the determinants of "accent" (salient moments) in the visual and auditory fields? and (2) Is the precise alignment of auditory and visual strata necessary to ensure that an observer finds the combination effective? In this experimental study, two convergent methods were used: a verbal scaling task and a similarity judgment task. Three alignment conditions were incorporated: consonant (accents in the music occur at the same temporal rate and are perfectly aligned with accents in the visual image), out-of-phase (accents occur at the same rate, but are perceptibly misaligned), or dissonant (accents occur at different rates). Results confirmed that VAME ratings are significantly different for the three alignment conditions. Consonant combinations were rated highest, followed by out-of-phase combinations, and dissonant combinations received the lowest ratings. Subject similarity judgments in response to these simple stimuli divided clearly into three dimensions: visual component, audio component, and alignment condition, further confirming the significance of the alignment of accent strata.

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, for example, 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, 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 1943-1944; Kerr 1945), though these studies are often far from rigorous in their method and analysis (McGehee and Gardner 1949; Cardinell and Burriss-Meyer 1949; Uhrbock 1961). Music therapists claim that music has a beneficial effect in the treatment of some handicapped individuals and as a part of physical rehabilitation following traumatic bodily injury (Brusilovsky 1972; Nordoff and Robbins 1973; an opposing viewpoint is presented by Madsen and Madsen 1970). Individuals use music to facilitate either relaxation or stimulation 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 (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" (points of emphasis) in the visual and auditory fields? 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 and Kendall 1996). In the field of perceptual psychology, interaction between the aural and visual sensory modalities is well documented (see, for example, Radeau and Bertelson 1974; Staal and Donderi 1983; Bermant and Welch 1976; Ruff and Perret 1976; Massaro and Warner 1977; Regan and Spekrijse 1977; and Mershon, Desaulniers, Amerson, and Kiever 1980). For a detailed discussion of film music research (Tannenbaum 1956; Thayer and Levenson 1984; and Marshall and Cohen 1988), see Lipscomb (1995) and Lipscomb and Kendall (1996). The latter paper was included in a special issue of *Psychomusicology* (vol. 13) devoted to the topic of film-music research, including investigations by a wide array of scholars (Thompson, Russo, and Sinclair 1996; Bolivar, Cohen, and Fentress 1996; Lipscomb and Kendall 1996; Bullerjahn and Gulderring 1996; Sirius and 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 (Boltz 1992; Boltz 2001; and Boltz, Schulkind, and Kantra 1991). Krumhansl and Schenck (1997) investigated the relationship between dance choreography by Balanchine and the music that inspired it, Mozart's *Divertimento No. 15*. In a study by Vitouch (2001), subjects, after seeing a brief film excerpt with one of two contrasting musical soundtracks, 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 and 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 A-V presentation, separate judgments were made on each of three semantic dimensions (Evaluative, Potency, and Activity; see Osgood, Suci, and 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 and 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

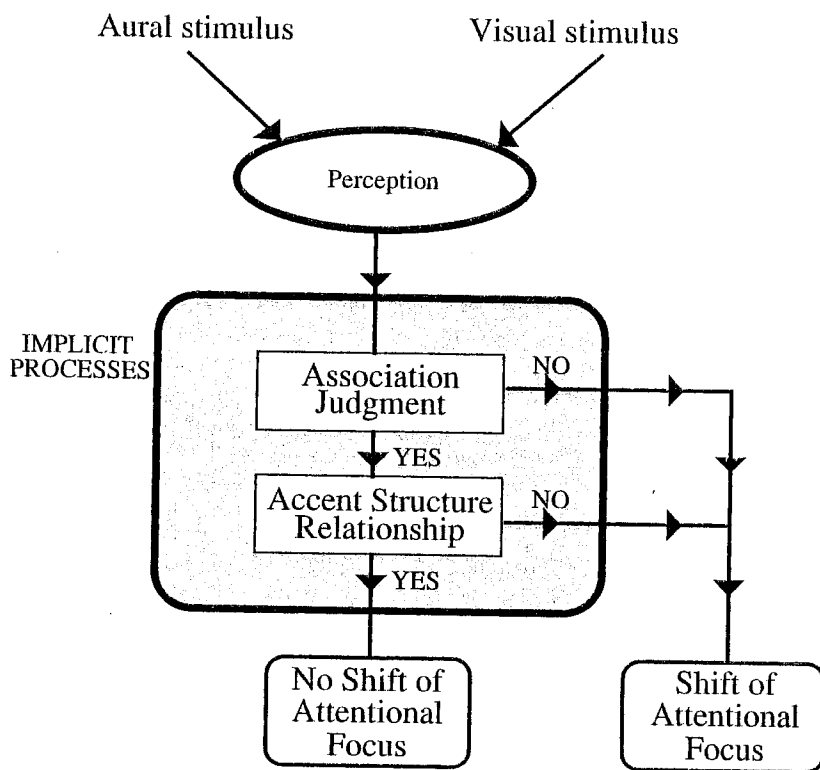


Figure 1. Lipscomb and Kendall's (1996) model of Film Music Perception.
Reprinted with permission of *Psychomusicology*.

