

UNIVERSITY OF CALIFORNIA

Los Angeles

**Cognition of musical and visual accent structure alignment
in film and animation**

A doctoral dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy in Music

by

Scott David Lipscomb

1995

© Copyright by
Scott David Lipscomb
1995

The dissertation of Scott David Lipscomb is approved.

Roger Bourland

Edward C. Carterette

William Hutchinson

James Thomas

Roger A. Kendall, Committee Chair

University of California, Los Angeles

1995

To my son, John David.

TABLE OF CONTENTS

	<u>page #</u>
LIST OF FIGURES	x
LIST OF TABLES	xiv
ACKNOWLEDGMENTS	xvi
VITA	xvii
ABSTRACT OF THE DISSERTATION	xix
CHAPTER ONE—INTRODUCTION	1
Research Questions	2
Film Music Background	3
Purpose and Significance of the Study	6
Basic Assumptions	7
Delimitation	8
Hypotheses	9
CHAPTER TWO—RELATED LITERATURE	12
Proposed Model and Its Foundation	15
Musical and Visual Periodicity	17
The Communication of Musical Meaning	19
A 3-Dimensional Model of Film Classification	23
Accent Structure Alignment	26
Determinants of Accent	26
Sources of Musical Accent	29
Sources of Visual Accent	34

A Common Language	38
Potential Sources of Musical and Visual Accent in the Present Study	39
Mapping of Audio-Visual Accent Structures	41
CHAPTER THREE—METHOD	42
Research Design	42
Subject Selection	43
Stimulus Materials	43
Experiment One	43
Experiment Two	45
Experiment Three	47
Exploratory Studies	47
Main Experiments	48
Alternative Hypotheses	51
Brief Summation	51
CHAPTER FOUR—EXPERIMENT ONE	53
Exploratory Study	53
Stimulus Materials	54
Auditory	54
Visual	54
Equipment	57
Subject Procedure	57
Data Analysis	58
Results	59

Statistical Analysis	63
Stimulus Selection for Experiment One	64
Main Experiment	65
Subjects	65
Equipment	66
Stimulus Materials	67
Stratification of Accent Structures	67
Experimental Task	68
Group One	68
Group Two	69
Results	70
Group One Data Analysis and Interpretation	70
Collapsing Alignment Conditions Across AV Composites	75
Analysis and Interpretation of Data Collapsed Across Alignment Con- ditions	78
Group Two	79
Data Analysis	79
Multidimensional Scaling	80
Cluster Analysis	80
Conclusions	82
CHAPTER FIVE—EXPERIMENT TWO	84
Exploratory Studies	84
Stimulus Selection	84
Establishing Accent Structure	85

Alignment Conditions	88
Main Experiment	92
Research Design, Hypotheses, and Subjects	92
Equipment	92
Stimulus Materials	92
Experimental Tasks	93
Results	94
Group One Data Analysis and Interpretation	94
Group Two Data Analysis	97
Multidimensional Scaling	97
Cluster Analysis	99
Conclusions	100
A Reassessment	102
CHAPTER SIX—EXPERIMENT THREE	104
Exploratory Studies	104
Stimulus Selection	104
Establishing Accent Structure	106
Alignment Conditions	109
Main Experiment	111
Research Design, Hypotheses, Subjects, and Equipment	111
Experimental Tasks	112
Results	112
Group One Data Analysis and Interpretation	112

Group Two Data Analysis	117
Multidimensional Scaling	118
Cluster Analysis	120
Conclusions	121
Confirmation of Perceived Difference Between Consonant and Out-of-Phase Composites	122
Interpretation of Data Analysis Across All Three Experiments	124
VAME Ratings	124
Synchronization	125
Effectiveness	127
Similarity Ratings	129
CHAPTER SEVEN—DISCUSSION, SUMMARY, AND CONCLUSIONS	132
Discussion	132
Limitations	139
Summary	141
Experiment One	144
Exploratory Studies	144
VAME Procedure	144
Similarity Judgment Procedure	146
Experiment Two	147
Exploratory Studies	147
VAME Procedures	148
Similarity Judgment Procedure	148
Experiment Three	149

Exploratory Studies	149
VAME Procedure	150
Similarity Judgment Procedure	150
Data Analysis Across All Experiments	151
Synchronization	151
Effectiveness	153
Conclusions	153
Confirmation or Disconfirmation of Null Hypotheses	153
Research Questions Answered	156
Suggestions for Further Research	158
ENDNOTES	163
REFERENCES	169

LIST OF FIGURES

Figure 2.1. Lipscomb & Kendall's (in press) model of Film Music Perception.	17
Figure 2.2. Scale used by composer Max Steiner to suggest geographical region in the opening titles of <u>Casablanca</u> .	23
Figure 2.3. Hypothetical example of films and animations placed within the proposed Lipscomb & Kendall (in press) Model of Film Music Classification.	25
Figure 2.4. Illustration of the creation of distinctive boundaries.	28
Figure 2.5. Visual illustrations of the Gestalt principles (after Dowling & Harwood, 1986, p. 154).	30
Figure 3.1. Visual representations of relationships between sources of accent.	46
Figure 4.1. Notation of the musical stimuli used in the exploratory study.	55
Figure 4.2. Visual stimulus patterns used in the exploratory study.	56
Figure 4.3a. Mean subject responses to the auditory portion of the exploratory study.	61
Figure 4.3b. Mean subject responses to the visual portion of the exploratory study.	61
Figure 4.4a. Equalized mean subject responses to the auditory portion of the exploratory study.	62
Figure 4.4b. Equalized mean subject responses to the visual portion of the exploratory study.	62
Figure 4.5. Scroll bar used to collect Group One subject responses.	69
Figure 4.6a. Mean subject ratings to the two VAME scales when viewing the combination of Visual #1 and Audio #1 across alignment conditions.	72
Figure 4.6b. Mean subject ratings to the two VAME scales when viewing the combination of Visual #1 and Audio #2 across alignment conditions.	73
Figure 4.6c. Mean subject ratings to the two VAME scales when viewing the combination of Visual #2 and Audio #1 across alignment conditions.	73

Figure 4.6d. Mean subject ratings to the two VAME scales when viewing the combination of Visual #2 and Audio #2 across alignment conditions.	74
Figure 4.7. Comparison of all VAME responses across AV composite and alignment conditions.	74
Figure 4.8. Standard deviations of mean responses to VAME scales across alignment conditions.	75
Figure 4.9. VAME ratings for all consonant, out-of-phase, and dissonant combinations across all AV composites.	77
Figure 4.10. Mean VAME ratings for Experiment One averaged across all levels of musical training.	79
Figure 4.11. Multidimensional scaling solution for the similarity judgments in Experiment One.	81
Figure 4.12. Cluster Analysis tree diagram—complete linkage (farthest neighbor)—for similarity ratings provided by subjects in Experiment One, Group Two.	81
Figure 4.13. Illustration of the mirroring relationship between elements in the upper cluster of composites incorporating Visual One.	83
Figure 5.1. Musical notation for audio excerpts of the Norman McLaren animations. Permission to use granted by Pioneer Entertainment (USA) L.P., based on their license from National Film Board of Canada.	86
Figure 5.2. Mean subject responses for the exploratory study to Experiment Two.	87
Figure 5.3. Equalized mean subject responses for the exploratory study to Experiment Two.	87
Figure 5.4. Mean subject VAME responses to the AV composites in Experiment Two.	95
Figure 5.5. Mean VAME responses from subjects in Experiment Two, collapsed across alignment condition.	96
Figure 5.6. MDS solution derived from mean similarity judgments in Experiment Two.	98

Figure 5.7. Cluster Analysis tree diagram—complete linkage (farthest neighbor)—for similarity ratings provided by Group Two subjects in Experiment Two.	100
Figure 5.8. Illustration of the mirroring relationship between the lowest six branches in the cluster tree diagram from Experiment Two (Figure 5.7).	100
Figure 6.1. Notation (reduced by author) representing excerpts from Bernard Herrmann’s musical score for “Obsession.” Permission to use granted by Hal Leonard Corporation and George Litto Productions, Inc.	107
Figure 6.2. Equalized mean subject responses to the auditory portion of the exploratory study.	109
Figure 6.3. Mean subject VAME responses to the AV composites in Experiment Three.	113
Figure 6.4. Mean synchronization ratings from subjects in Experiment Three, collapsed across alignment condition.	115
Figure 6.5. Mean effectiveness ratings for subjects of varying levels of musical training, collapsed across alignment condition (Experiment Three).	115
Figure 6.6. Mean ratings of effectiveness from Experiment Three, combining groups as appropriate.	117
Figure 6.7. MDS solution derived from mean similarity judgments in Experiment Three.	119
Figure 6.8. Line graph representing the mean responses in Table 6.5.	120
Figure 6.9. Cluster Analysis tree diagram—complete linkage (farthest neighbor)—for similarity ratings provided by Group Two subjects in Experiment Three.	121
Figure 6.10. Mean synchronization ratings for each alignment condition across all three experiments.	127
Figure 6.11. Mean effectiveness ratings for each alignment condition across all three experiments.	128
Figure 6.12. Relationship of mean similarity judgments for comparisons of identical stimulus pairs (i.e., consonant to consonant) and consonant-to-out-of-phase stimulus pairs in a) Experiment One, b) Experiment Two,	131

and c) Experiment Three.

Figure 7.1. Revised version of the Film Music Perception Paradigm.

136

LIST OF TABLES

Table 2.1. Potential sources of musical accent.	34
Table 2.2. Potential sources of visual accent.	37
Table 2.3. Proposed variables to be utilized in the initial pilot study labeled with direction.	40
Table 4.1. Relationship of <i>stimulus numbers</i> to <i>stimulus patterns</i> and IOIs.	59
Table 4.2a. The number of subjects that responded to the auditory stimuli in a manner that was either different from the hypothesized rate (i.e. not nested or subdivided), the spacebar presses did not occur at a steady rate (e.g. a pattern of long-short taps rather than a steady IOI), or the stimulus failed the Tukey pairwise comparison.	64
Table 4.2b. The number of subjects that responded to the visual stimuli in a manner that was either different from the hypothesized rate (i.e. not nested or subdivided), the spacebar presses did not occur at a steady rate (e.g. a pattern of long-short taps rather than a steady IOI), or the stimulus failed the Tukey pairwise comparison.	64
Table 4.3. The 14 alignment conditions for A-V composites in Experiment One.	66
Table 4.4. Number of subjects falling into each cell of the between-subjects design (Experiment One).	66
Table 4.5. AV composites arranged from highest response to lowest on the VAME scales.	75
Table 5.1. Excerpts of Norman McLaren’s animation used as stimuli in Experiment Two; all excerpted from Side 2 of the Pioneer Special Interest laserdisc entitled “The World of Norman McLaren: Pioneer of Innovative Animation” (catalog number: PSI-90-018; part of the Visual Pathfinders series).	85
Table 5.2. Number of subjects falling into each cell of the between-subjects design (Experiment Two).	92
Table 5.3. Abbreviations used to identify the various audio and visual components in Experiments Two.	93
Table 5.4. Labels used to identify each specific AV Composite based on the	93

abbreviations presented in Table 5.3.

Table 5.5a. Paired <i>t</i> -test values for mean synchronization ratings in Experiment Two (graphically represented in Figure 5.5).	103
Table 5.5b. Paired <i>t</i> -test values for mean effectiveness ratings in Experiment Two (graphically represented in Figure 5.5).	103
Table 6.1. Excerpts from side 2 of the Pioneer Special Edition laserdisc of “Obsession” (catalog number: PSE91-18).	106
Table 6.2. Abbreviations used to identify the various audio and visual components excerpted from “Obsession” for use in Experiments Three.	106
Table 6.3. Labels used to identify each specific AV composite to be used in Experiment Three; based on the abbreviations presented in Table 6.2.	111
Table 6.4. Number of subjects falling into each cell of the between-subjects design (Experiment Three).	112
Table 6.5. Mean subject similarity judgments to identities (comparing a consonant alignment with itself) and the consonant-to-out-of-phase comparison for each of the three AV combinations.	120
Table 6.6a. Paired <i>t</i> -test values for mean synchronization ratings in Experiment Three (graphically represented in Figure 6.4).	124
Table 6.6a. Paired <i>t</i> -test values for mean effectiveness ratings in Experiment Three (graphically represented in Figure 6.5).	124
Table 6.7. Number of subjects in each cell of the between-subjects design across all three experiments.	125

ACKNOWLEDGMENTS

The author would like to acknowledge the contributions of each and every member of my dissertation committee: Roger Kendall, Ed Carterette, William Hutchinson, James Thomas, and Roger Bourland. Without their guidance, the present work could not have been completed. Additional thanks to Dan McLaughlin and Phil in the UCLA Dept. of Animation for their expertise and access to the library of animation laser disks. A debt of gratitude is also due to Cheryl Keyes and Richard Keeling for assistance in recruiting subjects, Art Woodward & Eric Holman for statistical consultation, Christopher Husted for his knowledge of Bernard Herrmann's musical scores, Margaret Smith at Pioneer Entertainment, George Litto & Andria Nicole at Carlandria Publishers, Brenda Cline at Hal Leonard Corporation, John Hajda, Michael Harshberger, John O'Conner, David Martinelli, Kaye Lubach, Guiseppina Collicci, and all of the subjects who participated in the experiments described herein. The author recognizes the important role played by UCLA in the use of its facilities and continued financial support. Finally, and most important, I would like to express my admiration, respect, and sincere appreciation to Roger Kendall, the chair of my dissertation committee.

VITA

- August 2, 1959 Born, Lafayette, Indiana
- 1982 B.M., Music
specialization: Jazz Performance
University of Tennessee, Knoxville
- 1988–1993 Teaching Assistant & Research Assistant
University of California, Los Angeles
Los Angeles, CA
- 1990 M.A., Music
specialization: Systematic Musicology
University of California, Los Angeles
Los Angeles, CA
- 1990 to present Adjunct Faculty
Webster University
Vienna, Austria
- 1994 Temporary Adjunct Faculty
Southern Methodist University
Dallas, TX
- Teaching Fellow
University of California, Los Angeles
Los Angeles, CA

PUBLICATIONS AND PRESENTATIONS

- Lipscomb, S.D. (1989). Changing evaluations in audience perception of the motion picture experience. Paper presented at the meeting of the Society for Ethnomusicology, Southern California Chapter, Los Angeles, CA.
- Lipscomb, S.D. (1989). Film music: A sociological investigation into audience awareness. Paper presented at the meeting of the Society for Ethnomusicology, Southern California Chapter, Los Angeles, CA.
- Lipscomb, S.D. (1990). Music and visual components in film. Paper presented at the University of Southern California Music in Film Colloquium, Los Angeles, CA.

- Lipscomb, S.D. (1992). Perceptual judgment of the symbiosis between musical and visual components in film. Paper presented at the 2nd International Conference for Music Perception and Cognition, Los Angeles, CA.
- Lipscomb, S.D. (1994). Advances in music technology: The effect of multimedia on musical learning and musicological investigation. In D. Sebald (Ed.) Proceedings of the Conference on Technological Directions in Music Education, pp. 77-97. San Antonio, TX: IMR Press.
- Lipscomb, S.D. (1995). The personal computer as research tool and music educator. In K. Walls (Ed.) Proceedings of the 2nd Conference for Technological Directions in Music Education, pp. 169-173. San Antonio, TX: IMR Press.
- Lipscomb, S.D. (in press). Cognitive organization of musical sound. In D. Hodges (Ed.) Handbook of Music Psychology, 2nd ed. San Antonio, TX: Institute for Music Research.
- Lipscomb, S.D. & Hodges, D. (in press). Hearing and music perception. In D. Hodges (Ed.) Handbook of Music Psychology, 2nd ed. San Antonio, TX: Institute for Music Research.
- Lipscomb, S.D. & Kendall, R.A. (1995). Sources of accent in musical sound and visual motion. In I. Deliege (Ed.) Proceedings of the 3rd International Conference for Music Perception and Cognition, pp. 451-452. Liege, Belgium.
- Lipscomb, S.D. & Kendall R.A. (in press). Perceptual judgment of the relationship between musical and visual components in film. *Psychomusicology*, 13(1).

ABSTRACT OF THE DISSERTATION

**Cognition of musical and visual accent structure alignment
in film and animation**

by

Scott David Lipscomb

Doctor of Philosophy in Music

University of California, Los Angeles, 1995

Professor Roger A. Kendall, Chair

This investigation examined the relationship between musical sound and visual images in the motion picture experience. Most research in this area has dealt with associational aspects of the music and its affect on perception of still pictures or “characters” within film sequences. In contrast, the present study focused 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” (i.e. 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?

Three experiments were conducted using two convergent methods: a verbal attribute magnitude estimation (VAME) task and a similarity judgment task. Audio-visual (AV) stimuli increased in complexity with each experiment. Three alignment conditions were possible between the musical sound and visual images: 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 to the three alignment conditions. Consonant combinations were rated highest, followed by out-of-phase combinations, and dissonant combinations received the lowest ratings. However, as AV stimuli became more complex (Experiment Three), consonant composites were rated less synchronized and dissonant combinations were rated more synchronized than the simple AV composites in Experiment One. Effectiveness ratings failed to distinguish between consonant and out-of-phase conditions when considering actual movie excerpts. An analysis of variance over the VAME data from all three experiments, revealed that this difference between subject responses to simple animations and responses to complex film excerpts was statistically significant. A similar result was observed in the similarity scaling task. Responses to simple stimuli divided clearly on three dimensions: visual component, audio component, and alignment condition. However, as the AV composite became more complex, the third dimension appeared to represent AV congruence (i.e., appropriateness). Modifications to the proposed model of film music were suggested.

CHAPTER ONE

INTRODUCTION

In contemporary society, the human sensory system is bombarded by sounds and images intended to attract attention, manipulate state of mind, or affect behavior. The ubiquitous presence of television screens, providing aural and visual "companionship," whether one is standing in line at an amusement park or ordering a drink at the local pub. Patients awaiting a medical 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 to attract their specific clientele. Corporate training sessions and management presentations frequently employ not only communication through text and speech, but a variety of means for the purpose of attracting and maintaining attention, e.g. music, graphs, and animation. Even the latest generation of word processors for personal computers allows the embedding of sound files, charts, equations, pictures, and information from multiple applications within a single document.

Multimedia applications that utilize all of these capabilities are now commonplace in educational institutions and business offices. In each of the instances mentioned above, music is assumed to be a catalyst for establishing the mood deemed appropriate, assisting in the generation of actions desired, 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 often incorporate the use of music as a source of either relaxation or stimulation in leisure activities. With the increase in leisure time during the 1980s (Morris, 1988), products related to entertainment utilize music to great effect in augmenting the aesthetic affect of these experiences. Advertisement executives realize the impact music has on attracting a desired audience, e.g. the use of a rock beat to call baby-boomers to attention or excerpts from the classical repertoire to attract a more "sophisticated" audience.

A trip to the local movie theater will usually provide an excellent example of the manner in which a specially-composed musical score affects the perception of cinematic images. The present investigation studied this relationship of musical and visual components in animated sequences and motion pictures.

Research Questions

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.

Before considering this specific interrelationship, 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 sound-

track to line up precisely with points of emphasis in the visual modality in order for the subject to consider the combination 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.

Film Music Background

Combining music and visio-dramatic elements is hardly a new creation. In ancient Greece, music played a significant role in the dramatic tragedies of Aeschylus, Euripides, and Sophocles. Courtly displays of the Renaissance brought together many different artistic resources (e.g. singers, instrumentalists, dancers, scenery, costumes, stage effects, etc.) into a single spectacle for the sole purpose of aesthetic pleasure, as opposed to the religious significance ascribed to the Medieval sacred drama. However, the rebirth of a musico-dramatic form of entertainment for the masses must be credited to the opening of the first public opera house at Venice in 1637. The ideal of utilizing many artists and performers for a single dramatic purpose reached a pinnacle in the *Gesamtkunstwerk* of Richard Wagner, which Kuhn (1986) considered the harbinger of contemporary cinema.

Douglas Gomery (1985) asserted that movies combining motion pictures and sound did not appear "Minerva-like" in every theater of the roaring 20s. Rather, it was necessary to consider a period of about thirty years beginning in the final decade of the 19th century as technological innovations were advanced by the research and development teams at the American Telephone & Telegraph Corporation (AT & T) and the Radio Corporation of America (RCA). Seeking, respectively, to create better phone equipment and improve radio capabilities, equipment used in the recording and reproduc-

tion of sound was greatly enhanced (Gomery, 1985, p. 5). Initially, however, entrepreneurial experiments were the source of many inventive ideas.

Louis Lumière invented the cinematograph in 1894, followed by Thomas Alva Edison's introduction of his Kinetophone the following year. The latter did not attempt the synchronization of sound and image, but simply provided a musical accompaniment for the apparent motion produced by a sequence of still photos. In 1902, Léon Gaumont demonstrated his Chronophone, linking a projector to two phonographs and providing a dial for adjusting synchronization of sound and film. In 1913, a decade later, Edison was able to convince the operators of four Keith-Orpheum theaters in New York to install his updated version of the Kinetophone, with its complex system of belts and pulleys running between the stage and projection booth for purposes of synchronization. Finally, in 1923, Lee De Forest exhibited his Phonofilm system that placed the sound source directly on the film, eliminating the need for an external disc and its associated playing mechanism. This was a significant innovation, converting the light and dark portions of a soundtrack signal on the film itself into electrical pulses by means of a photoelectric cell. These pulses were then electronically amplified for reproduction in the theater.

The Jazz Singer (1927) popularized the notion of synchronized sound and lip movements. Four segments of Al Jolson singing in this picture paved the way for a 10-minute, all-talking comedy called My Wife's Gone Away (1928). The popularity of this new form of entertainment resulted in a boom for the fledgling industry, evidenced by a 600 percent increase in profits between 1928 and 1929 (Gomery, 1985, p. 23)!

