Cross-modal integration: Synchronization of auditory and visual components in simple and complex media

Dr. Scott D. Lipscomb

Institute for Music Research, University of Texas at San Antonio
6900 N. Loop 1604 West, San Antonio, TX 78249, USA
lipscomb@utsa.edu

Summary: Beginning in the 1950s, a series of psychophysical investigations revealed cross-modal influences using extremely simple auditory and visual stimuli. Most studies incorporating more complex stimuli have focused on the referential aspect of musical sound, i.e., the "cognitive congruency" of the music and the visual images (6). The present study investigates the alignment of accents (i.e., salient events) in the auditory domain with those in the visual domain and the effect of this alignment on subject perception of the A-V composite. The author reports results from two experiments utilizing varying levels of stimulus complexity: animations by Norman McLaren and motion picture excerpts. A-V alignment was manipulated as the independent variable with subject ratings as the dependent variable. A revision of Lipscomb & Kendall's model of film music perception (4) is presented, proposing a dynamic relationship between stimulus complexity and significance of A-V synchronization in the determination of subject ratings.

RELATED LITERATURE & RESEARCH QUESTION

Music has played an important role in the motion picture experience almost since its inception. Even so-called "silent films" were accompanied by musical performers. In recent years, there has been a significant amount of interest in the perceptual interaction between auditory and visual (A-V) systems in multi-modal contexts. Both psychologists and musicians are beginning to investigate the manner in which a stimulus perceived in one sensory modality may affect the cognitive processing of a stimulus in a separate modality.

To the present, there has been little empirical research studying the symbiotic relationship between the two perceptual modalities utilized when viewing a motion picture. In the field of perceptual psychology, interaction between the aural and visual modalities is well-documented (5, p. 12). Film music investigations, however, have focused almost exclusively on the *referential* (i.e., associational) aspect of music. A model of film music perception was proposed by Lipscomb & Kendall (4), suggesting that there are—at least—two implicit judgements made during the perceptual processing of the motion picture experience: an association judgement and a mapping of accent structures. The association judgement relies on past experience as a basis for determining whether the music is appropriate within a given context. The second implicit judgment (i.e., mapping of accent structures) is determined by the consistency with which important events in the musical score coincide with important events in the visual scene. The present study investigates accent structure alignment in animations by Norman McLaren and in excerpts from a motion picture by Brian DePalma entitled "Obsession" with a musical score composed by Bernard Herrmann. Results of previous research using these stimuli (2 & 3) suggest a dynamic relationship between the two implicit judgments delineated above. When viewing extremely simple auditory and visual stimuli (e.g., single-object animations accompanied by isochronous pitch sequences), accent alignment plays an important role in the determination of subject ratings of effectiveness. However, as the stimuli become more

complex, the importance of accent structure alignment appears to diminish and the association judgment assumes a dominant role. The present study addresses the question "Is it *necessary* for salient moments in the musical soundtrack to align precisely with salient moments in the visual image in order for the subject to consider the combination effective?

METHOD

A series of preliminary studies (5) assisted in the selection of three animation excerpts and three motion picture excerpts that were maximally different from one another. Excerpts from animations by Norman McLaren were taken from the Pioneer Special Interest laserdisc entitled "The World of Norman McLaren: Pioneer of Innovative Animation" (catalog number: PSI-90-018). Excerpts selected for use in the present study included 8-second portions of "Dots" (1940; frames 2285 to 2535), "Canon" (1964; frames 16500 to 16750), and "Synchromy" (1971; frames 41009 to 41259). Likewise, excerpts from the motion picture "Obsession" (1975) were taken from side 2 of a Pioneer Special Edition laserdisc (catalog number PSE91-18). Excerpts of the motion picture selected for use in the present study included the following 20-second scenes: "Portrait of Elizabeth" (00:04:47 to 00:05:07), "Flashback" (00:32:02 to 00:32:22), and "Reunion" (00:40:26 to 00:40:46). This Special Edition laserdisc recording provided a separate audio track that contained only the musical score, eliminating both dialogue and other ambient sounds.

Three alignment conditions were created for each excerpt: consonant, out-of-phase, and dissonant (7 & 8). As illustrated in Figure 1, *consonant* relationships (a) may be exemplified by accent structures that are perfectly synchronized. Accent structures that are *out-of-phase* (b) share a common temporal interval between consecutive points of emphasis, but the strata are offset such that they are perceptually out of synchronization. A *dissonant* relationship (c) occurs when the accent structures exhibit different temporal intervals between points of emphasis.