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, for example, Heinlein (1928), Hevner (1935 and 1936), Farnsworth (1954), Meyer (1956), Wedin (1972), Eagle (1973), Crozier (1974), McMullen (1976), Brown (1981), and Asmus (1985).

The second implicit judgment (mapping of accent structures) consists of matching emphasized points in one perceptual modality with those in another. Lipscomb and 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.

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 can 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 on-screen action. 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, and Austin 1958).

Therefore, in the context of the decision-making process proposed by Lipscomb and 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: 380) 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.²

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 (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 AV 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: 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 and Jackendoff's (1983) generative theory of tonal music, Deliege (1987: 326) stated that "in perceiving a difference in the field of sounds, one experiences a sensation of accent." Boltz and Jones (1986: 428) propose that "accents can arise from any deviation in pattern context."

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 (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 perceived accent (melodic contour may change from ascending to descending or the direction of an object's 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.

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

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

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

"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 (two to seven years of formal music training), and highly trained (more than seven 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 files 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: 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, that is, 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 (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), yet it aligned with the other accent periodicities only once every 4 seconds, which is beyond Fraisse's (1982) upper limit for the perceptual linking of stimuli. (The specific relationships are: 1000ms = 20 frames; 800ms = 16 frames, and 500ms = 10 frames.) 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, and Carterette 1987; Lipscomb and 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. Juxtaposition 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 (synchronization of a 500ms temporal interval in both the audio and visual modalities) and nested consonance (synchronization of a 500ms temporal interval

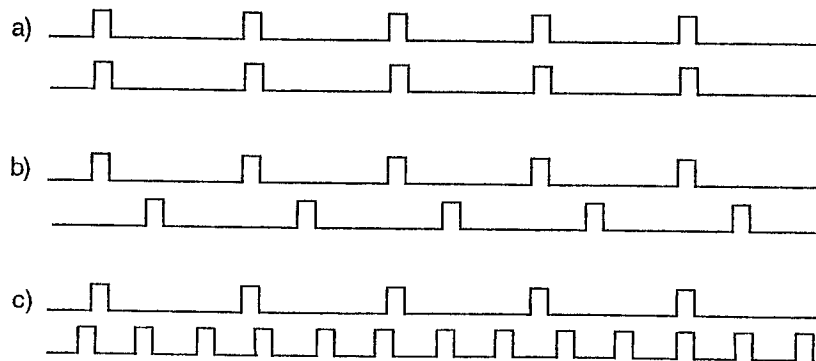


Figure 2. Visual representations of relationships between sources of accent.

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.

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 for incorporation into MEDS a module 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 thirty 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 (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 and 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 and Carterette 1992b and 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 (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/or the between-subjects variable: level of musical training.

Group Two was asked, in a paired-comparison task, 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, A-V composites—for placement within a multi-dimensional space (Kruskal 1964a, 1964b, and 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 (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 (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 Zetl (1990: 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 (for a detailed discussion of the exploratory studies, see Lipscomb 1995: 53–65). 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 and 4).

Subjects

Subjects for this experiment were forty UCLA students (ages 19 to 31) enrolled in general education classes in the Music Department, either Psychology of Music taught by Lipscomb in Fall, 1994 or American Popular Music taught by Keeling in



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.

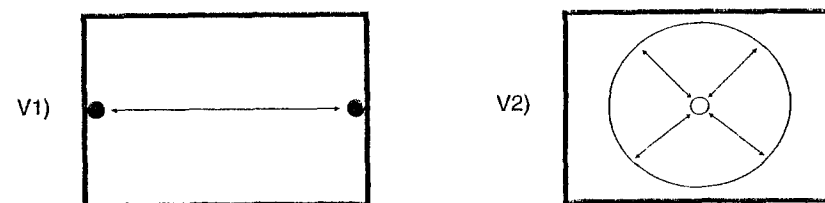


Figure 4. Visual stimuli selected for use in the Main Experiment. V1 exemplifies side-to-side continuous motion, while V2 illustrates apparent near-to-far-to-near continuous motion.

Summer Session II, 1994). The forty subjects (24 males and 16 females) 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.

Stimulus Materials

The A-V 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 (that is, *translation in the plane* along the x-axis), and V2 (Visual 2) represents front to back apparent motion (*translation in depth* along an apparent z-axis).