Not everyone, however, considered the use of synchronized sound to be an advancement in the artistry of filmmaking. For advocates of "pure cinema" who found its artistic source "within the silent discourse of images that constituted a unique 'language'" (Weis & Belton, 1985, p. 75), the introduction of dialogue posed a serious threat. One

must, however, distinguish between "sound" and "dialogue." Critics were not against the use of *any* sound. After all, sound in the form of musical accompaniment had been an integral part of the motion picture experience since the early parlor demonstrations of the Lumière Brothers. Even Charlie Chaplin, a dedicated advocate of silent films, made use of sound effects and music as an expressive means of enhancing the effect of the image. Dialogue, however, threatened the "figurative language" of the film images.

A group of Soviet film directors published a statement in a 1928 Leningrad magazine, proposing that "the first experimental work with sound must be directed along the line of its distinct nonsynchronization with the visual images" (Eisenstein, Pudovkin, & Alexandrov, 1985). This approach, they believed, held potential for the creation of a sound film (as opposed to "talkies") in which the interrelationship of visual and aural images would emulate orchestral counterpoint as two elements of an audio-visual (AV) montage.

In 1929, Wright & Braun (1985) decried talkies simply as filmed stage plays. Rudolf Arnheim (1938/1985) maintained that, in composite media, one aspect must dominate the others. He asserted the primacy of image in film as opposed to the dominance of speech in theatrical productions. He believed that sound "reduces the power of the images to express information by privileging human speech over that of objects" (Weis & Belton, 1985, p. 81). Counterrevolutionary to Eisenstein, Clair, and the others, Arnheim suggested a return to the silent era of film, restoring prominence unquestionably to the image.

Though these theories provided creative and intriguing philosophical possibilities for the creation of interesting works of art, consideration of cinematic presentation in the current entertainment marketplace showed that synchronization of sound and image has,

without a doubt, become the dominant presentation style. In contrast to the theoretical positions cited above, Hutchinson & Kuhn (1993) asserted that


the visuals of early film, like those of the later video, evinced a link with music through temporality; their utilization of discreet frame and sequenced visual composition create dynamic rhythmic and harmonic visual structures easily analogous to music (p. 545).

The present study proposes that rhythmicity of both auditory and visual components and the ability to synchronize salient moments in the two perceptual modalities provides a means of increased expressivity, rather than the destructive result predicted by Eisenstein et al., Wright & Braun, and Arnheim. Synchronization of musical and visual accent structures was the focus of the following experimental investigation.

Purpose and Significance of the Study

The purpose of the present study was to develop and test a model of film music perception based on the stratification (i.e., layering) of accent structures. It is a useful contribution to the music perception literature because of its sociological significance, improved validity of stimulus materials, chosen frame of reference, and because it presents a model of film music perception, testing concepts that have never before been addressed within an experimental context.

In support of the claim for sociological significance, there is no doubt that the multi-billion dollar motion picture industry provides entertainment on a daily basis for hundreds of millions of individuals in almost every corner of the world. Any cultural artifact of such prominence deserves serious study. Investigations for the purpose of increasing our present understanding of the relationship between musical sound and visual images, as an important aspect of the film experience, are certainly worthy of pursuit.

A relevant innovation of the present study its progression from reductionist, simplistic musical and visual stimuli to actual movie excerpts. This marked an advanced improvement over AV materials used in previous investigations of film music (as discussed in the next chapter). By applying the same experimental method to stimuli varying in levels of ecological validity (i.e., how closely the experimental situation relates to experience in the real world), the following experiment allowed analysis of subject responses across these varying levels of complexity for cross-comparison. Therefore, answers to the research questions were provided at each level of analysis, prior to the final macroscopic discussion of results. 

Most obviously significant, perhaps, is the fact that the present study focused on an aspect of the motion picture experience that had never 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 (see Chapter Two). This is the expressed purpose of the following investigation. The long-term 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.

Basic Assumptions

The theoretical foundation of this investigation was built upon the assumption that musical sound has the ability to communicate (Campbell & Heller, 1980; Senju & Ohgushi, 1987; Kendall & Carterette, 1990). Expressive musical performance results in the transmission of musical information from a composer to a listener, often involving

the interpretation of a performer sharing the same implicit rule system. The communicated message is not always verbalizable, rather the meaning is frequently “embodied” in the relation of tones one to another (Meyer, 1956, p. 35). Normally, the role of music is not to communicate connotative or denotative messages. Spoken language serves this purpose quite sufficiently. Langer (1942) suggested, rather, that “music articulates forms which language cannot set forth” (p. 233). Expression of the underlying psychological drama in cinematic contexts provides a perfect function for this form of musical communication.

It was assumed that measurement techniques used in the present series of investigations (i.e., verbal rating scales and similarity judgments) provided reliable response metrics based on “sufficiently reliable and valid” results in previous experimental contexts (Kerlinger, 1965, p. 578). In collecting subject data, perception was considered to be the criterion frame of reference. To avoid intentionally the reification of physical characteristics of acoustical vibration or physiological processes in the transmission of sound energy, development of the present model of film music perception required a primary focus on human cognitive organization.

Delimitation

The present study was primarily experimental and, therefore, did not focus on music theoretical analyses or philosophical discourse, except as needed in discussing experimental results. Musical attributes (e.g., pitch, loudness, timbre, & rhythm) and visual attributes (e.g., color, size, shape, etc.) were mentioned as aspects of the various stimulus materials, but the low-level processes involved in their perception were left to psychophysicists. Finally, all audio-visual (AV) composites used in the main experiments of the present series of studies consisted of *only* the visual image and musical score. All other sound elements were removed (e.g., dialogue, sound effects, etc.) to

avoid providing subjects with unwanted cues, confounding the experimental design. Therefore, though the stimuli exhibited a much higher level of ecological validity than past investigations, they were still reduced from what is experienced when viewing a motion picture in a theatrical setting. In addition, the subjects rated the stimuli in a laboratory setting. The effect of this environment on human behavior has not been clearly established.

Hypotheses

The following series of null hypotheses were subject to statistical confirmation or disconfirmation.

1. There will be no significant difference between subjects' verbal ratings of *synchronization* on the basis of accent alignment of the AV stimuli.
2. There will be no significant difference between the subjects' verbal ratings of *effectiveness* on the basis of accent alignment of the AV stimuli.
3. There will be no significant difference between the subjects' verbal ratings of *synchronization* on the basis of level of musical training.
4. There will be no significant difference between the subjects' verbal ratings of *effectiveness* on the basis of level of musical training.
5. There will be no significant interaction between accent alignment condition and level of musical training in subject ratings of *synchronization*.

6. There will be no significant interaction between accent alignment condition and level of musical training in subject ratings of *effectiveness*.
7. When considering the entire data set—adding the level of complexity as a between-subjects factor—there will be no significant interaction between level of complexity, alignment condition and musical training in the ratings of *synchronization*.
8. When considering the entire data set—adding the level of complexity as a between-subjects factor—there will be no significant interaction between level of complexity, alignment condition and musical training in the ratings of *effectiveness*.
9. There will be no significant difference between subject *similarity judgments* as a result of the various AV composites.
10. There will be no significant difference in subject *similarity judgments* as a function of level of musical training.

Alternative hypotheses are offered as predicted outcomes of the series of experiments presented herein. In the VAME task, it is hypothesized that verbal ratings (relating to both synchronization and effectiveness) will be significantly different. The ratings will be highest for the synchronized combinations, lowest for the out-of-sync combinations, and intermediate for the dissonant relationship. It is also hypothesized that there will be a significant difference (for both VAME scales) between the response means of groups representing varying levels of musical training. If significant interactions between levels

of musical training and alignment condition are found for either VAME scale, post-hoc analyses will determine the source of this difference. It is also hypothesized that both VAME ratings of consonant composites will decrease as a function of stimulus complexity in the analysis of the entire data set.

In the similarity scaling task, it is hypothesized that subject ratings will differ significantly as a function of the various AV composites. It is also hypothesized that there will be a significant difference in responses between groups of subjects representing various levels of musical training . The analysis of these responses is predicted to produce an MDS solution consisting of three dimensions: musical stimulus, visual stimulus, and accent alignment.

CHAPTER TWO

RELATED LITERATURE

To the present, there has been little empirical work specifically directed at studying the symbiotic relationship between the two perceptual modalities normally used in viewing films (Lipscomb, 1990). In the field of perceptual psychology, interaction between the aural and visual sensory modalities is well-documented. Radeau & Bertelson (1974) found that when a series of lights and tones is presented at a 15° spatial separation, the location judgments for both the lights and the tones are biased toward the location of the stimulus in the other modality. Staal & Donderi (1983) showed that introducing a congruent aural stimulus (i.e. one synchronized precisely with the visual stimulus in its temporal and spatial aspects) lowered the interstimulus interval at which their subjects perceived continuous apparent motion of one light instead of partial motion or succession of motion between two lights. As a result, they concluded that the presence of sound may alter the perceived duration of a light flash by affecting visual persistence (see also Bermant & Welch, 1976; Ruff & Perret, 1976; Massaro & Warner, 1977; Regan & Spekrijse, 1977; and Mershon, Desaulniers, Amerson, and Kiever, 1980).

There is a paucity of research available considering more complex audio-visual (AV) combinations. Three studies have utilized ecologically valid contexts in the consideration of the motion picture experience. Tannenbaum (1956), using Osgood, Suci, and Tannenbaum's (1957) three factors (i.e. Evaluative, Potency, and Activity) to collapse bipolar adjectives used in semantic differential scaling,¹ found that music does

influence the rating of dramatic presentations whether presented live on stage, in a studio taped version, or in a recorded version of the live performance. His results showed that the influence of music was most pronounced on the subject responses related to Potency and Activity dimensions. Overall evaluation of the play did not change significantly.

However, there were several problems concerning the musical aspect of his stimulus presentation. Firstly, the musical selection was not composed for the specific scene that it accompanied. In a rather ambiguous explanation of the selection process, Tannenbaum explained that the piece was chosen by a person who had "considerable experience in this kind of work" and confirmed by a panel of four "experts" (Tannenbaum, 1956, p. 96). The most serious problem, however, was the method employed to synchronize the audio and visual stimuli during the performances. A phonograph recording was played along with the visual image. As a result, synchronization of the dramatic action and musical accompaniment was left largely to chance, a procedure not at all representative of the relationship occurring in a well-edited motion picture. Finally, in an attempt to make the music seem more compatible with the visual action, during certain scenes the loudness was manually increased "for dramatic impact" (Tannenbaum, 1956, p. 96). This is hardly an acceptable substitute for a soundtrack composed specifically for the scene under investigation.

In a second study of interest, Thayer & Levenson (1984) recorded five different physiological measurements during exposure to a 12-minute black and white industrial safety film depicting three accidents. These measures included the subjects' interbeat interval of the heart, general somatic activity, skin conductance level (SCL), pulse transmission times to the finger and the ear, and finger pulse amplitude. In addition, the subject was asked to provide a continuous self-report of anxiety level by turning a dial on which the extremities were labeled "extremely calm" and "extremely tense." Two musi-

cal scores were composed for presentation with the film for comparison with the responses to a control (i.e. no music) condition. The "documentary music" was described as a mildly active progression of major seventh chords, purposely intended not to draw attention toward or away from any specific part of the visual scene. The "horror music" was described as a repetitive figure based on diminished seventh chords utilizing harsh timbres. In addition to the differences in musical style, placement of the music in the film differed radically between the two music conditions. While the "documentary music" was present throughout, the "horror music" was edited so that it preceded the first accident by 20 seconds, the second accident by 10 seconds, and the final accident by 30 seconds. In each instance, the music ended approximately 10 seconds after the accident at a natural cadence point (both musically and visually). Though the film produced significant responses in all five of the physiological measures when compared with subjects' pre-exposure levels, only SCL differentiated the three film score conditions. From this result, the investigators claimed to have provided "preliminary experimental support for the efficacy of musical scores for manipulating the stressfulness of films" (p. 44). Recall, however, that the subjects' continuous self-reports of *perceived* anxiety level did not differentiate between the three film conditions. Therefore, though Thayer & Levenson concluded from their data that the use of music caused either a heightened or reduced electrodermal response to the stressful stimuli, more evidence was needed to support a claim for the ability of a musical score to manipulate the stressfulness of a film in terms of the subjects' emotional responses.

In a third study, Marshall & Cohen (1988) selected a film utilizing abstract animation. They were interested in determining whether the information provided by a musical soundtrack would affect the judgments of personality attributes assigned by subjects to each of three geometric shapes presented as "characters" in the film. Marshall

composed two distinct soundtracks (described as "strong" and "weak") varying on a number of musical dimensions, e.g. major/minor mode, fast/slow tempo, high/low pitch, and single- or multi-note texture. Each soundtrack consisted of three main "themes." Synchronization of the aural and visual elements was kept constant by editing the soundtrack directly on to the videotape and the authors provided a brief description of the action occurring at the point when each of the themes was introduced. However, apart from the beginning point, no information was provided concerning the specific interaction of the aural and visual stimuli.

A second problem with these musical compositions is that, in their simplicity of content and repetitive structures, they failed to provide an accurate representation of the highly developed craftsmanship that goes into a typical movie score. Even using this limited musical vocabulary, however, the results were similar to those compiled by Tannenbaum (1956). In comparing five film conditions (i.e. film alone, "weak" music alone, "strong" music alone, "weak" music-film, and "strong" music-film), meaning of the music was found to be closely associated with the film on the Potency and Activity dimensions. Evaluative judgments, on the other hand, depended on a complex interaction of the musical and visual materials. However, in Marshall & Cohen's (1988) investigation, lack of validity seriously limits the ability to generalize results to actual motion pictures.

Proposed Model and Its Foundation

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. It is quite possible that, when watching a film, perception of the musical component remains almost completely at a subconscious level (Lipscomb, 1989).

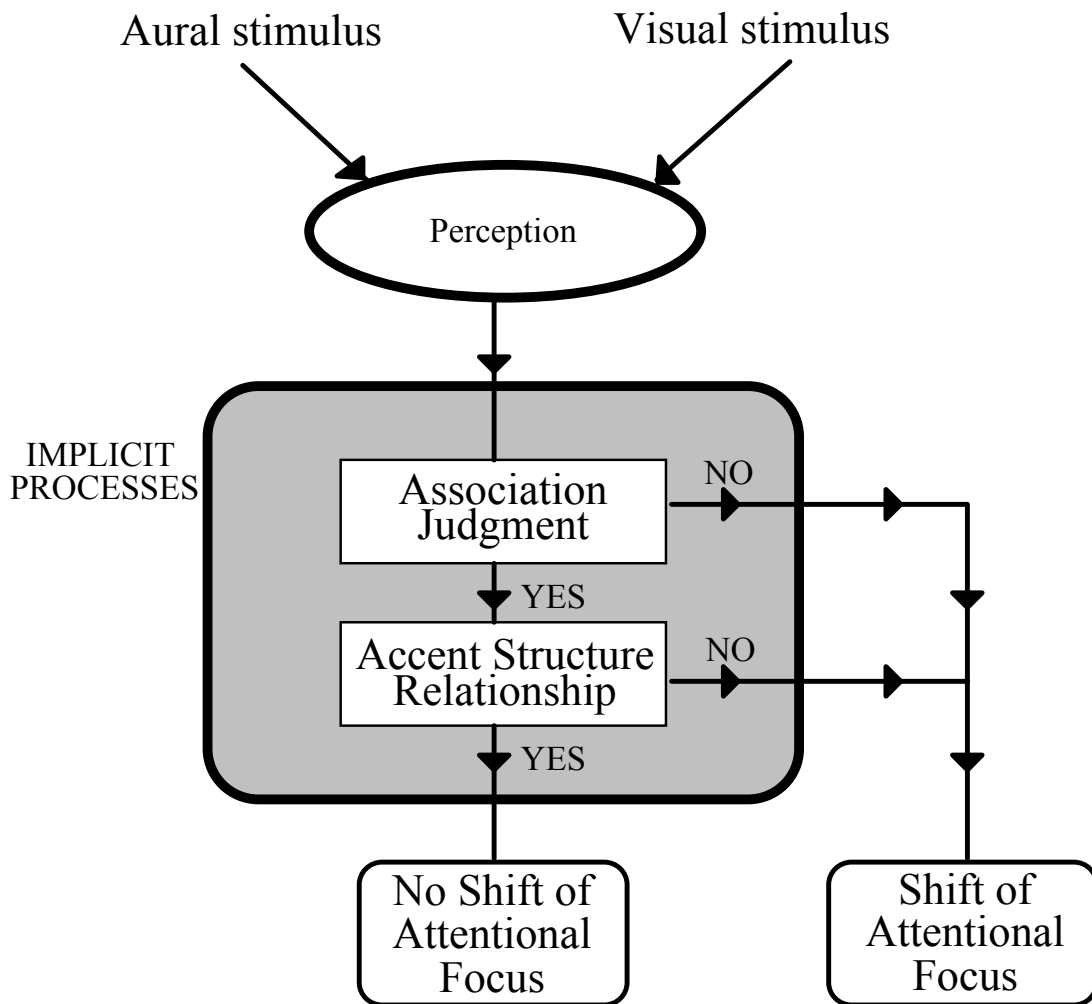
In the study mentioned earlier, Marshall & Cohen (1988) provided a paradigm to explain the interaction of musical sound and geometric shapes in motion entitled the "Congruence-Associationist model." They assumed that, in the perception of a composite AV presentation, separate judgments were made on each of three semantic dimensions (i.e. Evaluative, Potency, and Activity; see Osgood et al., 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 (in press) suggested 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 2.1). 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 (in press) proposed that, if the associations identified with the musical style were judged appropriate and the relationship of the aural and visual accent structures

were consonant, attentional focus would be maintained on the symbiotic composite, rather than on either modality in isolation.

Figure 2.1. Lipscomb & Kendall's (in press) model of Film Music Perception.



Musical and Visual Periodicity. The preceding paragraphs have emphasized not only referential aspects of music, but also the hypothesized importance of accent structure alignment between the aural and visual strata in movie contexts. There are 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 the 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. 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).

In the context of the decision-making process proposed by Lipscomb & Kendall (in press), 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, 1928; 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

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)

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 could determine the tolerance for misalignment.

The Communication of Musical Meaning. Many authors have claimed that music changes the "meaning" of a film (Tannenbaum, 1956; Harrell, 1986; Gorbman, 1987; Marshall and Cohen, 1988; Lipscomb, 1990), often leaving the term "meaning" undefined. Lipscomb & Kendall (in press) provided an operational definition within the context of their semantic differential study as "an indication of the subjects' changing evaluation of the audio/visual composite scaled in a semantic space." There are several possible means by which music can communicate meaning in a more general sense.

Papers of Charles Peirce (1931-35, Vol. 2) provided a classification of three types of signs used in the communication process, differentiated by the way in which they

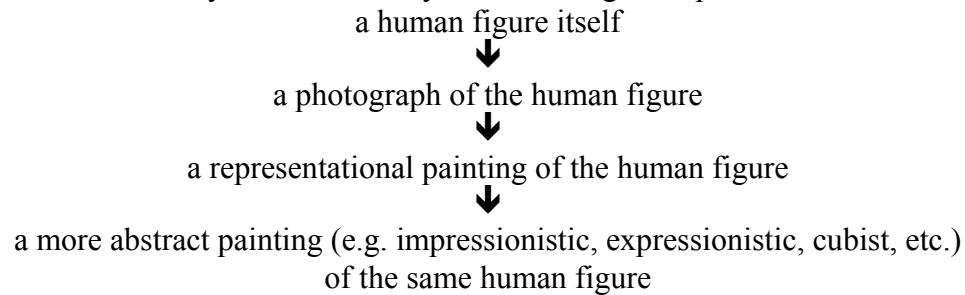
represented their referent. Dowling and Harwood (1986) provided the following explanation of Peirce's delineation:

An index represents its referent by having been associated with it in the past, as lightning and thunder with a storm. An icon represents through formal similarity to the referent, as a wiring diagram represents a circuit. Symbols represent by being embedded in a formal system, such as language
(p. 203).

Examples of musical scores illustrating each of these types of signification will be provided below.

Leonard Meyer (1956) provided a classification of meaning that applies specifically within a musical context. He referred to *referentialists* as those who believe that musical meaning is derived from its ability to refer to objects, ideas, and actions outside of the music itself. In contrast, musical *expressionists*, while considering music essentially intramusical (i.e. areferential), argue that the musical relationships "are in some sense capable of exciting feelings and emotions in the listener" (Meyer, 1956, p. 3). Both referential and expressionist aspects of meaning are important to the film composer. For instance, the general conventions mentioned earlier (e.g. "romantic" string lines, "majestic" brass, etc.) assist the composer in creating music considered appropriate for a given setting. Using these conventions, the composer must rely on the listener to provide the appropriate emotional response, based on past experience. Similarly, the relationship between musical sounds themselves is important in supplying a means for the musical expression of psychological drama (e.g. tension expressed musically as dissonant chord clusters) and in providing a means of unification (e.g. the recurring theme or motive that becomes associated with a character or idea).

It is important to understand that referentialism and expressionism are not mutually exclusive. The two concepts may be thought to lie on a continuum between abstract and concrete forms of musical expression (Lipscomb & Kendall, in press). This continuum mirrors the relationship exemplified within the visual arts in which a sequence from concrete to abstract may be illustrated by the following example:



Attributes of the object are removed (or made more abstract) as the process continues from concrete to abstract, until only the criterial attributes² remain. The object is no longer **the** fountain, but **a** fountain, relying on an observer to fill in the details.

A film excerpt in which the director exploits this distinction for dramatic purpose occurs in Arthur Freed's An American in Paris (1951). Initially, a sketch of a human figure is shown on screen. Gradually, an image of Gene Kelly assuming the same position as that represented in the drawing is superimposed onto the picture as the sketch fades out. This process gives the illusion that the inanimate drawing has "come to life" (i.e. become less abstract or, conversely, more concrete).