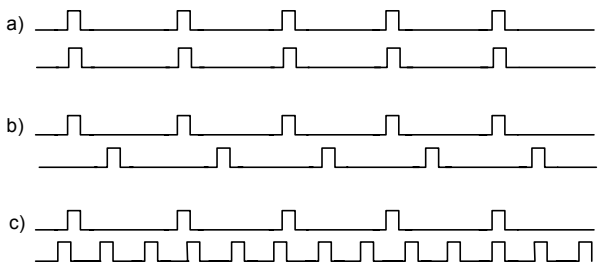

FIGURE 1. Visual representations of relationships between sources of accent.

In all of the excerpts described above, the consonant alignment condition consisted of the visual image and the musical score as intended by the composer or animator. The out-of-phase alignment condition was determined by a series of preliminary studies (5) in order to ensure that the alignment was perceptually as out-of-sync as possible. As a result, the audio track was adjusted for each excerpt as follows: "Dots" (-100 ms), "Canon" (532 ms), "Synchromy" (532 ms), "Portrait of Elizabeth" (672 ms), "Flashback" (-890 ms), and "Reunion" (425 ms). A positive number of milliseconds means that the audio accents followed the visual accents by the given temporal interval, while a negative number means that the audio accent preceded the visual accent by the given temporal interval. In previous investigations (2, 3, & 5), ecological validity of the stimulus materials was considered paramount and was maintained in the dissonant combinations by superimposing music from another part of the film over the visual excerpt. However, as a result, dissonant alignment conditions differed from the consonant and out-of-phase conditions not only in accent structure alignment, but also in other musically significant ways. To focus specifically on accent structure alignment in the

present study, dissonant alignment conditions were created by "time expanding" the audio track. Using Sonic Foundry's *Sound Forge* software and the "Time Expand/Compress" plug-in, the author was able to create temporally altered versions of the audio tracks without changing the perceived pitch of the music. The soundtracks for the McLaren animations were expanded to 115% of their original duration, while the audio tracks for the "Obsession" excerpts were expanded to 110% of their original duration.

Twenty UTSA students enrolled in music classes participated in this study. Subjects were categorized according to gender and level of musical training for consideration as grouping variables. Three levels of musical training were utilized: low training (less than two years of formal study and/or private lessons), moderate training (2 to 7 years), and highly trained (greater than 7 years). Stimuli were presented to groups of four to seven subjects at a time using a Samsung VR 5855 videocassette recorder and an RCA F276776BC 24-inch television monitor. Each group was assigned to one of three random presentation orders. Subjects responded to every A-V combination on two Verbal Attribute Magnitude Estimation (VAME) scales (1): "not synchronized—synchronized" and "ineffective—effective." In the instructions provided to subjects, *synchronization* was defined as "… how often important events in the music coincide with important events in the visual image." In contrast, *effectiveness* was defined as simply the "… subjective evaluation of how well the [music and visual image] go together."

RESULTS

A repeated measures analysis of variance revealed no significant difference in either between-groups variable: musical training ($F_{(2,17)}$=0.696; p=0.512); gender ($F_{(1,18)}$=0.457; p=0.508). However, the statistical analysis did reveal a highly significant within-groups difference between subject ratings as a function of the A-V combination ($F_{(17)}$=11.582; p<.0005). Figure 2 and Figure 3 represent the mean subject ratings of both VAME scales for the McLaren animations and "Obsession" excerpts, respectively. In these figures, consonant alignment conditions are identified by the title of the excerpt. Out-of-phase alignment conditions are identified by appending "_p" to the title and dissonant alignment conditions are identified by appending "_x" to the title.