As in the visual arts, musical communication occurs with varying types and degrees of abstraction. Both iconic and indexical signs may be considered referential means of musical communication. *Icons* provide an example of the most concrete means of transmitting musical meaning by emulating the physiognomic structure of physical motion. John Booth Davies (1978) described *physiognomic perception* as "certain states

or conditions of both human and non-human objects [that] seem inherently to express particular qualities, because they possess the same kind of structure" (p. 104). He illustrated this notion by suggesting that a willow tree may look "sad," because both the tree and a sad person appear passive and droopy. This type of signification is utilized frequently in animated cartoons. Inevitably, as a character falls toward certain destruction, the musical score will incorporate a descending melodic line or glissando.

The two phrases representing the bird character in Prokofiev's *Peter and the Wolf* provide an example of two types of meaning. The first phrase is an imitation of birdsong that serves as an *index* "pointing to" the object "bird," while the second phrase may be thought to provide an iconic (i.e. physiognomic) representation of the motion of a bird in flight. Indexical meaning moves one step toward abstraction since accurate transmission of the intended meaning relies heavily on a previously established association between musical sound and another (often extramusical) object or idea.

Two outstanding examples of indexical meaning come from Hal B. Wallis' classic film Casablanca (1942; musical score by Max Steiner). During the opening titles, the initial phrase of music was based on a scale consisting of semitones and augmented seconds that is intended to "point to" the foreign locale (Figure 2.2). Indexical meaning was utilized in the musical score throughout this film as a means of distinguishing between the French populous and German soldiers occupying the city. A minor key rendition of "Deutschland über alles" became associated with the German antagonists, while "La Marseilles" came to represent the French citizens.

Perhaps the most effective musical use of indexical meaning ever used in a motion picture occurred in a bar scene (at "Rick's") when the conflict between these two nationalist groups was expressed musically. Both of these referential melodies occurred simultaneously as the French patrons drowned out the soldiers' rendition of "Deutschland

über alles” by boldly singing their own national anthem. It is defensible to say that the tension during these moments, in which no dialogue and very little action occurred, was unsurpassed in any other scene of the movie.

Figure 2.2. Scale used by composer Max Steiner to suggest geographical region in the opening titles of Casablanca.



The highest degree of musical abstraction, however, is fulfilled by works that incorporate *absolute expressionism*, relying solely on the relationship between musical sounds for meaning. For an example of a composite musico-dramatic form that fully exemplifies this type of musical communication, one may turn to some form of modern dance or ballet that makes no attempt at story-telling. In this context, the various bodily movements and musical sounds take on a significance purely because of the syntactical relationship of one sound or movement to another. This relationship is directly affected by the interaction between the two modalities (i.e. sight and sound) and the alignment of their accent structures. Within the context of a motion picture varying degrees of *iconic*, *indexical*, and *symbolic* signs are used to communicate the composer's intended musical message.

A 3-Dimensional Model of Film Classification. Using both the associative and syntactical aspects of Lipscomb & Kendall's (in press) model, it would theoretically be possible to place a given AV composite within a 3-dimensional space defined by the degree of visual referentialism, the degree of musical referentialism, and the degree of synchronization between audio and visual accent structures. The referential dimensions

must be considered separately due to the fact that the level of referentialism in the visual image is independent of the level of referentialism in the music.

In contrast, accent structure alignment between the two modalities may be incorporated into a single relationship, because the degree of synchronization will be determined by the *relationship* of each dimension to the other. The visual reference dimension may be represented by a continuum from highly referential (e.g. realistic, representational images) to highly areferential (e.g. abstract imagery with no previous associative "meaning"). The musical reference dimension may be similarly represented by a continuum from highly referential (e.g. music with an unambiguous association) to highly areferential (e.g. music with no previous associative "meaning"). The accent structure alignment dimension may be depicted by a continuum from precisely matched accent structures (e.g. known as "Mickey Mousing" because of its predominance in animated cartoons) to asynchronous combinations.³

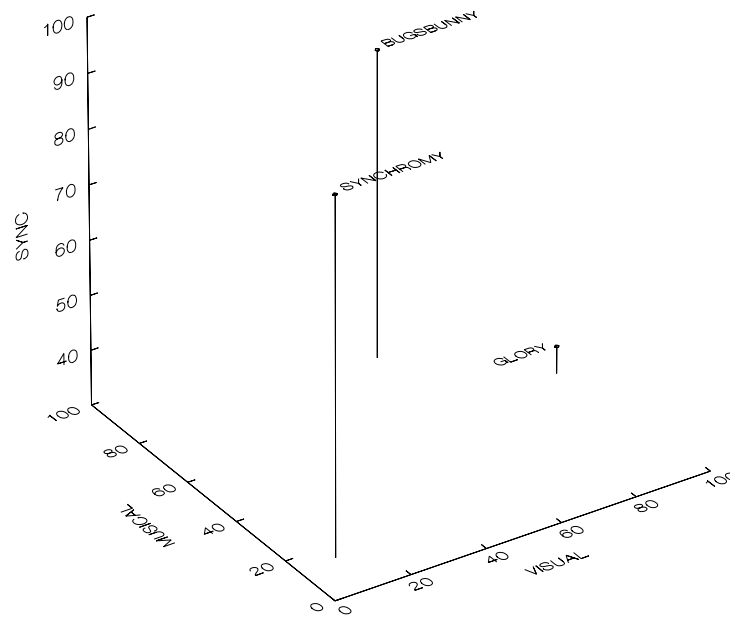
The first two dimensions derive their designation from Leonard B. Meyer's *referentialism*, while the matching of accent structures may be considered a form of *absolute expressionism*, in which the musical meaning is considered to "arise in response to music ... without reference to the extramusical world of concepts, actions, and human emotional states" (Meyer, 1956, p. 3).

A visual representation of the resulting 3 dimensions of Lipscomb & Kendall's model is provided in Figure 2.3. Three examples from the film repertoire are represented within this area. Norman McLaren's "Synchrony" (1971) is a piece of experimental animation in which abstract shapes appear on the screen as identical images pass over the photoelectric cell on the "sound track" portion of the film celluloid. The resulting tones are not intended to have a conventional association for the typical viewer. Therefore, the level of referentiality is quite low in both the visual and the musical dimensions. Since,

in this case, “meaning” is derived largely as a function of synchronization between the sound and images, it is placed extremely high on the dimension of AV synchronization. The second film example in Figure 2.3 represents one of the final scenes from Edward Zwick & James Horner’s “Glory” (1989) in which a troop of soldiers prepares to march to war as the themes that have become associated with the various characters and situations throughout the film are heard on the soundtrack. This scene is extremely “concrete” (i.e. low degree of abstraction or high degree of referentiality) in its representation of human drama while the music is referential both thematically and stylistically. However, the synchronization of accent structures is relatively low throughout the scene. The third point plotted on the graph represents a typical Bugs Bunny cartoon, in which the visual images are fairly representational, the music is loaded with indexical meaning, and synchronization between the music and visual images is extremely high.

As stated previously, every past investigation into the role of film music has dealt exclusively with the referential aspects of this model. The purpose of the present investigation was to develop and test a method of quantifying the degree to which the auditory and visual strata align, providing a rating system for AV accent structure synchronization.

Figure 2.3. Hypothetical example of films and animations placed within the proposed Lipscomb & Kendall (in press) Model of Film Music Classification.



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 was possible to describe an A-V composite in terms of accent strata. It was this stratification of accent structures and the relationships of these strata one to another that was investigated in the present study.

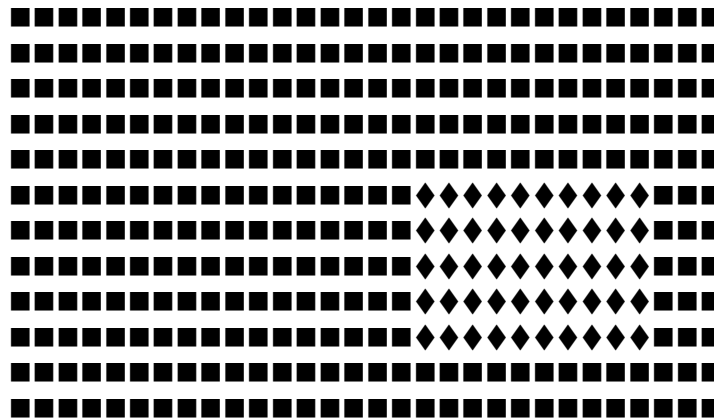
Determinants of Accent. In the search for determinants of accent, potential variables were established by considering the various aspects of visual objects and musical phrases that constituted perceived boundaries. Fraisse (1982, p. 157) suggested that grouping of constituent elements results "as soon 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). Though none of these statements provided a truly operational definition of accent, it was clear that the identification of sources of accent had to be considered from the perceptual frame of reference. Hence, the purpose of the following paragraphs was to determine parameters that were capable of change in both the auditory and visual domains, resulting in perceived accent..

The idea that variation in a stimulus stream creates a point of emphasis fits well with the basic tenets of information theory (Shannon & Weaver, 1949). A system exhibiting a high level of *redundancy* (e.g. a 1 KHz sine tone with a duration of 250ms sounding one time per second for a long period of time) provided no reason to assume that a change would occur, though expectations generated in a listener may anticipate that, at sometime in the future, the *certainty* of this series of pips will be dropped to a lower level through the introduction of a change. If, at some point during the period of time in which this stream of pips continues, a silent interval of 1.25 seconds was introduced between tones (creating an interonset interval of 1.5 seconds), information content at that moment became higher to any listener who had endured the series of equally-spaced, equi-durational tones and a perceived accent point was created. If, however, the 1.5 second interonset interval (hereinafter referred to as IOI) replaced the 1 second IOI at a periodic

interval (for instance, after every tenth tone), then segregation of the sound stream into units of ten tones occurred and one could speak of the longer temporal separation as a boundary between larger groups of ten tones (Vos, 1977; Povel & Okkerman, 1981; Monahan & Carterette, 1985).

In the visual domain, Ann Treisman and her colleagues (Treisman & Gelade, 1980; Treisman, 1982; Treisman, 1988; Treisman & Gormican, 1988) and Julesz (1984) found that distinctive boundaries were created from *elementary properties*, such as brightness, color, and orientation of line. Figure 2.4 provides an illustration of a rectangular boundary created by differences in shape (i.e. orientation of line). Notice how there appears to be a rectangular box of diamonds ("◆") that stands out clearly from the surrounding "■"s.⁴ Treisman and Julesz proposed the existence of two distinct processes in visual perception. A *preattentive process* initially scans the field of vision for the useful elementary properties of the scene, e.g. presence of objects and overall features of these objects (color, orientation, size, and direction of movement). During this stage, variation in any of the elementary properties may be perceived as a *border*. However, complex differences resulting from combinations of simple properties are not detected until the later *attentive process*, that directs the "spotlight of attention" to specific features of an object, emphasizing the salient features while repressing those objects and features that are considered to be of lesser importance. This form of data reduction is a necessary function of the perceptual process, allowing selection of those elements of sensory input at any given moment that are deemed worthy of attention, filtering out those that are less relevant to the present activity.

Figure 2.4. Illustration of the creation of distinctive boundaries.



Such perceptual organization has been the focus of a large body of research inspired by the work of Gestalt psychology. In contrast to the stimulus-response model of the contemporaneous Behaviorist school, the Gestalt psychologists considered the perceiver as a gatherer *and interpreter* of data from the environment. This holistic approach to perception was based on the ability of a human observer to organize individual elements into complete units and structures. The Gestalt principles of proximity, similarity, common fate, and good continuation (*Prägnanz*) provided a basis for the organization of perceived objects into groups.

Köhler (1929) provided visual illustrations of these principles similar to the diagrams shown in Figure 2.5. Deutsch (1982) and Dowling & Harwood (1986) provided musical examples of these grouping principles. The following discussion will focus on the determination of musical and visual parameters that successfully formed perceptual boundaries (i.e. emphasized points in the stream of sensory input), providing a theoretical basis for generating a set of stimuli for utilization in the testing of perceptual tolerance

for alignment and misalignment of these boundaries between the auditory and visual strata.

Sources of Musical Accent. In 1938, Carl E. Seashore stated that there are four, *and only four*, physical characteristics of a sound: frequency, amplitude, duration, and signal shape (p. 16). His model of musical sound proposed that these attributes correlated isomorphically with perceptual variables: pitch, loudness, duration, and timbre, respectively. Although each of these parameters may be used as a source of accent (or grouping boundary), it is important to emphasize that these variables are not orthogonal. In fact, experimental studies have shown that there is a high level of interaction between the elements in Seashore's model (see Kendall, 1987, for a thorough discussion).

Cooper & Meyer (1960; cited in Monahan & Carterette, 1985) identified three types of accent used in the perceptual grouping of musical sounds. A *dynamic* accent provides emphasis by means of a higher intensity level, i.e. the more intense a tone, the louder it sounds and the more it seems to be dynamically accented. An *agogic* accent results with temporal variation. In this case, a tone is durationally (or temporally) accented when it has a longer duration than surrounding tones or when its expected onset is delayed up to a certain point. Finally, within a tonal system, certain tones (e.g. the tonic and dominant) are given more *melodic* weight.

Figure 2.5. Visual illustrations of the Gestalt principles (after Dowling & Harwood, 1986, p. 154).

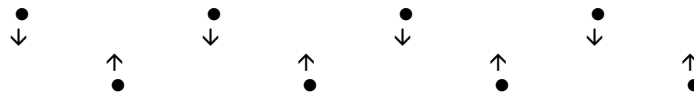
a) proximity



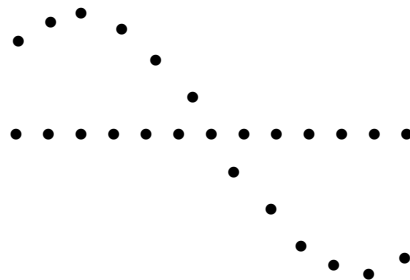
b) similarity



c) common fate



d) good continuation (*Prägnanz*)



Four physical factors were proposed by Thomassen (1982) as sources of perceived accent. *Dynamic accentuation* results with an increase in sound level. Accent through *temporal differentiation* results when durations are varied and, he pointed out, is the major source available to the harpsichordist for stressing points of musical significance. *Melodic accentuation* is due to a tone's hierarchical position within the tonal system. To these three sources of accent (similar to those delineated by Cooper & Meyer, 1960), he added *harmonic accentuation*, resulting from variation in the spectral envelopes of successive complex tones.

In addition, he cited a study by Ortmann (1926) that showed how *serial position* interacts with the length of a melodic sequence to create a point of accent. The first tone

of a short sequence was perceived as accented, while the last tone of longer sequences received the accent. Thomassen's experimental data confirmed that melodic contour provides the strongest source of perceived accent (when IOI, tone duration, amplitude envelope, and spectral characteristics are controlled), followed by interval size, and (perhaps) a slight difference in ascending and descending pitch intervals. He concluded that successive changes in pitch tended to accentuate the middle note if the changes are in opposite directions (e.g. C4 **F4** C4). However, if the pitch changes were in the same direction, the impression of accent fell equally on the 2nd and 3rd tones (e.g. C4, **E4**, **G4**). He also determined that when the size of two successive intervals clearly diverged within the context of a 3- or 4-tone melodic sequence, the largest of the intervals was perceived as accented. The effect of this relative magnitude (i.e. interval size) was, however, less pronounced than the effect of melodic contour (see Monahan & Carterette, 1985).

Drawing heavily from Noam Chomsky's (1965 & 1975) linguistic generative-transformational grammar, Lerdahl & Jackendoff (1983) proposed three types of musical accent. A *phenomenal accent* is "any event at the musical surface that gives emphasis to a moment in the musical flow" (p. 17). This would include attack points and relatively consonant pitch events. *Structural accents* result at melodic and harmonic points of gravity within a phrase (e.g. cadences and harmonic resolutions). The final type of accent falls on those tones that occur on strong beats in the metrical hierarchy and is called, accordingly, *metrical accent*. Lerdahl & Jackendoff provided grouping preference rules to explicate specific musical events that result in perceptual grouping (pp. 345-6); two of which are quoted below:

GPR2 (Proximity)

- a. (Slur/Rest) the interval of time from the end of n2 to the beginning of n3 is greater than that from the end of n1 to the beginning of n2 and that from the end of n3 to the beginning of n4, or if
- b. (Attack-Point) the interval of time between the attack points of n2 and n3 is greater than that between the attack points of n1 and n2 and that between the attack points of n3 and n4.

GPR3 (Change)

- a. (Register) the transition n2-n3 involves a greater intervallic distance than both n1-n2 and n3-n4, or if
- b. (Dynamics) the transition between n2-n3 involves a change in dynamics and n1-n2 and n3-n4 do not, or if
- c. (Articulation) the transition between n2-n3 involves a change in articulation and n1-n2 and n3-n4 do not, or if
- d. (Length) n2 and n3 are of different lengths and both pairs n1, n2 and n3,n4 do not differ in length.

These grouping preference rules appear to agree with the conclusions of Cooper & Meyer (1960) and Thomassen (1982), adding the additional parameter of articulation. This additional source of differentiation stressed the fact that an IOI results, in most cases, from the combination of a tone duration interval and a silent interval (Povel & Okkerman, 1982; Monahan, Kendall, & Carterette, 1987).

Drake, Dowling, & Palmer (1991) utilized four sources of musical accent in their study of the effect of accent structure alignment. *Metric accent structure* results in periodically spaced points of emphasis due to the meter of a given musical phrase. A *melodic accent* falls on events following a jump in pitch interval or events at a change in direction of the melodic contour. *Rhythmic grouping accent structure* causes an accent to be placed on the first and last events of a rhythmic group. Finally, the *foreground/background* relationship is exemplified by differences at the musical surface (e.g. tones that are louder, higher in pitch, or longer in duration).

Drake et al. found that accuracy of melodic performance in a reproduction task for both children (singing) and university music students (playing the piano) deteriorated when rhythmic grouping and melodic accent structures were out of synchronization. No

deterioration occurred, however, when only the metric structure was displaced. They also found that fewer errors were made by the college-age pianists on the accented notes than on unaccented ones, concluding that "accents may help listeners structure music in time by drawing their attention to important points in time" (p. 332).

The results of the studies cited above and the specific sources of musical accent delineated may be considered in terms of the four branches of Seashore's (1919) Tree of Musicality used in his well-known test of musical aptitude. The Melodic, Dynamic, Qualitative, and Temporal branches correspond, respectively, to the perceptual attributes of pitch, loudness, timbre, and duration. Using these four characteristics as a starting point, the potential sources of musical accent listed in Table 2.1 can be identified. *Melodic accent* may occur due to a tone's position within the tonal system (Cooper & Meyer's "melodic accent," Thomassen's "melodic accentuation," and Lerdahl & Jackendoff's "phenomenal accent"), pitch height (Thomassen's "melodic accentuation" and Lerdahl & Jackendoff's GPR3a "Register"), serial position within the melodic phrase (Ortmann, 1926), articulation (Lerdahl & Jackendoff's GPR3c "Articulation"), and melodic contour (Dowling & Fujitani, 1971; Dowling, 1978). The latter can occur due to either interval size differentiation (Thomassen's "interval size," Lerdahl & Jackendoff's GPR3a "Register," or Drake et al.'s "melodic accent structure") or direction change (Thomassen's "contour direction change" and Drake et al.'s "melodic accent structure"). *Dynamic accent* occurs when a given tone is louder than those in its immediate vicinity (Cooper & Meyer's "dynamic accent," Thomassen's "dynamic accentuation," Lerdahl & Jackendoff's GPR3b "Dynamics," and Drake et al.'s loudness "foreground/background"). Timbre variation may result in a *qualitative accent* (Thomassen's "harmonic accentuation"). *Temporal accents* occur with increased duration (Cooper & Meyer's "agogic accent," Thomassen's "temporal differentiation," Lerdahl & Jackendoff's GPR2b "Attack-Point"

and GPR3d "Length," and Drake et al.'s durational "foreground/background"), rhythm (Drake et al.'s "rhythmic grouping accent structure"), meter (Lerdahl & Jackendoff's "metrical accent" and Drake et al.'s "metric accent structure"), and delayed onset of a tone (Cooper & Meyer's "agogic accent" and Lerdahl & Jackendoff's GPR2b "Attack-Point").

Table 2.1. Potential sources of musical accent.

Melodic	Position in tonal system Pitch height Serial position Articulation Contour interval size direction
Dynamic	Loudness
Qualitative	Timbre
Temporal	Duration Rhythm Meter Onset delay

Sources of Visual Accent. David Marr's (1982) text concerning the computational tasks of visual processes began with the question "What does it mean, to see?" He answered this question, initially, by identifying vision as a process of discriminating from retinal images *what* is present in the visual field and *where* each object is located. Mortimer Mishkin and his associates (Mishkin, 1972; Mishkin, Lewis, & Ungerleider, 1982; Ungerleider & Mishkin, 1982; Mishkin & Ungerleider, 1983) proposed that these two tasks are performed by distinct anatomical pathways. Information about *form* is discerned by the parvocellular interblob system, while *color* is conveyed by the parvocellular-blob system, both of which terminate in the inferior temporal cortex. An object's *location in space* is largely the task of the magnocellular system, terminating in the posterior parietal cortex, which is important in the process of spatial organization

terior parietal cortex, which is important in the process of spatial organization (Kandel, 1991, p. 464).

James Thomas (1986) identified the retinal image as the immediate stimulus for visual perception, consisting of four variables: wavelength, time, and two spatial dimensions (p. II-3). These components may be delineated, respectively, as spectral, temporal, and spatial characteristics within the visual field. The first and last were accounted for in Mishkin's model, while the second introduced the important aspect of motion perception.

A *plenoptic function* (from *plenus*, complete or full, and *optic*) of early vision was proposed by Adelson & Bergen (1991). They suggested that "all the basic visual measurements can be considered to characterize local change along one or more dimensions of a single function that describes the structure of the information in the light impinging on an observer" (p. 4). The plenoptic function is a theoretical concept in which an idealized eye is conceived of as existing at every *x*-axis, *y*-axis, and *z*-axis location and the light rays passing through the center of the pupil at every possible angle, for every wavelength, at every point in time are recorded. Needless to say, in reality, this would be no small feat! The mathematical formula for calculation of the plenoptic function is provided below:

$$P = P(x, y, \lambda, t, V_x, V_y, V_z)$$

where:

P = intensity distribution

x and y represent the Cartesian coordinates of a picture plane

λ = wavelength

V_x = eye position (left to right)

V_y = eye position (up and down)

V_z = eye position (front to back)

Information provided by each of these dimensions was exemplified by Adelson & Bergen (1991) through a series of illustrations. A black and white photograph taken by a pinhole camera provides information concerning only the intensity of light from a single view-

point, at a single time, averaged over the wavelengths of the visible spectrum and thus may be parameterized using the Cartesian coordinates of a picture plane.⁵ A color photograph adds information regarding variation of intensity with wavelength. Addition of the temporal dimension can be exemplified by a color movie, while the various perspectives of a viewer could only be accounted for by a color holographic (i.e., 3D) movie. According to Adelson & Bergen, "a true holographic movie would allow the reconstruction of every possible view, at every moment in time, from every position, at every wavelength, within the bounds of the space-time-wavelength region under consideration" (1991, p. 5). They claimed that the plenoptic function is equivalent to this last type of visual world representation. In essence, the plenoptic function adds *viewer perspective* to the characteristics of the visual field suggested by Thomas (1986) and Mishkin et al. (Mishkin, 1972; Mishkin, Lewis, & Ungerleider, 1982; Ungerleider & Mishkin, 1982; Mishkin & Ungerleider, 1983).

A review of relevant literature suggested numerous potential variables for use in considering visual perception, as listed in Table 2.2. *Object shape* (or *form*) and *size* result from higher-order convergence of perceived *boundaries* and *contours* (Attneave & Arnoult, 1956; Zusne, 1970; Sutherland, 1973; Marr, 1982; Jackendoff, 1987; Hubel & Wiesel, 1962). Identification of an object's *location* and *orientation* provide additional information about recognized objects within the visio-spatial dimension (Hubel & Wiesel, 1962; Sutherland, 1973; Marr, 1982; Jackendoff, 1987). The perception of *color* results from light reflections on the retina, that may be specified in terms of *wavelength* and *intensity*. *Lighting* characteristics of the environment must, therefore, be taken into account (Evans, 1974). The resulting perception may be discussed in terms of three basic color sensations: *hue*, *saturation*, and *brightness* (Ostwald, 1931/33; Munsell, 1942; Sutherland, 1973; Evans, 1974; Marr, 1982). Combinations of these visual characteris-

tics may result in the perception of *patterns* (Martin, 1972; Sutherland, 1973) and/or varying *textures* (Marr, 1982).

Table 2.2. Potential sources of visual accent.

Spatial	Shape/Form Size Orientation Location
Spectral	Color Lighting Pattern Texture
Temporal	Motion horizontal vertical rotations in depth rotations in the plane translations in depth translations in the plane

Temporal aspects of the visual system assist in identifying changes in both the spatial and spectral characteristics as a function of time, resulting in the perception of *motion* (Marr, 1982; Jackendoff, 1987). In addition to simple horizontal or vertical motion, which have been described as "common motion" by Börjesson & von Hofsten (1972 & 1973), Kilpatrick (1952) delineated four types of "rigid motions." *Rotations in depth* may be exemplified by an object rotating about an axes perpendicular to the line of sight (e.g. a spinning globe). Objects rotating about the line of sight (e.g. a cartwheel motion) illustrate *rotations in the plane*. *Translations in depth* are simply motions toward or away from the observer, while *translations in the plane* consist of objects moving across the frontal plane (e.g. from left to right).⁶

Low-order features (e.g. angle size, object size, orientation, and location within the visual field) appear to be abstracted by the visual system in several stages (Hubel & Wiesel, 1962). Marr (1982) considered this process in terms of consecutive levels which he described as a primal sketch, a 2½D sketch, and a 3D model. The *primal sketch* was based on the principle that the detection of discontinuities (i.e. "change" as discussed above as a source of accent) of intensity in the retinal image is necessary for the perception of form. *Local markers* or *boundaries*, at this low level, include *segments* of edges (identified simply by position, length, and orientation), *termination* of edges, and *discontinuity* of orientation (i.e. corners). In the *2½D sketch* the surface representations visible to the observer take on the additional features of object contour and depth. However, only with the *3D sketch* does the representation become *volumetric*, i.e. the objects occupy volume in space. Also, this final level is considered to be "object-centered," rather than "viewer-centered," meaning that, regardless of viewer position, object shapes and size constancies may be stored in long-term memory for recognition on subsequent occasions from any perspective (see Jackendoff, 1987, 168-78 for a complete discussion).

The present investigation sought to learn about visual accent structure at this highest level of abstraction and was, therefore, concerned with those aspects of the visual field resulting from convergence of the low-order receptive fields.

A Common Language. Livingstone & Hubel (1987) proposed that components of a visual scene are organized by a perceiver into coherent groups based on each object's particular set of values (e.g. brightness, texture, depth, etc.). When objects move as a function of time, these values may be considered to have a *direction* and *velocity* that may be used as an additional source of information in separating a given object from surrounding objects.

Herbert Zetl (1990) suggested that such *directional forces* are "probably the strongest forces operating within the screen ... [and] lead our eyes from one point to another within, or even outside, the picture field" (pp. 119-20). He referred to these forces as *vectors* with a specific *direction* and *magnitude*.

It is possible that, by using the concept of vectors with these two primary characteristics, a common terminology may be used to discuss both the musical and visual parameters utilized in the present study *as a function of time*. For example, pitch height of a melody can be described as ascending (direction = up), descending (direction = down), scalar (magnitude = small), or arpeggiated (magnitude = moderate). Loudness characteristics of a given musical passage could likewise be described as crescendo (direction = up), decrescendo (direction = down), and amount of change per unit time (i.e. magnitude). In the visual domain, motion across the spatial dimension can be represented easily as direction (e.g. up/down and left/right) and magnitude (i.e. amount of movement).

Once a motion vector (either auditory or visual) is perceived, expectations are generated for continuation. When this expectancy is blocked or inhibited (i.e. a change of direction or magnitude occurs), a point of emphasis is created (Meyer, 1956; Fraise, 1982; Deliege, 1987). When points of accent occur simultaneously (e.g. a tone that is louder, a different timbre, and at a point of melodic contour direction change with an object that changes color, shape, and direction of motion at the same instant), the resulting perceived accent will take on added significance (Monahan, Kendall, & Carterette, 1987; Drake, Dowling, & Palmer, 1991).

Potential Sources of Musical and Visual Accent in the Present Study. A limited number of the potential variables listed in Table 2.1 and Table 2.2 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 of all, 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). Secondly, change in the direction of a motion vector will cause a perceived 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 2.3, along with proposed values for the direction and magnitude characteristics. The value assigned to each potential variable listed in Table 2.1 and Table 2.2 that does not appear in Table 2.3 will be kept constant as a means of controlling its influence on subject responses.

Table 2.3. Proposed variables to be utilized in the initial pilot 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 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	purier/unchanging/more impure	"
brightness	brighter/unchanging/darker	"

Mapping of Audio-Visual Accent Structures. Using variables common to both the visual and auditory systems, the present investigation relied on subjects to perform a cross-modality transfer task (i.e. matching of audio and visual stimuli). This procedure was an offspring of S.S. Stevens' (1956) work using magnitude estimation. Rather than asking subjects to assign numbers to a varying stimulus, however, Stevens (1959) asked them to match the loudness of a sound to the magnitude of vibration produced by a device attached to the finger. His subjects participated in two procedures. First, they were required to vary the level of vibration until they felt that it was equal to that of the loudness of the tone. Then a second set of measurements was taken while the subject was provided with a vibration level to which they were to match the loudness. Rule (1969) confirmed the ability to match across modality by having subjects assign numbers to circles of different sizes. A plot of the size of the circle in inches by the subjects' magnitude estimations resulted in an almost perfectly straight line, confirming that changes in the visual domain were accurately reflected in the subjects' ratings.

In the field of music cognition, W. Jay Dowling has done a great deal of work investigating the abstraction of melodic contour on the initial hearing of novel pieces of music, claiming that specific intervals or tone chromas are usually stored only after repeated listening (Dowling & Fujitani, 1971; Dowling, 1978).

Davies & Jennings (1977) showed that both musicians and nonmusicians were able to make fairly accurate drawings as a means of representing melodic contour. Their study showed not only that subjects were able to abstract the criterial attributes of a musical phrase, but also that they were capable of mapping that information across modalities in the process of representing that aural information visually, in the form of a drawing. Such cross-modal analogies form the basis of the ability to match musical and visual accent structures.

CHAPTER THREE

METHOD

Research Design

This study was a quasi-experimental investigation of the aesthetics involved in the perception of an audio-visual composite, utilizing both highly abstract musical and visual materials, as well as actual motion picture excerpts. The method of every experiment described in the following chapters consisted of a post-test only, repeated measures factorial design. Each experiment, presented in a separate chapter for organizational clarity, was preceded by a series of exploratory studies that assisted in selecting stimulus materials and creating alignment conditions for the various pairs of audio-visual (AV) stimuli. The main experiments incorporated two independent methods (subject preference ratings and similarity judgments) of gathering experimental data. The null hypothesis for each of these procedures maintains that there will be no change in subject responses (i.e. verbal ratings or similarity judgments) as a result of the synchronization of musical and visual accent structures, as represented below.

$$H_0: \mu V_1 A_1 S_1 = \mu V_1 A_1 S_2 = \mu V_1 A_1 S_3 = \dots = \mu V_x A_y S_z$$

where:

V = visual component

A = audio component

S = A-V synchronization

Subject Selection

All subjects (except where otherwise noted) were students from the University of California, Los Angeles. Every participant was required to have seen at least four main-stream, 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. Across all experiments, musical training was the single between-subjects grouping variable that was considered. Subjects were grouped according to 3 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). Musical training was selected as a grouping variable, because in the perception of movies or animations it is hypothesized that musically-trained individuals exhibited a higher level of awareness for audio-visual synchronization than those with less musical training.

Stimulus Materials

In addition to the two methods utilized in gathering data, there were three proposed stages of the present series of investigations, each incorporating a different stimulus set, increasing gradually in complexity. The experiments progressed from a highly atomistic, abstract initial study to the final experiments which utilized excerpts from an actual motion picture. Each of these stages will be discussed in the following paragraphs. Since the method remained consistent for every experiment, the focus of the following descriptions is on the disparate sets of stimuli. After describing these varying stimulus materials, the procedural method is clearly explicated.

Experiment One. Before the initial experiments, a series of exploratory studies was run to determine auditory and visual stimuli that are consistently interpreted by subjects as generating the intended accent points. For this procedure, the sources of

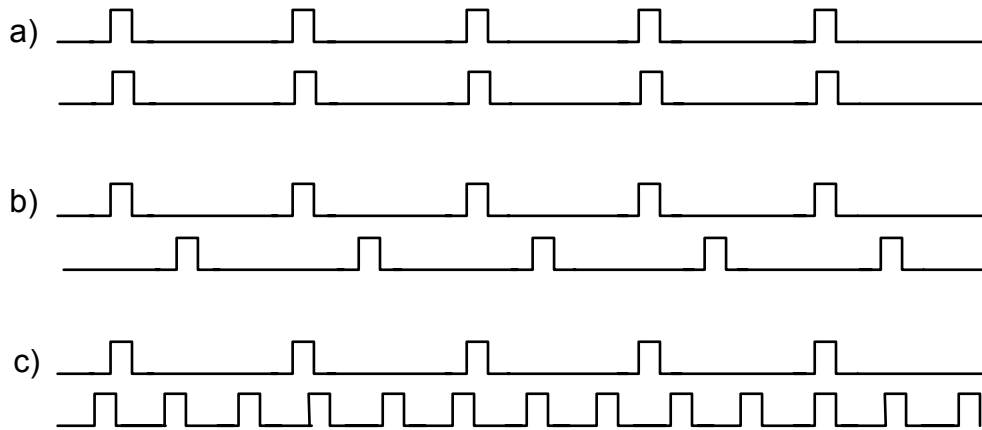
musical accent delineated in the previous chapter was used as a theoretical basis for creating MIDI files to be used as musical stimuli. Likewise, the visual sources of accent discussed previously were utilized as a basis for generating computer animations for use in this experiment. Both the sound files 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 frames per second rate of 20fps⁷, 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 another 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, in press). Figure 3.1 presents a 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 3.1a) may be exemplified by accent structures that are perfectly synchronized. Accent structures that are *out-of-phase* (Figure 3.1b) 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 3.1c).⁸ 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.

Experiment Two. The second stimulus set consisted of 8-second excerpts from Norman McLaren's "Dots" (1940), "Canon" (1964) and "Synchromy" (1971). This group of stimuli maintained the abstract geometric shapes of the previous experiment, rather than moving toward representational film, but complexity within the audio and visual fields will rise due to increasing musical validity and the presence of multiple objects (with individual motion vectors) simultaneously.⁹

Figure 3.1. Visual representations of relationships between sources of accent.



These experimental animation projects were chosen because of the present author's belief that both the musical and visual accent structures are perceived easily. In addition, it is well documented that McLaren's creative process involved a great deal of concern for alignment of the audio and visual aspects of his animations (Collins, 1990). In fact, for the three animations utilized in this study McLaren used an innovative technique to draw the soundtrack directly onto the film, assuring alignment of musical and visual events.

Specific excerpts were chosen with two criteria in mind: a clear sense of perceived rhythmicity (i.e. points of accent due to changing characteristics, as described previously) and varying tempos (i.e. the periodic rate at which these accents occur, cued by the metronomic pulse of the music track). In generating audio-visual composites for the experimental procedure, the same levels of alignment were utilized as in Experiment One (consonant, out-of-phase, and dissonant). In Experiment Two and Three, however, the terms took on a slightly different meaning. A *consonant* relationship consisted of the composer's intended audio-visual combination, resulting in the synchronization of accent

structures. An *out-of-phase* relationship resulted when the composer's intended audio-visual pair was offset by a perceptibly salient amount—to be determined in the exploratory studies. Composites utilizing a soundtrack from one work and the visual images of another were considered *dissonant*, due to the variance between their accent structures.

Experiment Three. In the final stage of the present investigation, three actual motion picture excerpts were selected, adding the referential dimension (e.g., human characters, location, dramatic action, etc.). Selection criteria was the same as that utilized in Experiment 2. The stimuli consisted of 25-second excerpts from the movie “Obsession” (1975) with a musical score composed by Bernard Herrmann.

Exploratory Studies

A series of exploratory studies were 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 of this dissertation, Roger Kendall's *Music Experiment Development System* (MEDS, version 3.1e) was utilized to play the auditory and visual examples and collect subject responses.¹⁰ 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.¹¹ This facility allowed the subjects to register their perceived pulse simply by tapping along on the spacebar.

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 for Experiment One, stimuli were continuously looped for a period of about 30-seconds so that subjects had an adequate period of time to determine accent

periodicities. In Experiments Two and Three, subjects were allowed to practice tapping as many times as they wish before recording their responses.

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 determine the audio and visual stimuli that produce the most reliably consistent sense of accent structure.

Main Experiments

The methods utilized in Experiments One, Two, and Three remained consistent, differing only in the stimulus materials, as described above. Therefore, a single delineation of the procedure is provided below with parenthetical remarks added when necessary to differentiate between the three main experiments.

There are two methodological innovations that warrant a brief discussion. First, a system of “convergent methods” was utilized to answer the research questions. Kendall & Carterette (1992a) proposed this alternative to the single methodology 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, 1993 & 1992b).¹² In contrast to semantic differential

scaling, 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 groups of subjects were required for each experiment. Group One was asked to watch every audio-visual composite in a randomly-generated presentation order and provide a VAME response, according to the following instructions:

In the following experiment, you will see and hear a series of [animations / movie excerpts] combined with a variety of musical accompaniments. After each combination, you will be asked to respond on two scales: 1) the degree of synchronization between the auditory and visual components and 2) the effectiveness of the pairing.

The rating of ‘synchronization’ refers to how often important events in the music coincide with important events in the visual scene. The rating of ‘effectiveness’ simply concerns your subjective evaluation of how well the two go together. You will provide a rating by moving the button on a scroll bar. For example, if you consider an audio-visual combination to be in PERFECT synchronization, you would click the button with the mouse and drag it to the end of the scroll bar labeled ‘synchronized.’ Sometimes you will be asked to provide a rating of the synchronization first, other times you will be asked initially for a rating of effectiveness, so pay attention to the labels below the scroll bar. After providing both ratings, you will be given the opportunity to either see the same combination again or proceed on to the next example.

The following three examples are practice combinations, so that you can get used to the procedure. If you have any questions, feel free to ask them after completing these warm-up examples.

Before you begin, please put the headphones on ...

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. An analysis of variance (ANOVA) was used as the method for determining whether or not there was a significant difference between the responses as a function of accent alignment.

In a paired-comparison task, Group Two was asked to provide ratings of “similarity” (i.e. on a continuum from “not same” to “same”), according to the instructions below:

In the following experiment, you will see and hear a series of [animations / movie excerpts] accompanied by musical sound, presented in pairs. After each pair, you will be asked to respond on a single scale: not same–same. You will provide a rating by moving the button on a scroll bar. For example, if you consider a pair of audio-visual combinations to be IDENTICAL, you would click the button with the mouse and drag it to the end of the scroll bar labeled ‘same.’ Therefore, throughout this experiment, ‘same’ will be defined as ‘identical’ and ‘not same’ will be defined as ‘maximally different.’ In making this assessment, you are judging the audio-visual combination, not either the sound or image in isolation.

The following six examples are practice pairs, so that you can get used to the procedure. If you have any questions, feel free to ask them after completing these warm-up examples.

Before you begin, please put the headphones on ...

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 multi-dimensional space (Kruskal, 1964a & 1964b). According to Wilkinson, Hill, Welna, and Birkenbeuel (1992), the function of MDS is “to compute

coordinates for a set of points in a space such that the distances between pairs of these points fit as closely as possible to measured dissimilarities [or similarities] between a corresponding set of objects” (p. 109). 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 and—in Experiments Two and Three—the audio track was the one intended to accompany the given visual scene). 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 Zetl (1990, p. 380).

Finally, it was hypothesized that similarity ratings provided by Group Two would result in a multi-dimensional space consisting of three dimensions: musical stimulus, visual stimulus, and accent alignment. The third dimension will exert less influence over the solution as the stimuli progress from simple abstract images (Experiment One) to real movie excerpts (Experiment Three), because of the relative complexity within both the musical and visual vector fields, increasing the potential sources of perceived accent at any point in time.

Brief Summation

In the past, a great deal of attention has been paid to the appropriateness of a musical soundtrack for a cinematic sequence based on referential associations (i.e. the suitability of a given musical style for a scene with a specific emotional-dramatic intent). However, one purpose of the present study is to suggest that there are aural and visual forces working at a syntactical level that are also worthy of consideration during both compositional (i.e. creative) and analytical (i.e. critical) processes. The method described herein provided a means of quantifying perceptual responses to audio-visual composites that differ in temporal interval between accent points and intermodal alignment of these strata.

CHAPTER FOUR

EXPERIMENT ONE

Since there has been little research concerning the synchronization of auditory and visual stimuli in the context of film and animation, the first experiment in the present series of investigations utilized stimuli that can be classified as simplistic. Auditory examples 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 pilot study portion to ensure reliability in responses to the selected stimuli. In addition, since this is the first explication of a specific application of the experimental method described in the previous chapter, more details are presented in order to provide procedural clarification.

Exploratory Study

An exploratory study was designed to confirm that the events considered “accented” (i.e., points of greater salience) according to the literature cited earlier were, in fact, perceived as such by a group of subjects. The purpose of these two initial experiments was to determine which of the auditory and visual stimuli most reliably resulted in the perception of a periodic pulse at the hypothesized rate. The rate of this pulse is referred to as *interonset interval* (IOI).

The null hypotheses predict that there will be no change in subject response to auditory (A) or visual (V) stimuli utilizing three different accent periodicities (i.e., IOI).

$$H_0: \mu A_{500} = \mu A_{800} = \mu A_{1000}$$

$$H_0: \mu V_{500} = \mu V_{800} = \mu V_{1000}$$

Subjects for the exploratory study, were seven individuals residing in the Dallas area.

Stimulus Materials

Auditory. Seven musical pitch sequences were created using Coda Software's Finale[®] musical notation software and are presented in Figure 4.1. Each example used either a different type (or combination of different types) of melodic sources of accent. The melodic accent in Pattern #1 is caused by pitch contour direction change. Pattern #2 illustrates accent periodicities related to a combination of both direction change and a change in interval size. A dynamic accent was added to the pitch contour direction change in Pattern #3. Pattern #4 exhibits a duration change and a pitch contour direction change. Changing timbre (i.e., a synthesized brass sound, as opposed to the piano sound used for all other patterns) is added to the pitch contour direction change in Pattern #5. Patterns #6 and #7 are essentially duplication of the sources of accent used in Patterns #1 and #5, utilizing a triplet rhythmic pattern rather than the typical duple meter of the other audio stimulus patterns. These pitch sequences were saved as MIDI files for use in the investigation.