As seen in Figure 2, the highest mean rating is consistently that given in response to the consonant alignment condition. The "Dots" excerpt provides a response trend seen consistently for both ratings of synchronization and effectiveness in the previous investigations (2, 3, & 5): consonant alignment rated highest, dissonant condition rated lowest, and out-of-phase condition rated in-between. However, VAME responses for "Canon" and "Synchromy" exhibit high ratings for consonant alignment conditions and extremely low ratings for both the out-of-phase and dissonant alignment conditions. In Figure 3, it becomes apparent that in actual motion picture excerpts, subject ratings of effectiveness are no longer consistently highest for consonant alignment conditions and lowest for dissonant alignment conditions. In fact, the out-of-phase alignment condition in the "Portrait of Elizabeth" excerpt is given a higher mean rating than the consonant alignment condition. Recall that the out-of-phase alignment conditions were created as a result of a preliminary study in which the A-V pair given the lowest synchronization rating was selected for use! "Reunion," however, does exhibit the same response trend as the McLaren animations. It would appear that, in this scene, synchronization of the musical beat with running footsteps and dramatic changes in musical content synchronized with changes in camera angle

provide explicit cues for synchronization that are not readily apparent in the other two excerpts. Note that relatively high levels of effectiveness are given to all alignment conditions for "Portrait of Elizabeth" and "Flashback," providing support for the suggestion that the associative content of the musical score plays a more significant role than accent structure alignment in such highly complex A-V stimuli.

As a result, future research in this area must attempt to develop a method for the quantification of complexity both within a given visual scene and within a given musical excerpt. Perhaps research in artificial intelligence can assist in the identification of "significant objects" within each modality and individual motion vectors can then be calculated for each of these objects, determining whether auditory motion vectors are coincident with visual motion vectors.

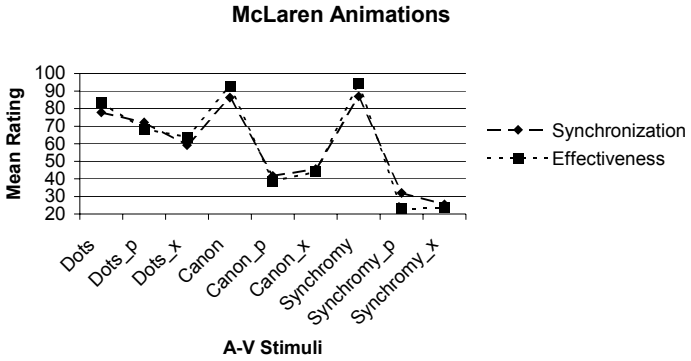

FIGURE 2. Mean VAME ratings for each animation.

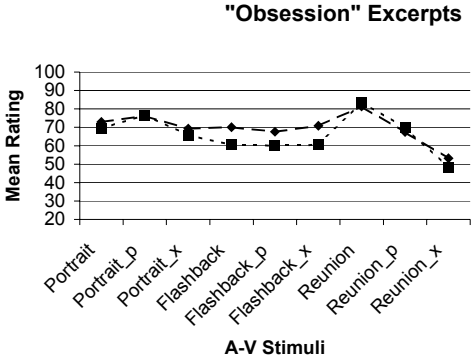

FIGURE 3. Mean VAME ratings for each motion picture excerpt

ACKNOWLEDGMENTS

The author wishes to express appreciation to the Institute for Music Research, The University of Texas at San Antonio, the UTSA Division of Music, and UCLA's Department of Ethnomusicology for the use of equipment and technology without which this research could not have been carried out. Special thanks to Dr. Alex Ramirez and his talented staff in the UTSA Office of Academic Technology for their assistance … without Gerard Bustos and Eric Quiroz, the multimedia materials utilized in this experiment and the conference presentation would have been of significantly lower quality.

REFERENCES

1. Kendall, R.A. & Carterette, E.C., *Music Perception* **10**, pp. 445-467 (1993).
2. Lipscomb, S.D, "Synchronization of musical sound and visual images: Issues of empirical and practical significance in multimedia development," in *Proceedings of the Conference of the Acoustical Society of America*, Norfolk, VA, 1998.
3. Lipscomb, S.D., *Journal of the Acoustical Society of America* **101**, p. 3190 (1997).
4. Lipscomb, S.D. & Kendall, R.A., *Psychomusicology* **13**, pp. 60-98 (1996).
5. Lipscomb, S.D., "Cognition of musical and visual accent structure alignment in film and animation," unpublished dissertation, *University of California, Los Angeles*, CA (1995).
6. Marshall, S.K. & Cohen, A.J., *Music Perception* **6**, pp. 95-112 (1988).
7. Monahan, C.B., Kendall, R.A., & Carterette, E.C., *Perception & Psychophysics* 41, pp. 576-600 (1987).
8. Yeston, M., *The stratification of musical rhythm*, New Haven, CT.: Yale University Press, 1976.