Visual. Seven computer animation sequences were created using Autodesk 3-D Studio[™]. The resulting apparent motion patterns (i.e., a neon green circle moving across a black background) are represented visually in Figure 4.2.¹³ Visual Pattern #1 shows the circle moving counter-clockwise from left to bottom to right to top. Visual Pattern #2 utilizes the same accent points, but the motion is in a clockwise direction. Visual Pattern #3 shows the ball moving vertically from top to bottom, while visual Pattern #4 displays

the ball moving horizontally from left to right. Pattern #5 produces the illusion of a z-axis in its emulation of front to back motion. Pattern #6 uses a morphing technique causing the ball (larger in this pattern and Pattern #7) to stretch left into an oval shape, return to center in the shape of a circle, stretch right into an oval, and return to the center circle shape. Pattern #7 reveals the larger ball centered on the screen changing colors.


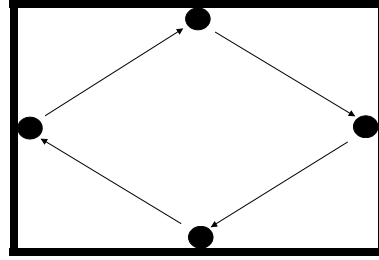
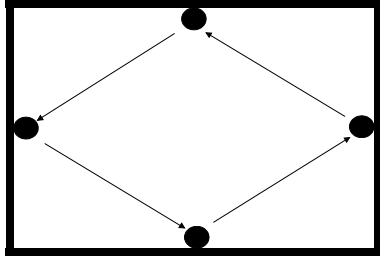
Figure 4.1. Notation of the musical stimuli used in the exploratory study. 

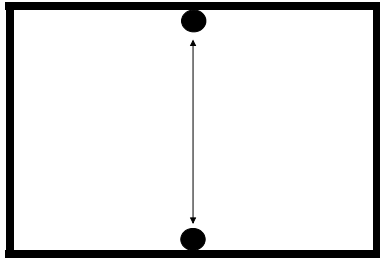
Figure 4.2. Visual stimulus patterns used in the exploratory study.

Pattern #1—counter clockwise

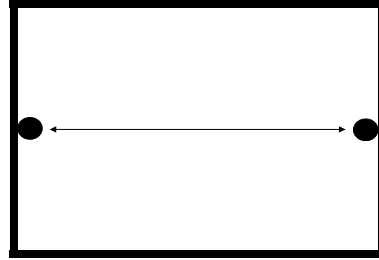
Pattern #2—clockwise



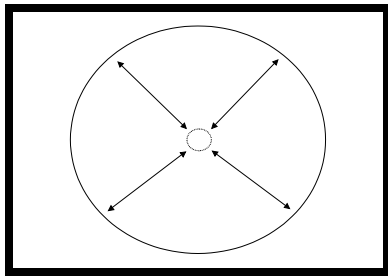
Pattern #3—vertical



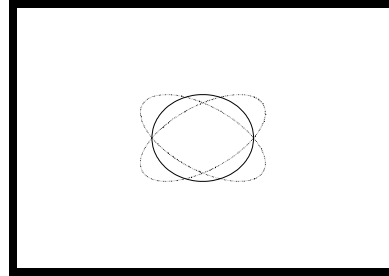
Pattern #4—horizontal



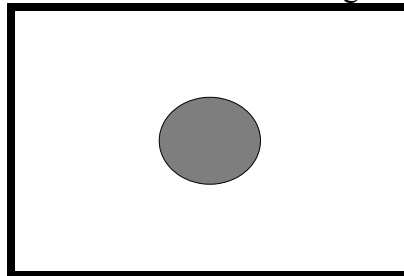
Pattern #5—back to front



Pattern #6—shape change



Pattern #7—color change



Each audio and visual example consisted of periodically spaced accent points, exemplifying either an IOI of 500ms, 800ms, or 1000ms. These values are well within Fraisse's (1982) "temporal limits for the perceptual grouping of sound events" (i.e., lower limit = 120ms; upper limit = between 1500 and 2000ms), as discussed previously.

Equipment

The stimuli for the exploratory study were presented on an IBM-compatible 486-33MHz computer with a CTX Super VGA monitor and a ProAudio Spectrum 16-bit internal sound card. Subjects heard the MIDI-generated pitch sequences through a pair of Sony DMR-V100 headphones.

Subject Procedure

Every subject was asked to perform two tasks; one involving the auditory stimuli discussed above and the other utilizing the visual stimuli. Subjects were divided randomly into 2 groups. One group performed the auditory task first and the other performed the visual task first, according to the following instructions¹⁴:

Thank you for volunteering to be a subject for this experiment. In the next few minutes, you will [hear / see] a series of ['melodies' / animations].

Your task is to simply tap the SPACEBAR along with the pulse created by [each melody / the visual images]. When the [music / animation] begins, [listen to / watch] it carefully and, when you feel that a pulse has been established, tap along with it....

Each of the two procedures consisted of 21 examples (7 musical or 7 visual stimulus patterns [Figures 4.1 & 4.2] x 3 IOI conditions [500ms, 800ms, and 1000ms]). The auditory and visual patterns were looped (i.e., repeated continuously) for a temporal duration of approximately 30 seconds each, allowing ample time for the determination of

perceived pulse and a sufficient period of tapping for each subject. The stimuli were presented to each subject in random order.

Data Analysis

Temporal intervals between presses of the spacebar (IOIs) were saved to the computer hard disk for the purpose of analysis. This data stream consisted of a series of IOIs for every subject in response to each of the 21 auditory and each of the 21 visual stimuli. The initial IOIs prior to establishment of a steady pulse and the final IOI for every stimulus were eliminated from the data files. The initial taps were frequently unequal, possibly due to onset of motor activity. The final tap was sometimes cut short by data storage process.

The number of IOIs in response to each stimulus was, therefore, a function of two variables: 1) the amount of time from the onset of stimulus presentation to the first spacebar press at a steady IOI and 2) the IOI between key presses (e.g., it is possible to press the key more times within a 30 second time period when tapping at a rate of 500ms than when tapping at a rate of 1000ms). The total number of clicks varied from subject to subject, so the mean click rate was used as the best estimate of perceived IOI.

Examining the original data, it was clear that there were instances in which extraneous key presses occurred, e.g., where a steady IOI stream was interrupted by a single value approximately twice the duration of the established pulse. In this case, it is believed that the subject simply missed the spacebar or did not depress it completely.

Also, there were infrequent occurrences of two to four IOI values differing from those of the surrounding steady stream of inter-keypress durations. In such cases, because these values were surrounded on either side by a series of consistent IOIs, it is believed that the subject either experienced a brief attentional lapse or had momentary motor difficulty with the task of depressing the spacebar. These deviant IOIs were con-

sidered to be “noise” and, as such, were eliminated from the data set prior to calculating subject mean scores. Such extreme scores (e.g., a value of 1000ms in a stream of 500ms IOIs) would have had a dramatic effect on the mean IOI.

Results

Before discussing the results of the exploratory investigation, it is important to distinguish between two terms that will be used throughout. A *stimulus pattern* refers to any of the seven audio or visual sequences, as represented in Figures 4.1 & 4.2. In contrast, the term *stimulus number* refers to any of the stimulus patterns at a specific IOI. This relationship is represented in Table 4.1 for clarification.

Table 4.1. Relationship of *stimulus numbers* to *stimulus patterns* and IOIs.

Stimulus Number	Stimulus Pattern & IOI (in ms)
1-7	1-7 @ 500ms
8-14	1-7 @ 800ms
15-21	1-7 @ 1000ms

The subject response means are represented graphically in Figures 4.3a & 4.3b. A general finding in the subject data—particularly evident in the responses to auditory stimuli—was a halving or doubling of the hypothesized IOI. For example, if the hypothesized IOI was 800ms, a subject might tap every 1600ms (doubling or “nesting”) or every 400ms (halving or “subdividing”).

This effect may be considered analogous to the “octave effect” found in pitch matching experiments and confirms Lerdahl & Jackendoff’s (1983) “metric structure” subdivisions. There appears to be a preference for duple subdivision. When asked to tap along with complex rhythmic patterns, subjects tend to migrate toward a 2:1 ratio. Povel (1981) has suggested that there are two steps involved in the process of encoding rhythms. First, the listener attempts to identify a beat framework (i.e., a regular beat

pattern) with interbeat intervals of not much more than 1.5 seconds. Then, this pattern of beats is divided further into an equal number of subdivisions. Once again, the preference is for a 2:1 ratio. Fraisse (1982) points out that this preference is well-justified since there is a prevalence of such subdivisions (80-90%) in Western music composed in the period from Beethoven to Bartók. Metrical organization (i.e., the search for underlying regularity) with a preference for nested hierarchical relationships constitutes a cognitive framework for the temporal dimension of music perception, similar to the framework that the scale provides for the pitch dimension.

To consider such nesting and subdivision significantly different from the hypothesized IOI proposed in the exploratory study would be a misrepresentation, since these subject responses were clearly related to the hypothesized IOI. Therefore, if the IOI could be accounted for in terms of either nesting or subdividing the hypothesized IOI, the scores were “equalized” (i.e., either multiplied or divided to bring them in line with the hypothesized rate), as represented in Figures 4.4a & 4.4b.

In all but four cases, the majority of subjects clicked at the hypothesized IOI. Three of the four deviations from majority agreement with the hypothesized rate involved visual pattern #5 (i.e., visual stimulus numbers 5, 12, & 19). In each of these cases, all subjects clicked at an IOI double that of the hypothesized rate. Therefore, this pattern was considered to be extremely reliable in creating a perceived pulse, but the pulse perceived was double that of the hypothesized IOI. Therefore, the accent structure perceived by the subjects was used as a basis for the alignment in the Experiment 1, rather than the hypothesized IOI. This is an important confirmation of the fact that, in this investigation, the perceptual frame of reference was considered criterial. For statistical analysis of the exploratory study data, however, these IOIs were mathematically transformed to the hypothesized rate.

Figure 4.3a. Mean subject responses to the auditory portion of the exploratory study.

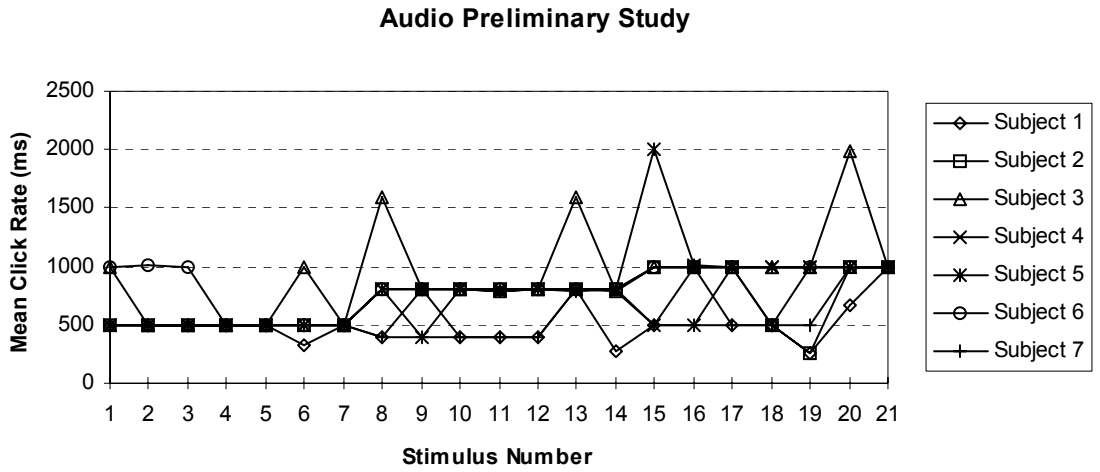


Figure 4.3b. Mean subject responses to the visual portion of the exploratory study.

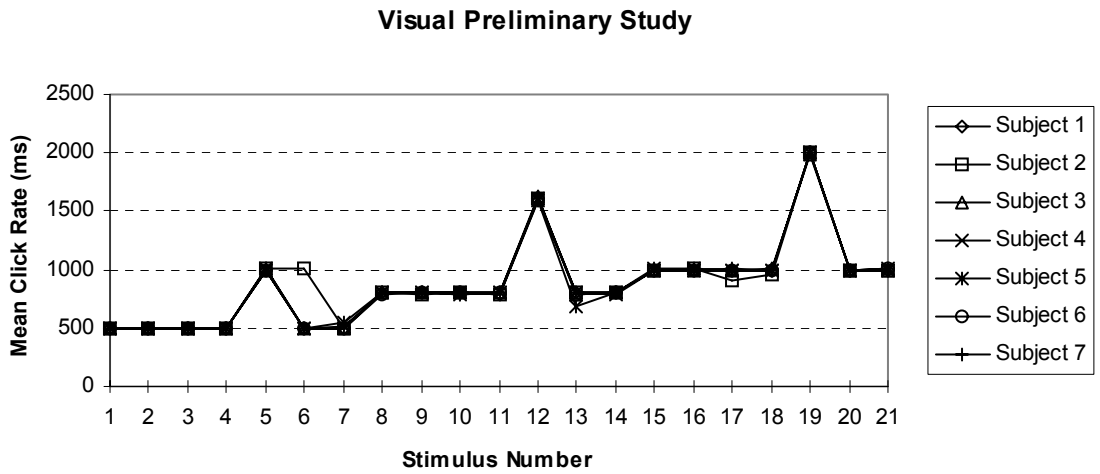


Figure 4.4a. Equalized mean subject responses to the auditory portion of the exploratory study.

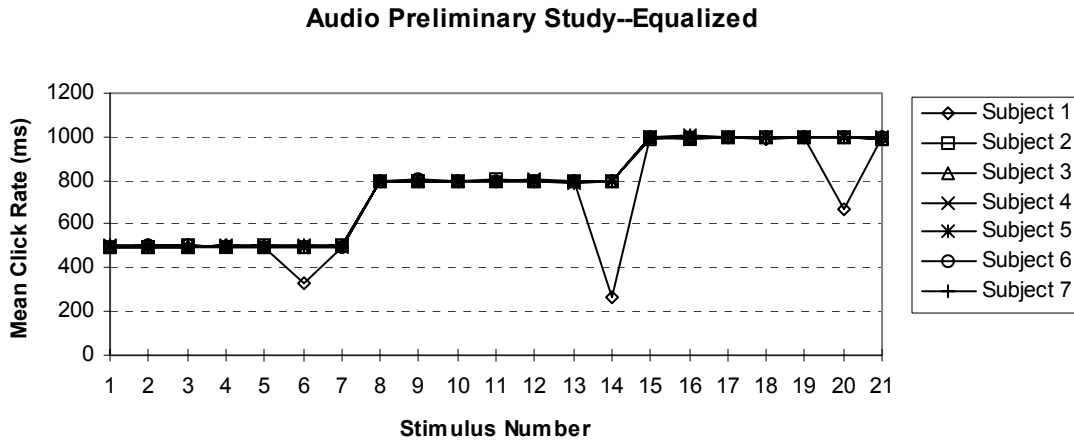
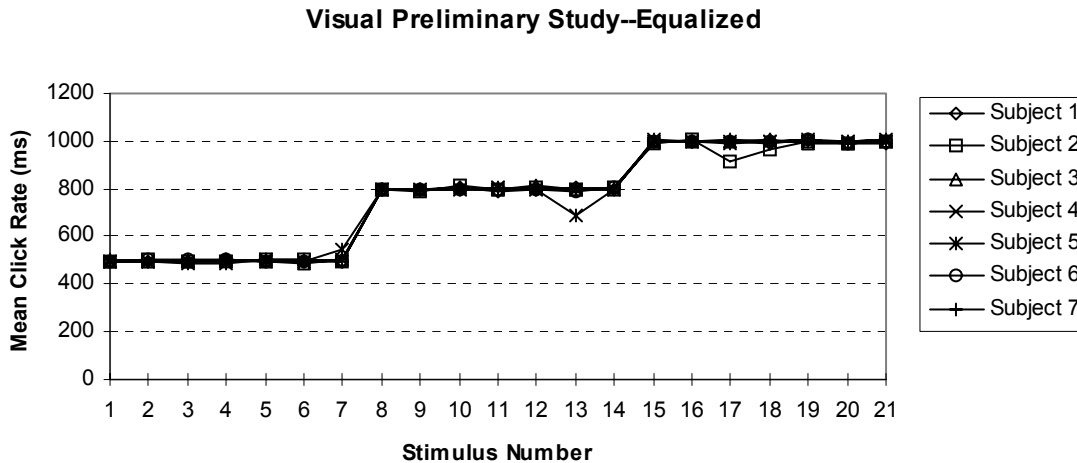


Figure 4.4b. Equalized mean subject responses to the visual portion of the exploratory study.



The other single instance in which the majority of subjects tapped at a rate different from the hypothesized IOI was audio stimulus number 18. Since audios #4 and #11—stimuli using this same pitch pattern—produced the hypothesized result, one can only speculate about the reasons for this anomaly. Perhaps the IOI of 1000ms seemed too long and the majority of subjects were more comfortable subdividing this interval

into two parts (Fraisse, 1982), though this was not the case with any of the other auditory stimulus patterns.

Statistical Analysis. These equalized data were subjected to a repeated measures analysis of variance (3 IOIs x 7 stimulus patterns). The null hypothesis for the auditory stimuli was rejected, since there was a significant difference between the subject responses to the three IOIs ($F_{(2, 12)} = 4211.067, p < .0005$). However, there was not a significant difference either between the seven stimulus patterns ($F_{(6, 36)} = 1.138, p < .36$) or in the interaction between IOI and stimulus pattern ($F_{(12, 72)} = 1.010, p < .449$). Similarly, responses to the visual stimuli exhibited a significant difference between the IOIs ($F_{(2, 12)} = 12100, p < .0005$), but not between the stimulus patterns ($F_{(6, 36)} = 1.744, p < .139$) or in the interaction between temporal interval and stimulus pattern ($F_{(12, 72)} = .959, p < .495$). Post hoc comparisons, using Tukey's HSD ($\alpha = .05$), revealed that both the audio and visual stimuli separated clearly on the basis of hypothesized IOI, i.e., all stimuli hypothesized to have 500ms accent intervals belonged to the same population and were significantly distinct from all stimuli hypothesized to have either 800ms or 1000ms accent intervals. The single exception to this rule was Audio Stimulus #14, due to the extremely low mean IOI of one subject that cannot be accounted for in terms of either nesting or subdivision. Tables 4.2a & 4.2b present the number of individual subjects who responded in a manner that was either different from the hypothesized rate (i.e., not nested or subdivided), did not press the spacebar at a steady rate (e.g., a pattern of long-short taps rather than a steady IOI), or the subject response mean failed the Tukey pairwise comparison.

Table 4.2a. The number of subjects that responded to the auditory stimuli in a manner that was either different from the hypothesized rate (i.e. not nested or subdivided), the spacebar presses did not occur at a steady rate (e.g. a pattern of long-short taps rather than a steady IOI), or the stimulus failed the Tukey pairwise comparison.

	Pattern #1	Pattern #2	Pattern #3	Pattern #4	Pattern #5	Pattern #6	Pattern #7
500ms						1	
800ms						1	1
1000ms				1		1	

Table 4.2b. The number of subjects that responded to the visual stimuli in a manner that was either different from the hypothesized rate (i.e. not nested or subdivided), the spacebar presses did not occur at a steady rate (e.g. a pattern of long-short taps rather than a steady IOI), or the stimulus failed the Tukey pairwise comparison.

	Pattern #1	Pattern #2	Pattern #3	Pattern #4	Pattern #5	Pattern #6	Pattern #7
500ms							1
800ms	1					1	
1000ms							

Stimulus Selection for Experiment One.

A single selection criterion was used to determine the most reliable audio and visual patterns to be used in constructing the audio-visual (AV) composites for Experiment One. Only those patterns to which there were no responses deviating from the hypothesized rates (i.e., empty columns in Tables 4.2a and 4.2b) were considered for inclusion. In considering these responses, nesting and subdividing were not considered deviant. Audio patterns 1, 2, 3, & 5 and visual patterns 2, 3, 4, & 5 met these criteria.

It was desirable to further reduce the number of musical and visual patterns since there were 14 alignment possibilities for each combination of a musical pattern and a

visual pattern, as shown in Table 4.3. By limiting the number of musical and visual components to two, it was possible to create both a semantic differential task that could be completed by most subjects in 45 minutes or less. This limitation ensured that experimental time constraints would be met and also reduced the potential of subject fatigue becoming a factor in the design. Of the four patterns that met the initial criterion stated above, one audio pattern (#1) and one visual pattern (#3) were eliminated, because they caused the most subjects to tap at either a nested or subdivided IOI, rather than at the hypothesized rate, suggesting lower reliability. In piloting the AV composites, it became clear that the timbre change in audio pattern #5 caused most subjects to experience a streaming effect¹⁵ (Bregman, 1990; Bregman & Campbell, 1971; van Noorden, 1975) that resulted in confusion when attempting to make judgments concerning the alignment of auditory and visual points of accent. Finally, visual pattern #2 was eliminated, because it has four accent points within the repeated sequence (i.e., left side, top, right side, bottom) as opposed to the two accent points existing within the repeated sequence of the remaining visual and auditory patterns. Removing this pattern from the data set allowed the number of accent points within the repeated portion of all stimuli to be controlled. Therefore, of the seven auditory and seven visual patterns, audio patterns # 2 and #3 and visual patterns #4 and #5 were selected for use in Experiment One (see Figures 4.1 & 4.2).

Main Experiment

Subjects

Subjects for this experiment were 40 UCLA students (ages 19 to 31) taking general education (G.E.) classes in the Music Department—either Psychology of Music (Lipscomb; Fall 1994) or American Popular Music (Keeling; Summer I 1994).¹⁶ The 40

subjects were randomly assigned to two groups before performing the experimental tasks. Group One (n = 20) responded on a verbal rating scale and Group Two (n = 20) provided similarity judgments between pairs of stimuli. The number of subjects falling into each of these categories is represented in Table 4.4.

Table 4.3. 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

Table 4.4. Number of subjects falling into each cell of the between-subjects design (Experiment One).

<u>Exp. Task</u>	<u>Musical Training</u>		
	<i>Untrained</i>	<i>Moderate</i>	<i>Trained</i>
VAME	10	7	3
Similarity	10	8	2

Equipment

Experimental stimuli were presented on an IBM-compatible 486-50MHz computer with a ProAudio Spectrum 16 sound card and a 16-bit video accelerator card using

Cirrus Logic 6026 technology. The subjects heard the audio track through a pair of Sennheiser HD222 headphones.

Stimulus Materials

The AV composites utilized in Experiment One were created by combining the two audio and the two visual stimuli selected in the exploratory study into all possible pairs ($n_{AV} = 4$).¹⁷ For ease of discussion, the method of referring to these stimuli will now be changed. Audio 1 (A1; shown as audio stimulus pattern #2 in Figure 4.1) consists of a repeated ascending melodic contour; i.e., CDEFCDEF). Audio 2 (A2; shown as audio stimulus pattern #3 in Figure 4.1) consists of an undulating melodic contour; i.e., CDEFGFED. Visual 1 (V1; shown as visual stimulus pattern #4 in Figure 4.2) represents left to right apparent motion, i.e., movement along the x-axis. Visual 2 (V2; shown as visual stimulus pattern #5 in Figure 4.2) represents front to back apparent motion, i.e., movement along an apparent z-axis.

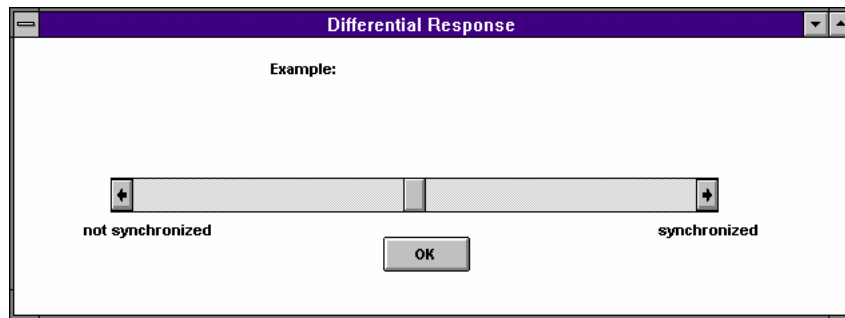
In addition to these various AV composites, the method of audio-visual alignment was systematically altered. Three levels of superimposition were utilized, as discussed in Chapter 3: consonant, out-of-phase, and dissonant. It was important to create composites in which the AV alignment was out-of-phase 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 “pulse” (i.e., accent) where certain moments in the stimulus stream were considered to be 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 the 14 different alignment conditions shown in Table 4.3 for each AV composite. 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 about 15 seconds, before requiring a subject response. Order of stimulus presentation was randomized for every subject.

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,” according to the instructions provided in the previous chapter. After viewing one of the randomly-generated composites, the subject was given a choice of either providing a response or repeating the stimulus. As mentioned in the instructions, the order in which the two VAME scales were presented was also randomized.

Figure 4.5. Scroll bar used to collect Group One subject responses.



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. The resulting similarity matrix was used to calculate coordinates within a multidimensional space, such that the distances between pairs of these points fit as closely as possible to the similarity judgments provided by the subjects in Group Two.

Results

Group One Data Analysis and Interpretation. The data set for Experiment One (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. A repeated measures ANOVA was performed on the subject responses to each of the VAME rating scales provided by Group One, 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 4.6a to 4.6d. Each graph represents a different AV combination. The x-axis labels refer to specific AV composites exemplifying a specific alignment condition (i.e., synchronization), determined by the following algorithm:

$$V[\text{visual IOI} \div 100]A[\text{audio IOI} \div 100]_{\text{Alignment}}[\text{AV Composite}]$$

where:

visual IOI \div 100 = 5 for 500ms; 8 for 800ms; 10 for 1000ms

audio IOI \div 100 = 5 for 500ms; 8 for 800ms; 10 for 1000ms

Alignment = C for Consonant; O for Out-of-Phase; D for Dissonant

AV Composite = 1 for Visual 1 & Audio 1; 2 for Visual 1 & Audio 2; 3 for Visual 2 & Audio 1; 4 for Visual 2 & Audio 2

Examining Figures 4.6a to 4.6d, there is a striking consistency in response pattern across AV composites. This consistency is confirmed by Figure 4.7, providing a comparison of these same responses across all four AV combinations by superimposing Figs 4.6a to 4.6d 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). 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 4.7) are consistently given higher ratings than the nested consonant composites (e.g., alignment conditions V5A10_C and V10A5_C in Figure 4.7),¹⁸ with the exception of V10A5_C4 which received a mean rating almost equal to that of V10A10_C4.¹⁹

The two types of consonance differ also in the range of mean scores. Considering identical consonances, the mean ratings are very similar across all composites, while the largest spread of mean ratings is consistently that of the nested consonances. This relationship is confirmed in Figure 4.8, representing the standard deviation between the mean scores to each of the VAME scales across alignment conditions. Notice that the smallest standard deviations are those associated with identical consonant composites (V5A5_C, V10A10_C, & V8A8_C) and the largest standard deviations are those associated with the nested consonant composites (V5A10_C & V10A5_C). This suggests that there was

more influence of the specific AV combination on the nested consonances than on the identical consonances. The latter appear to have been rated consistently high regardless of the audio and visual components. These relationships explain the statistically significant interaction between AV composite and alignment condition mentioned in the previous paragraph.

Figure 4.6a. Mean subject ratings to the two VAME scales when viewing the combination of Visual #1 and Audio #1 across alignment conditions.

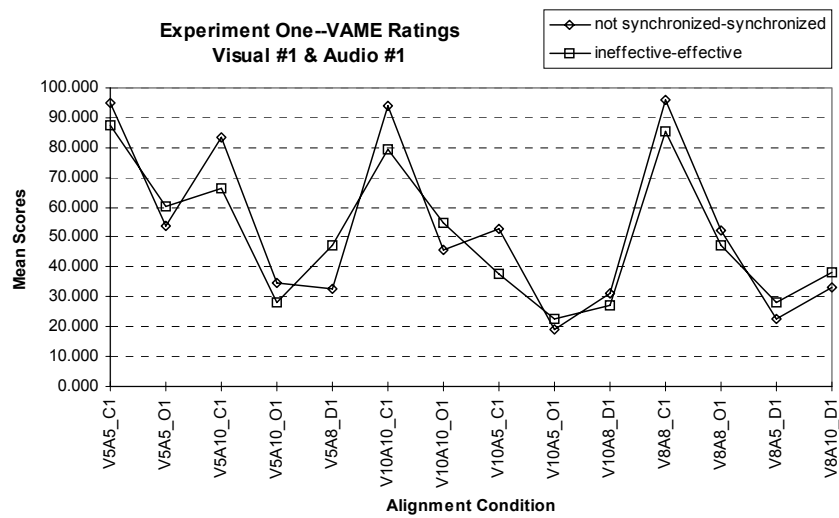


Figure 4.6b. Mean subject ratings to the two VAME scales when viewing the combination of Visual #1 and Audio #2 across alignment conditions.

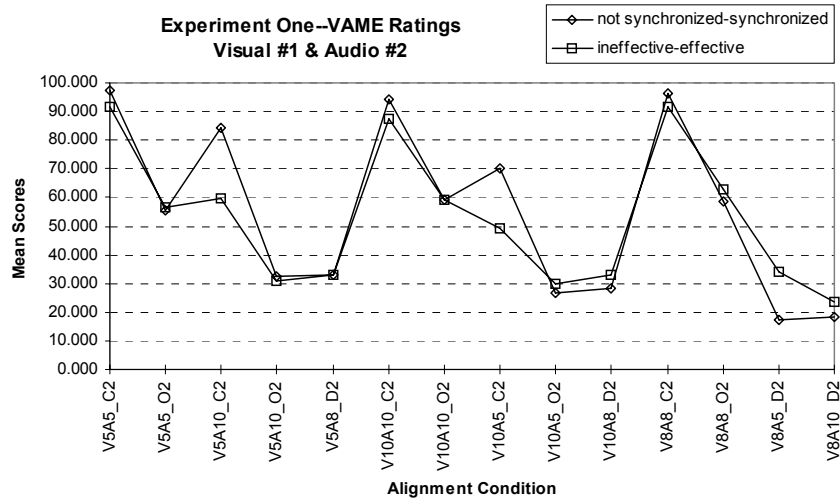


Figure 4.6c. Mean subject ratings to the two VAME scales when viewing the combination of Visual #2 and Audio #1 across alignment conditions.

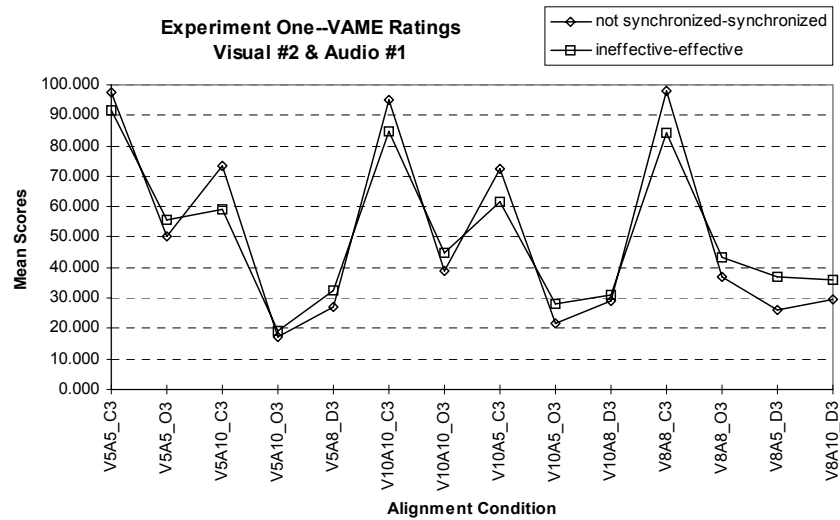


Figure 4.6d. Mean subject ratings to the two VAME scales when viewing the combination of Visual #2 and Audio #2 across alignment conditions.

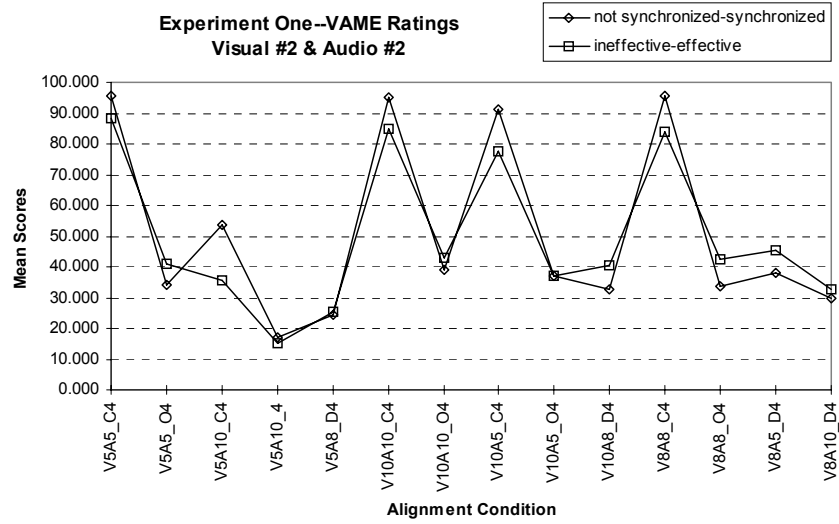


Figure 4.7. Comparison of all VAME responses across AV composite and alignment conditions.

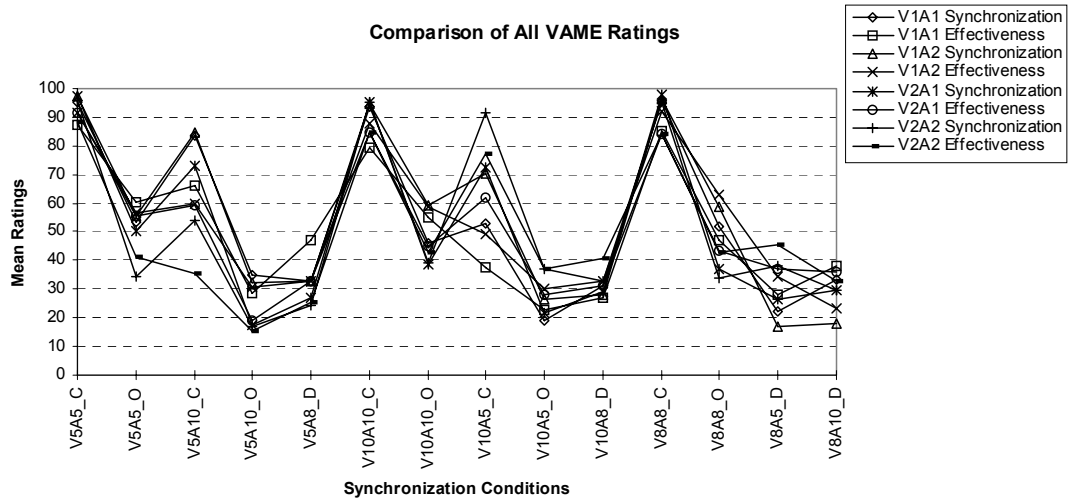
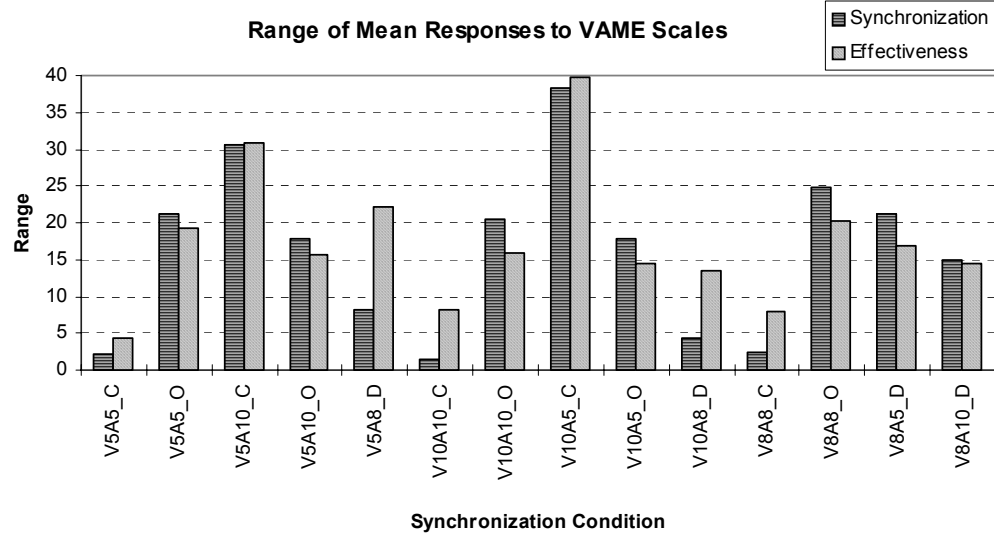


Figure 4.8. Standard deviations of mean responses to VAME scales across alignment conditions.



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. Therefore, the relationship between subject responses on the two VAME scales and accent structure alignment may be represented as shown in Table 4.5.

Table 4.5. AV composites arranged from highest response to lowest on the VAME scales.

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

Collapsing Alignment Conditions Across AV Composites. One research question explicitly addresses alignment of the audio and visual components of an AV composite.

Therefore, the subject VAME responses were collapsed across alignment conditions. When compared to a single measurement, such multiple measures of a single condition provided 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 statistically significant (synchronization— $F_{\lambda(6,102)} = 1.98959, p < .056$; effectiveness— $F_{\lambda(6,102)} = 2.18760, p < .117$). Further justification for collapsing the data in this manner is provided graphically in Figure 4.9, which presents 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. This same method of data analysis was utilized in Experiments Two and Three. Therefore, the null hypothesis for these collapsed data sets may be represented as:

$$H_0: \mu C = \mu O = \mu D$$

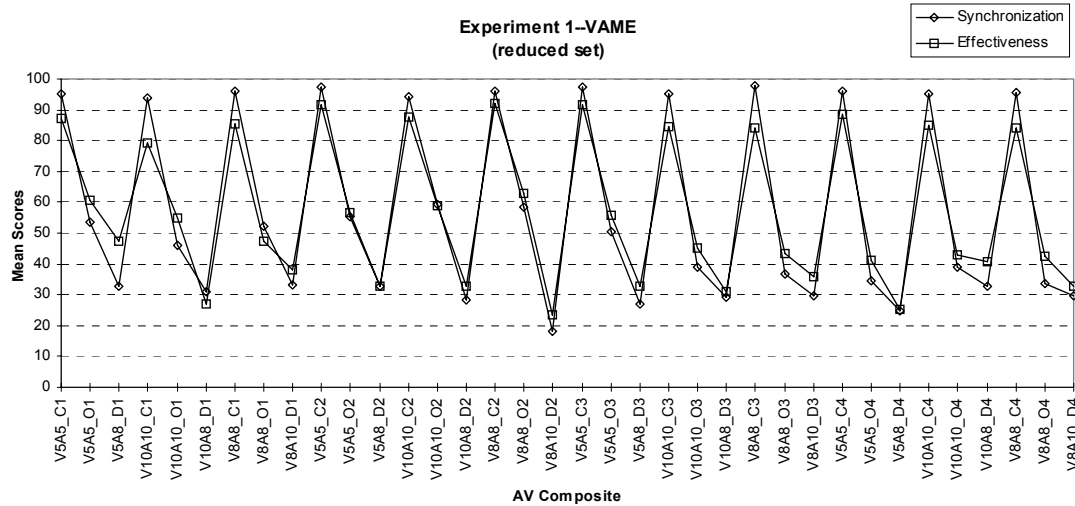
where:

C = consonant

O = out-of-phase

D = dissonant

Figure 4.9. VAME ratings for all consonant, out-of-phase, and dissonant combinations across all AV composites.



Collapsing the variables in this manner also alleviated a potential difficulty in pairing auditory and visual stimuli for Experiments Two and Three. Because the visual and auditory stimuli for Experiment One were created by the author, it was possible to completely control the accent periodicities (i.e., IOIs) so that alignment, misalignment, and nesting were possible using the same stimulus set. However, since Experiment Two (experimental animations by Norman McLaren) and Experiment Three (excerpts from the film “Obsession”) incorporated more ecologically valid stimulus materials, the amount of control is lessened appreciably. For the remainder of this investigation, only three alignment conditions were considered: consonant, out-of-phase, and dissonant, eliminating the nesting conditions. 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.

Before proceeding, the subject responses obtained in Experiment One were subjected to the same analyses as the later experiments, i.e., collapsing across alignment condition and eliminating the nesting conditions. Running a second ANOVA on this same data set caused the probability of alpha error (α) to increase. The data from each experiment was analyzed independently for significant differences and then one final ANOVA was run across the entire data set, including subject responses from all three experiments. Since the alpha error level was set *a priori* to .025, the resulting level of confidence remained above 95% (i.e., $.975 \times .975$).

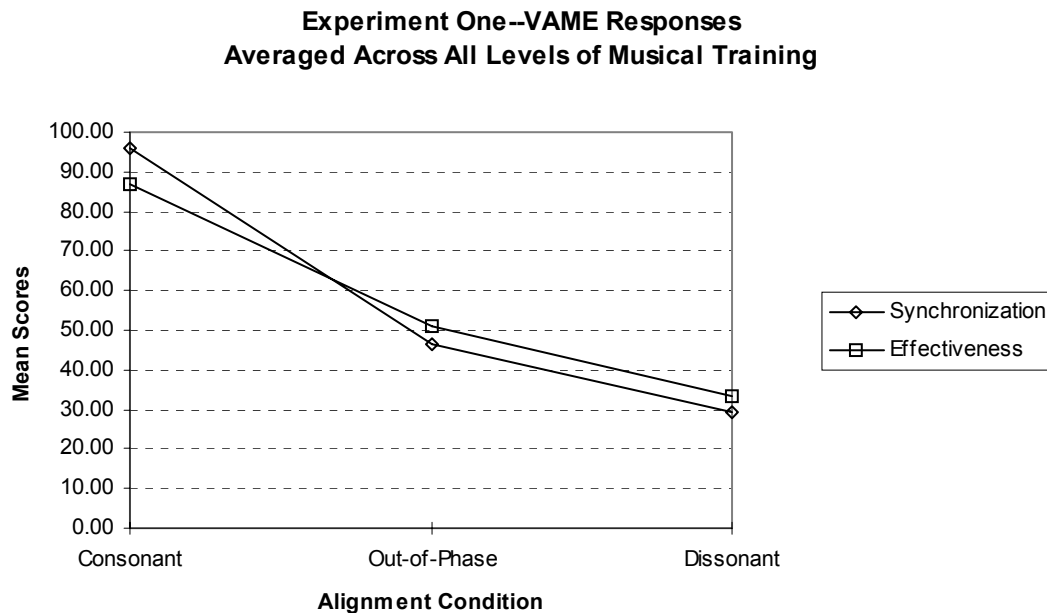
The single exception to this rule was the analysis of the data from Experiment One. One ANOVA was already run on the complete data set from the present experiment. Along with the following ANOVA on the collapsed data set and the final ANOVA across all three experiments, the resulting level of confidence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

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, there is 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 ($F_{\lambda(4,32)} = 1.575, p < .2048$) 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 4.10). 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 suggested 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 4.10. 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 compari-

sons) and one between-groups variable (3 levels of musical training).¹⁹ 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 4.11 provides the 3-dimensional solution, 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).²⁰ ²⁰ To further tease out the group membership among the various AV composites, the same triangular matrix was submitted for cluster analysis.

*Cluster Analysis.*²¹ Cluster analysis provided a method for dividing a data set into subgroups without any *a priori* knowledge considering the number of subgroups nor their specific members. Using Euclidean distance metric and the complete linkage (farthest

neighbor) method, the tree diagram presented in Figure 4.12 graphically illustrates the clustering of AV composites.

Figure 4.11. Multidimensional scaling solution for the similarity judgments in Experiment One.

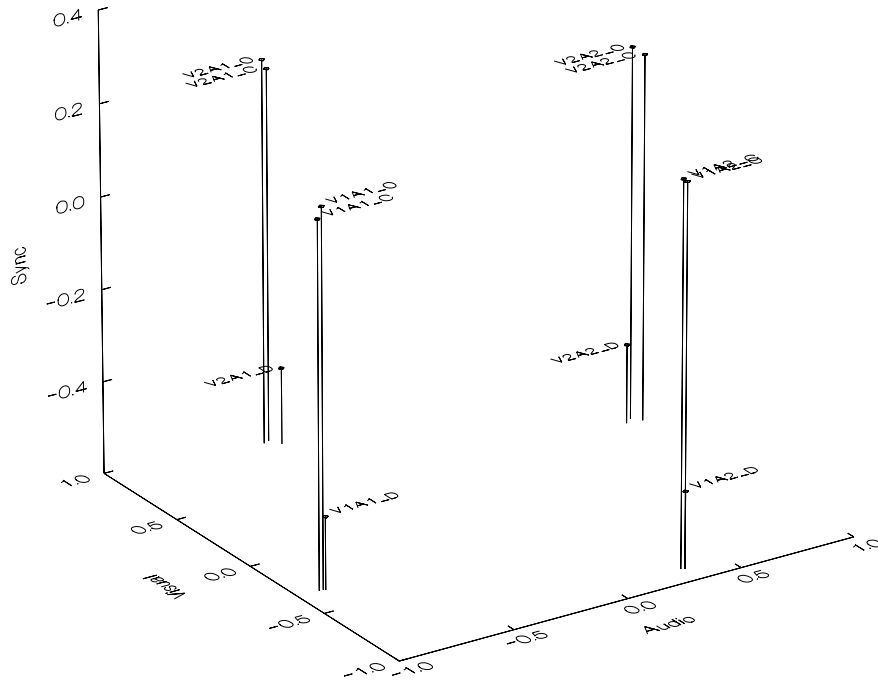
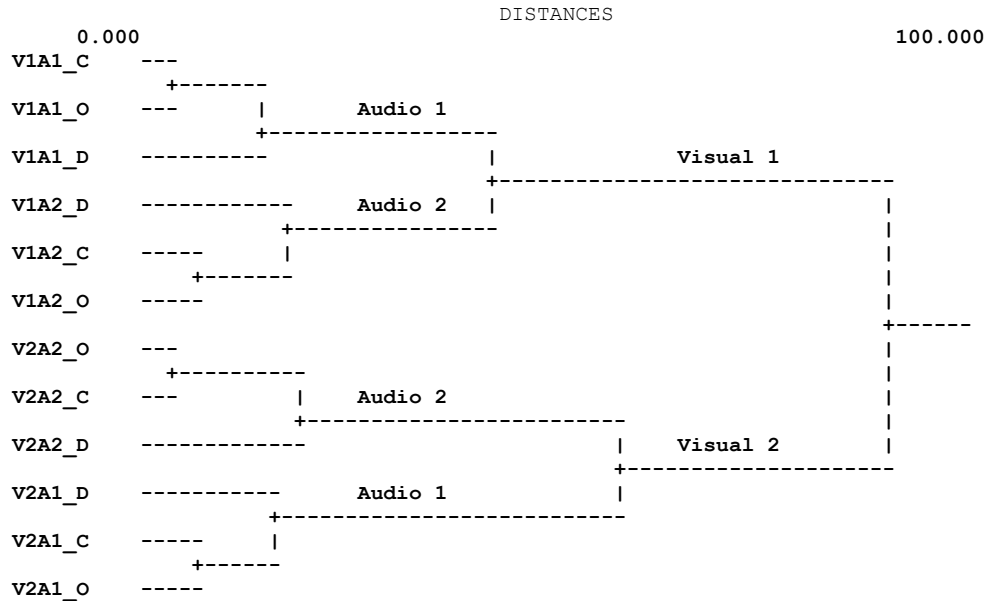


Figure 4.12. Cluster Analysis tree diagram—complete linkage (farthest neighbor)—for similarity ratings provided by subjects in Experiment One, Group Two.



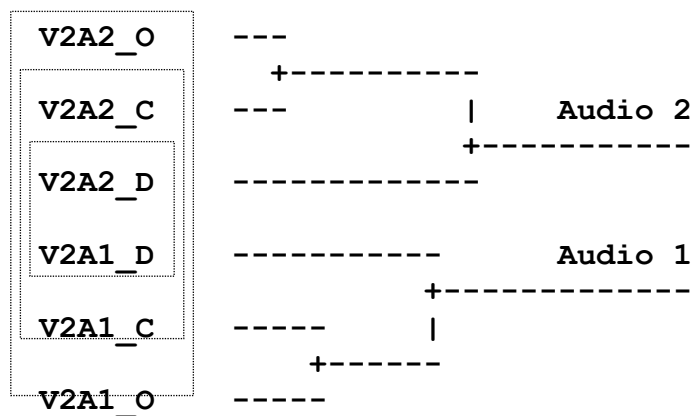
As is readily apparent when considering this cluster diagram from right to left, the initial branching of AV composites into subgroups clearly separates the composites according to the visual component, i.e., all composites on the upper branch utilize Visual One (V1) and all composites on the lower branch 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 4.13). 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. Also notice that, when considering the main (i.e., visual) branches in Figure 4.12, 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.

Conclusions

Summarizing the results of experimental tasks utilized in Experiment One, 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.

Figure 4.13. Illustration of the mirroring relationship between elements in the upper cluster of composites incorporating Visual One.



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.

CHAPTER FIVE

EXPERIMENT TWO

Stimulus materials used in Experiment Two were more representative of visual images and sounds that might be experienced in an actual theatrical setting, rather than the simple, single-object animations of Experiment One. This increase in complexity, both in the auditory and visual domains, consequently increased experimental validity.

Exploratory Studies

Stimulus Selection. Two 8-second excerpts were selected by the author from each of three Norman McLaren animations (“Dots” [1940], “Canon” [1964], and “Synchrony” [1971]). Since the Pioneer LD-4400 to be used in Experiments Two and Three was a single-sided laserdisc player (i.e., disks must be manually flipped), it was necessary to select animation excerpts that were on either one side of the laserdisc or the other so that the experimental stimuli could be presented to each subject in random order, eliminating any order effect.²² A small exploratory study, using five UCLA graduate students, was carried out to assist in selecting three examples that were maximally different. Subjects viewed all possible pairs of stimuli in a random presentation order and rated them on a scale of “not same - same.” The triangular matrix of mean responses was submitted for cluster analysis, enabling the selection of an excerpt from each animation that was maximally different from the others. As a result, the excerpts represented with

Frame Start and Frame End values in Table 5.1 were selected for use in Experiment Two. Musical notation for the audio portions of these excerpts is provided in Figure 5.1.

Table 5.1. Excerpts of Norman McLaren’s animation used as stimuli in Experiment Two; all excerpted from Side 2 of the Pioneer Special Interest laserdisc entitled “The World of Norman McLaren: Pioneer of Innovative Animation” (catalog number: PSI-90-018; part of the Visual Pathfinders series).

<i><u>Title (Date)</u></i>	<i><u>Chapter #</u></i>	<i><u>Start Frame</u></i>	<i><u>End Frame</u></i>
Dots (1940)	1	2285	2535
Canon (1964)	4	16500	16750
Synchromy (1971)	6	41009	41259

Establishing Accent Structure. The author’s analysis of the animations used in Experiment Two suggested the following accent periodicities (IOIs): “Dots” (500ms), “Canon” (667ms), and “Synchromy” (begins at 1333ms, halved abruptly to 667ms half-way through). These IOIs were logical since—at 12 frames per second (fps)—6 frames occupied an interval of exactly 500ms, while 8 frames took up 2/3 of a second (approximately 667ms). A second exploratory study was carried out to ensure that subjects perceived these accent periodicities. Nine UCLA undergraduates performed two tapping tasks, similar to that described in the exploratory study to Experiment One. Stimulus presentation order was randomly determined for every subject. In one task they tapped along with the musical sound for each of the animation excerpts, while the other task required them to tap along with the visual images in an attempt to pick up the underlying periodic accent structure. Each subject was allowed to practice tapping along with the stimuli as many times as s/he wished before recording his/her response. After completing each example, the subject was allowed to save the recorded IOIs or try again. The resulting mean tap rates are presented in Figure 5.2. Subject responses were equalized, as in Experiment One, to account for nesting and/or subdividing. The mean scores of the subjects tapping at a rate nesting or subdividing IOI perceived by the majority of subjects

were transformed by either multiplying or dividing by 2, as appropriate. The equalized mean scores for the audio task—“Dots” (502.75ms), “Canon” (660.63ms), and “Synchrony” (1327.86ms) confirmed that the subjects did perceive the predicted periodic accent structure (Figure 5.3). The mean tapping rate of Subject 5 (918.13ms) was eliminated as an outlier from the IOI calculation for “Synchrony,” since it could not be accounted for in terms of nesting or subdividing and deviated so drastically from the IOI perceived by every other subject.

Figure 5.1. Musical notation for audio excerpts of the Norman McLaren animations. Permission to use granted by Pioneer Entertainment (USA) L.P., based on their license from National Film Board of Canada.



Figure 5.2. Mean subject responses for the exploratory study to Experiment Two.

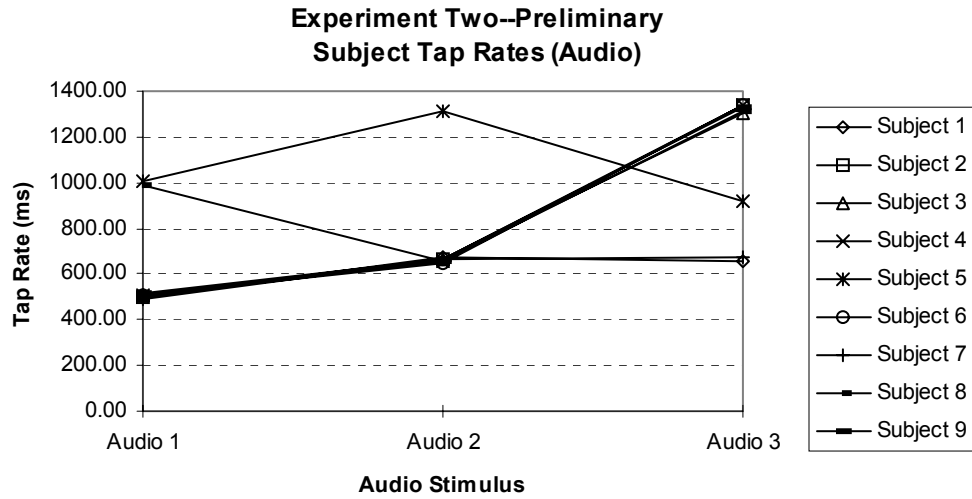
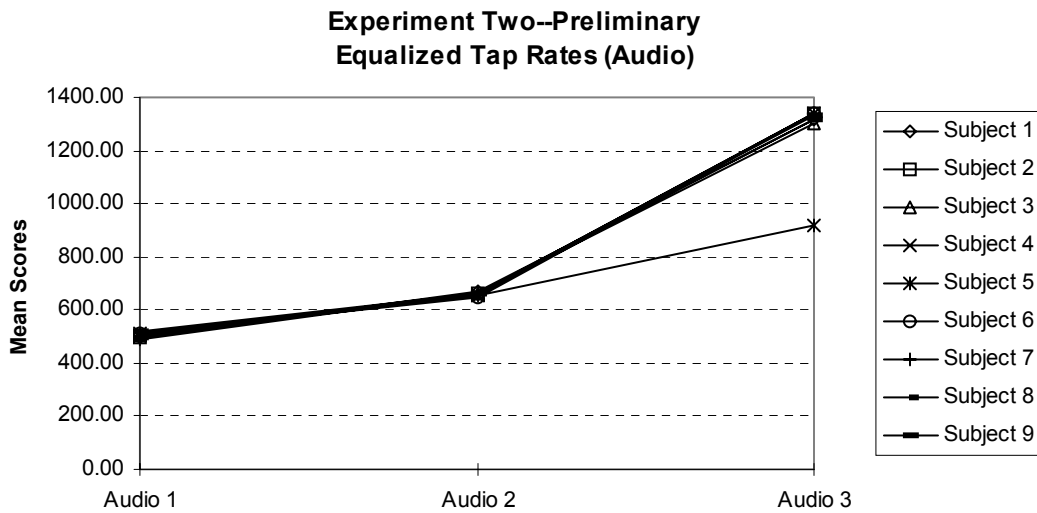


Figure 5.3. Equalized mean subject responses for the exploratory study to Experiment Two.



Subject responses to the visual tapping task were more variant and much less precise than responses to the audio task. In fact, several subjects commented that the task

seemed quite unnatural (i.e., tapping along with visual images). Though the subjects did not tend to tap at a steady IOI consistently, observation of their responses revealed several common strategies, allowing determination of events that appeared to consistently signal increased salience. As a result, the following events seemed to increase reliably the probability that the subject would tap the spacebar at a given moment: appearance of an object or shape on the screen, sudden change in size or shape, and objects in motion coming to a rest.

Alignment Conditions. Since subject responses to the auditory tapping task were more consistent than responses to the visual task and provided a metric upon which to base AV alignment, the accent periodicity of the audio track was used as a basis for determining alignment conditions for Experiment Two. A third exploratory study was run to determine the “most synchronized” (consonant) and the “least synchronized” (out-of-phase) alignment condition for each intended AV combination.

The auditory portion of each of the three animations was digitally sampled (22.05 KHz; 16-bit mono) and saved as a WAVE file. The visual image was accessed directly from a Pioneer LD-V4400 laserdisc player, while the sound was digitally reproduced by the computer sound card. The visual images were relayed to the computer monitor using a Hauppauge video overlay. Since the audio and visual stimuli were coming from two separate computer-controlled sources (i.e., the computer sound card and the laserdisc player), it was important to ensure that the synchronization created by this computer-controlled setup matched exactly that intended by McLaren. Four UCLA graduate students in the Department of Ethnomusicology and Systematic Musicology assisted the author in this determination. Once the intended AV alignment had been recreated using the laserdisc and the appropriate WAVE audio file, a series of altered WAVE files were

created so that the soundtrack was moved backward and forward in time in relation to the visual image.

The amount of temporal shift utilized in stimuli for this final exploratory study was based upon the perceived IOIs in the auditory tapping task described above. This IOI was divided by five (to avoid simple duple or triple nesting effects) and used as the base amount of temporal shift to create 11 alignment conditions for each AV composite. For example, since the perceived IOI for “Dots” was 500ms, the base temporal shift (i.e., IOI divided by 5) was 100ms. In addition to the intended alignment, five combinations were created in which the audio accents were delayed (by 100ms, 200ms, 300ms, 400ms, and 500ms) and five combinations were created in which the auditory accents preceded the visual accents (by 100ms, 200ms, 300ms, 400ms, and 500ms). “Canon” and “Synchrony” had perceived IOIs of approximately 667ms and 1333ms, respectively. As a result, the base amount of temporal shift for “Canon” was 133ms (i.e., by 133ms, 266ms, 399ms, 532ms, and 665ms), according to the system outlined above. However, if the same method were utilized for the excerpt from “Synchrony,” the resulting offsets would be 266ms, 532ms, 798ms, 1064ms, and 1330ms. Using 266ms as a base amount resulted in a situation where the largest amount of temporal offset (1330ms) was approaching Fraisse’s (1982) upper boundary for grouping consecutive events (i.e., 1500ms). Instead of using such a large increment as the base amount of temporal shift, the investigator decided to use the nested amount of 667ms (identical to that of “Canon”).

Since there were eleven AV alignments for each of the three animation excerpts, the stimulus set consisted of 33 AV combinations. Ten UCLA undergraduate students were asked to rate each of these AV composites (presented in random order) on a scale of “not synchronized - synchronized.” Mean subject responses revealed that, in every case, the highest ratings of synchronization were given in response to the alignment intended

by McLaren. This AV composite was selected as the consonant alignment condition. The combination exhibiting the lowest mean rating of synchronization was chosen for use as the out-of-phase alignment condition. For “Dots,” the out-of-phase combination was that in which the audio accents occurred 100ms before the visual accents. In both “Canon” and “Synchrony,” the out-of-phase combination was that in which the audio accents occurred 532ms after the visual accents.

The small temporal interval for offset in the out-of-phase alignment condition for “Dots” may seem surprising when compared to that selected for “Canon” and “Synchrony.” An explanation for this difference may be found through a general comparison of the three animations. The excerpt from “Dots” consists of a complex and rhythmically-interesting sequence of shapes (i.e., dots) appearing, disappearing, and “interacting” on the screen, accompanied by sound bursts that map onto every visual event. The portion of “Canon” used in the experiment showed four wooden blocks moving from square to square of a checkerboard, creating the illusion of a ballroom dance to the accompaniment of the musical canon (or “round”) “Frère Jacques.” The “Synchrony” excerpt consisted of a single vertical strip in the center of the screen that matched exactly the hand-drawn markings passing over the photoelectric cell on the soundtrack portion of the film that, resulting in a very simple scalar musical passage. As a result, the excerpt from “Dots” exhibited a higher level of complexity both in the audio track and the visual image (informally quantified in terms of the number of events per unit time) than the other two animations. For this reason, subjects appear to have found that offsetting the audio and visual components by a smaller amount in such a complex AV composite resulted in a lower level of perceived synchronization than when offset by a larger temporal interval.

As mentioned previously, the stimuli for Experiments Two and Three differed from those in Experiment One, because the use of ecologically valid stimuli—by necessity—limited the amount of control that the experimenter had over possible alignment conditions. In Experiment One, it was possible to create identical consonant, nested consonant, out-of-phase (nested), out-of-phase (identical), and dissonant combinations using the same visual and musical patterns. However, using actual animation or movie excerpts presented certain restrictions. For example, though the intended visual and audio portions of a given animation may have been combined in either a consonant or out-of-phase relationship, it was—by definition—impossible to create a dissonant combination using this same AV combination. Recall that a dissonant combination required that accent periodicities of the audio and visual components occurred at different IOIs. No matter how sounds and visual images with the same IOI were to be aligned, their accent periodicities would continue to occur at exactly the same rate as the originally-intended combination and could not be dissonant. It was equally impossible, by definition, to create a consonant or out-of-phase alignment condition from sound and images that shared a dissonant relationship.

Experiments Two and Three incorporated three alignment conditions for each of the animations. The consonant condition paired the intended audio with an animation, exactly as intended by McLaren. The out-of-phase condition paired the same audio with the animation, but offset by an amount determined by the exploratory study described above. The dissonant condition could not combine images from “Canon” and audio from “Synchrony” (or vice versa) because the accent periodicities were nested (i.e., “Canon” has an IOI of 667ms, while “Synchrony” has an IOI of 1333ms). Therefore, the dissonant condition for both of these excerpts utilized the audio track from “Dots.” The dissonant condition for the “Dots” excerpt, however, could incorporate the audio track from

either “Canon” or “Synchrony.” The audio track from the former was selected, because subjects tapped more consistently to this audio track in the auditory tapping exploratory study than to the latter.²³

Main Experiment

Research Design, Hypotheses, and Subjects

The research design and hypotheses for this experiment are identical to those stated for Experiment One. Subjects for this experiment were 40 UCLA students (ages 19 to 25) taking general education (G.E.) classes in the Music Department—either African-American Heritage (Keyes, Winter 1995) or Analysis of Traditional Music (Keyes, Winter 1995).²⁴ The number of subjects falling into each of these categories is represented in Table 5.2.

Table 5.2. Number of subjects falling into each cell of the between-subjects design (Experiment Two).

<u>Exp. Task</u>	<u>Musical Training</u>		
	<i>Untrained</i>	<i>Moderate</i>	<i>Trained</i>
VAME	8	7	5
Similarity	6	9	5

Equipment

Experimental stimuli were presented on an IBM-compatible 90 MHz Pentium computer with a Turtle Beach Monterey sound card and a 16-bit Hauppauge video overlay. The Pioneer LD-4400 laserdisc player was computer-controlled via an RS-232C serial interface connection and the visual image was routed through the video overlay so that it could be presented on the computer monitor. The subjects heard the audio track through a pair of Sennheiser HD222 headphones.

Stimulus Materials

The stimuli consisted of 9 AV composites resulting from the combination of visual images and audio tracks selected in the pilot study above. In the following discussion, the abbreviations listed in Table 5.3 are used to refer to audio and visual components and their various combinations. Using these abbreviations, the 9 excerpts (3 visual excerpts x 3 alignment conditions) are referred to as represented in Table 5.4.

Table 5.3. Abbreviations used to identify the various audio and visual components in Experiments Two.

<u>Abbreviation</u>	<u>Description</u>
<i>Visual</i>	
V1	visual excerpt from “Dots”
V2	visual excerpt from “Canon”
V3	visual excerpt from “Synchrony”
<i>Audio</i>	
A1	audio track from “Dots”
A2	audio track from “Canon”
A3	audio track from “Synchrony”

Table 5.4. Labels used to identify each specific AV Composite based on the abbreviations presented in Table 5.3.

<u>Alignment Condition</u>	<u>AV Composite Label</u>
<i>Consonant</i>	V1A1_C
	V2A2_C
	V3A3_C
<i>Out-of-Phase</i>	V1A1_O
	V2A2_O
	V3A3_O
<i>Dissonant</i>	V1A2_D
	V2A1_D
	V3A1_D

Experimental Tasks

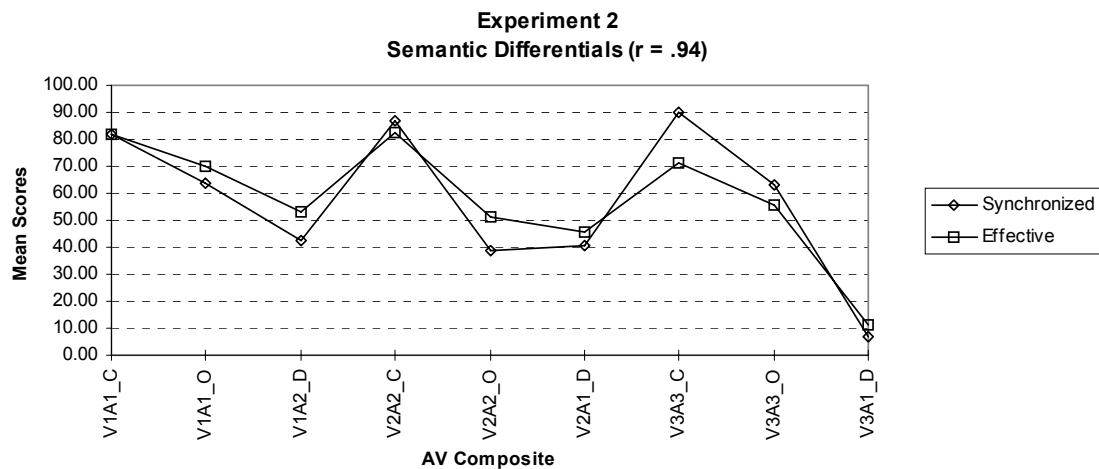
Subjects were once again randomly assigned to one of two groups: Group One (n = 20) responded on a verbal rating scale and Group Two (n = 20) provided similarity judgments between pairs of stimuli. The experimental procedures were identical to those described in the discussion of Experiment One and were carried out according to the instructions provided in the Methods chapter. Since the total number of stimuli was smaller for this experiment, subjects in Group One responded to only 9 AV composites on the VAME scales of “not synchronized–synchronized” and “ineffective–effective.” Group Two viewed a total of 45 pairs of AV composites, rating them on a scale of “not same–same.” In each of the main experiments, subjects were provided a series of practice examples (3 in the VAME task and 6 in the similarity scaling task), as a means of familiarizing them with the experimental procedure. Stimulus presentation order was randomly determined for every subject.

Results

Group One Data Analysis and Interpretation. As in Experiment One, a repeated measures ANOVA was performed on the subject responses to each of the VAME rating scales provided by Group One, considering one within-groups variable (3 AV alignment conditions) and one between-groups variable (3 levels of musical training). The alignment condition data were calculated by taking the average of the 3 AV composites utilizing the same alignment condition (e.g., all composite pairs—V1A1_C, A2A2_C, & V3A3_C), resulting in 3 separate measures of consonant, out-of-phase, and dissonant combinations for every subject. Figure 5.4 provides the mean subject responses for both VAME scales. Once again, the consonant pairs consistently received the highest ratings. In the composites utilizing V1 and V3, the relationship between out-of-phase and dissonant combinations was consistent with that seen in Experiment One (i.e., the out-of-phase

condition received a rating lower than the consonant pair, but higher than the dissonant pair). However, the out-of-phase and dissonant ratings for composites incorporating V2 (“Canon”—image of blocks moving on a checker board) were both rated almost equally low. In fact the mean synchronization rating for the out-of-phase composite was actually lower than that of the dissonant combination. This deviation may be a result of the fact that A1 was the most rhythmically complex of the three audio tracks. It is possible that this complexity (i.e., number of musical events per unit time) resulted in a perception that enough of the musically salient events aligned with visually salient events in V2 which made the dissonant alignment condition appear more synchronized than hypothesized.

Figure 5.4. Mean subject VAME responses to the AV composites in Experiment Two.

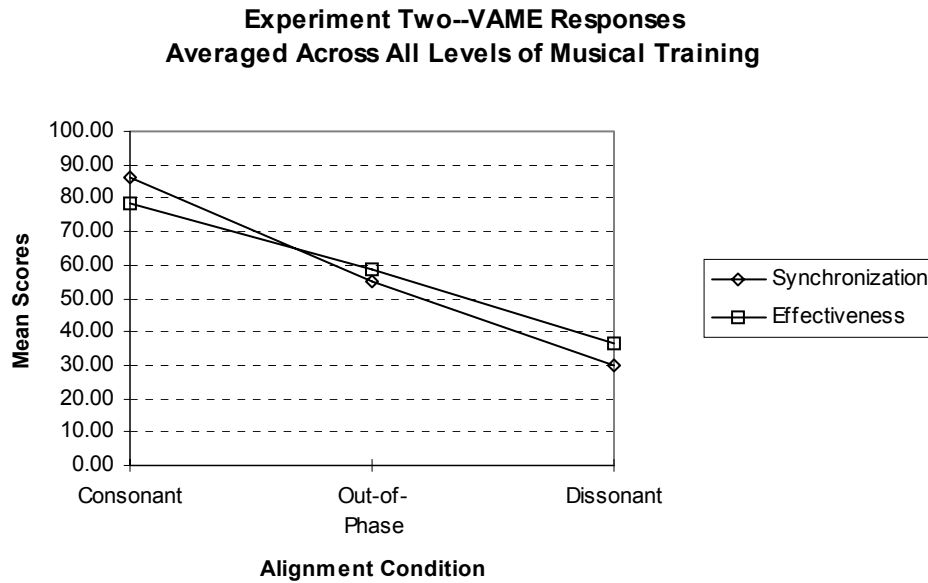


Considering the collapsed data ($\alpha = .025$), neither the synchronization ratings nor the ratings of effectiveness exhibited any significant effect of either the level of musical training (synchronization— $F_{(2,17)} = 0.137$; $p < .873$; effectiveness— $F_{(2,17)} = .409$; $p < .671$) or the interaction between musical training and alignment condition (synchronization— $F_{(4,34)} = 1.643$, $p < .1861$; effectiveness— $F_{(4,34)} = 1.045$, $p < .3987$). However,

there was a highly significant effect of alignment condition on both VAME scales (synchronization— $F_{(2,34)} = 73.704, p < .0001$; effectiveness— $F_{(2,34)} = 34.586, p < .0001$).

These results confirmed that subjects continued to distinguish between the three alignment conditions when considering moderately complex AV stimuli, regardless of their level of musical training. Figure 5.5 shows the mean subject responses to the consonant, out-of-phase, and dissonant composites, averaged across all subjects. The highest ratings were given to the consonant composites, the lowest ratings were given to the dissonant combinations, and the out-of-phase composites received a rating approximately halfway between the other two conditions. Notice that the mean effectiveness ratings were once again less extreme than the mean synchronization ratings, i.e., when the synchronization rating was high (consonant alignment) the effectiveness rating was slightly lower and when the synchronization rating was relatively low (out-of-phase and dissonant conditions) the effectiveness rating was slightly higher. Also, as in Experiment One, there was a high correlation between responses on the two VAME scales gathered in Experiment Two ($r = .94$), suggesting a strong relationship between subject ratings of synchronization and effectiveness as audio-visual alignment conditions change.

Figure 5.5. Mean VAME responses from subjects in Experiment Two, collapsed across alignment condition.

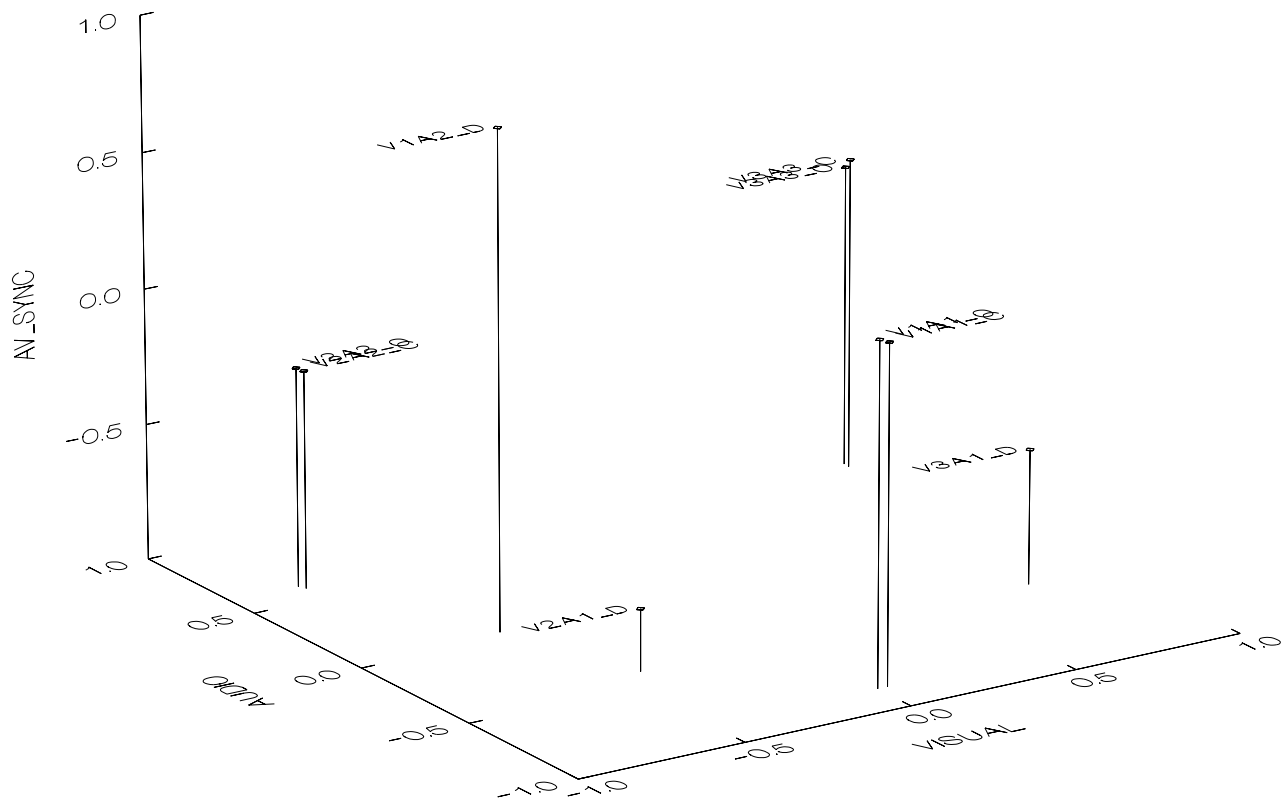


Group Two Data Analysis. To ensure that there was no difference in similarity ratings as a function of level of musical training, a repeated measures ANOVA was performed on the responses provided by Group Two, considering a single between-groups variable (3 levels of musical training) and one within-groups variable (45 paired comparisons). There was no significant effect of either musical training ($F_{(2,17)} = .16, p < .852$) or the interaction between musical training and similarity ratings ($F_{(88,748)} = .72, p < .971$). However, the difference between mean similarity ratings was found to be highly significant ($F_{(44,748)} = 42.37, p < .0005$).

Multidimensional Scaling. The MDS solution for Experiment Two is provided in Figure 5.6. This 3-dimensional solution accounted for 99.95% of the variance at a very low level of stress (i.e., .00749). The three dimensions can be identified as labeled in the figure below, though the resulting configuration was not as easily interpreted as that

observed in Experiment One. All composites incorporating V3 are located on the positive side of the “Visual” dimension while all those incorporating V2 are on the negative side. The V1 composites utilizing the intended soundtrack (both consonant and dissonant) are located near the zero-point, while the dissonant combination appears to be pulled toward the other composites accompanied by A2. All composites utilizing A1 are on the negative side of the “Audio” dimension, while those incorporating either A2 or A3 are on the positive end of the scale.²⁵ Finally, the consonant and out-of-phase alignment conditions for each intended AV composite are placed near the zero-point of the “AV Sync” dimension, while the dissonant combinations are placed near the extremes of the dimension.

Figure 5.6. MDS solution derived from mean similarity judgments in Experiment Two.



In the figure, notice how the consonant and out-of-phase alignment conditions of the intended combination are very tightly clustered within the 3-dimensional space and separated by a relatively large distance from the dissonant combination using the same visual image. This dimension might simply be thought to differentiate between intended combinations of audio and visual components (i.e., consonant and out-of-phase) and unintended combinations (i.e., dissonant)—dimension based on “appropriateness” of the AV pair. However, the exploratory study confirmed that subjects did reliably perceive accent periodicities in the audio portion of the AV composite. The animations used in this experiment exhibited a tight mapping of the audio accent structure and the visual images. Alignment of audio and visual accent structures is proposed, therefore, as an

important factor in the determination of whether an AV combination is appropriate.²⁶ In order to confirm or disconfirm this explanation, the subject data was subjected to cluster analysis.

Cluster Analysis. Using Euclidean distance metric and the complete linkage (farthest neighbor) method, the tree diagram presented in Figure 5.7 graphically represents the clustering of AV composites in Experiment Two. In the resulting tree diagram, similar to the MDS solution for this same experiment, the divisions are not quite as clear as the clustering of subject responses in Experiment One. The first branch separated composites incorporating V3 from those using V1 and V2. The next division on the Visual 3 branch separates the dissonant (unintended) composite from the consonant and out-of-phase composites (intended AV combinations). The upper branch, however, separates at its next division according to audio component, rather than continuing to separate initially on the basis of visual image.

The Audio 1 and Audio 2 branches are then divided on the basis of alignment (or appropriateness), separating the dissonant (unintended) composite from the consonant and out-of-phase composites (intended AV combinations). Cross-cluster neighbors reveal that the audio and visual components were the main determining factor in joining the branches (e.g., Audio 2 and Audio 1 branches are joined by composites sharing the same visual component [V1A2_D & V1A1_O] and the Visual 3 and Visual 1 & 2 branches are joined by composites sharing the same audio component (V2A1_D & V3A1_D). Notice, in this latter case, that the cross-cluster neighbors share the same alignment condition (dissonant), as well. In fact, careful observation of the lowest six branches of the tree diagram (Figure 5.8) reveals a mirroring similar to that observed in Experiment One, proceeding from the center with the dissonant composites (V2A1_D & V3A1_D), neighbored by the consonant (V1A1_C & V3A3_C) and then out-of-phase

(V1A1_O & V3A3_O) composites. Therefore, in addition to audio and visual components, alignment condition did appear to play a role in subject similarity ratings given in response to a series of moderately complex AV composites, though its importance was subordinate to the influence exerted by either the audio and/or visual components.

Figure 5.7. Cluster Analysis tree diagram—complete linkage (farthest neighbor)—for similarity ratings provided by Group Two subjects in Experiment Two.

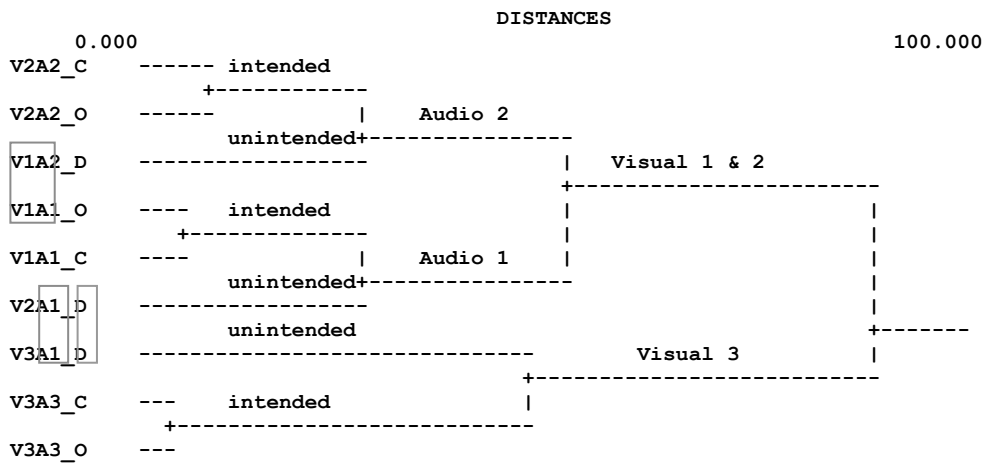
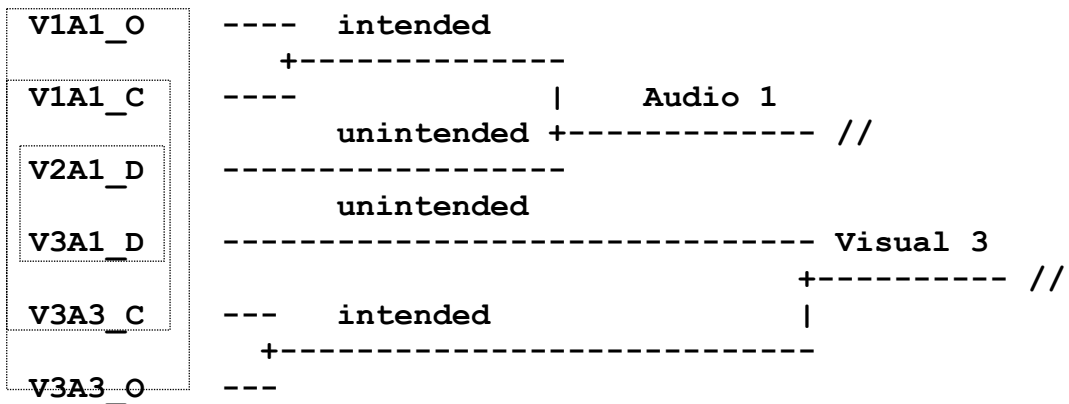


Figure 5.8. Illustration of the mirroring relationship between the lowest six branches in the cluster tree diagram from Experiment Two (Figure 5.7).



Conclusions

The results of Experiment Two continued to substantiate the role played by alignment condition in both the VAME ratings and similarity judgments given in response to a series of AV composites, though its influence appeared to have weakened with the increase in stimulus complexity. Regardless of level of musical training, subjects in Group One distinguished clearly between the 3 alignment conditions on both VAME scales, giving the highest ratings to consonant composites, lowest ratings to dissonant combinations, and out-of-phase combinations received a mean rating in-between the other two. As in Experiment One, mean scores for effectiveness continued to be less extreme than those for synchronization, i.e., when the mean synchronization rating was high (consonant alignment) the mean effectiveness rating was slightly lower and when the mean synchronization rating was relatively low (out-of-phase and dissonant conditions) the mean effectiveness rating was slightly higher. The high correlation between synchronization and effectiveness ratings further confirmed the close relationship between these two concepts.

The MDS solution derived from subject similarity ratings of Group Two once again revealed three interpretable dimensions, though these dimensions were not as clearly segregated as the solution for the responses to the less complex stimuli in Experiment One. As in the previous experiment, the first two dimensions were related to the visual and audio component, though the audio dimension may be more accurately described as including an aspect of musical complexity. The third dimension—AV synchronization—was unique in that the consonant and out-of-phase composites clustered near the zero-point of this dimension, while the dissonant composites were located at the extremes (both positive and negative). After closer consideration, the investigator determined that this third dimension might be described as a continuum of “appropriate-

ness” of AV combination, taking into account both stylistic congruency and accent alignment.

Cluster analysis of the mean responses of Group Two confirmed the roles played by audio component, visual component, and alignment condition in subject similarity judgments. The branching structure of the tree diagram was not as symmetrical as the solution in Experiment One, but the resulting clusters were easily interpreted. In response to these moderately complex stimuli, an interaction between musical and visual components appeared to have developed. Though the initial branching was based on the visual component, the second branch revealed a dominance of music (A1 vs A2) over visual image (V1 vs V2). This contrasted with the results of Experiment One in which groups were formed wholly according to—in decreasing order of significance—visual component, audio component, then alignment condition.

A Reassessment. The fact that the similarity scaling task failed to show a strong influence of accent structure alignment begs explanation. Were subjects perceiving accent structure relationships or were their responses based solely on the audio and visual component and some measure of their mutual congruency? A close look at Figure 5.5 reveals that subjects in the VAME task were, in fact, capable of distinguishing between the consonant and out-of-phase alignment conditions of both VAME scales. A series of same-sample pairwise *t*-tests²⁷ were run to test the significance of the differences between response means to the various alignment conditions on the verbal scale of synchronization. All pairs were found to be significantly different from one another based on the statistical data provided in Table 5.5a. The data in Table 5.5b is presented to confirm the fact that the same statistically significant difference occurred in the subjects’ mean effectiveness ratings. Had it not been for this determination, the present investigation would have had to conclude that the alignment condition variable had been hope-

lessly confounded with congruency of the AV composite, forcing a reconsideration of the stimulus materials used. However, since subjects were clearly responding to these moderately complex stimuli on the scale of synchronization in a pattern identical to the responses of the subjects in Experiment One (to the simple animations), they were able to distinguish between AV composites on the basis of alignment condition. Most crucial to the continuation of the present investigation was the fact that the ratings for the consonant and out-of-phase alignment conditions were significantly different. This confirmed explicitly that, though two AV composites may have shared the same audio and visual components, subjects in the VAME task *did perceive the difference in alignment*. As hypothesized in the introductory chapter, this aspect of the AV composite appeared to play a significantly smaller role as the stimulus complexity increased, even moderately.

Table 5.5a. Paired *t*-test values for mean synchronization ratings in Experiment Two (graphically represented in Figure 5.5).

Comparison	<i>t</i>-value	df	2-tail significance
<i>Consonant to Out-of-Phase</i>	6.67	19	<.0005
<i>Consonant to Dissonant</i>	13.31	19	<.0005
<i>Out-of-phase to Dissonant</i>	5.04	19	<.0005

Table 5.5b. Paired *t*-test values for mean effectiveness ratings in Experiment Two (graphically represented in Figure 5.5).

Comparison	<i>t</i>-value	df	2-tail significance
<i>Consonant to Out-of-Phase</i>	4.60	19	<.0005
<i>Consonant to Dissonant</i>	7.56	19	<.0005
<i>Out-of-phase to Dissonant</i>	4.67	19	<.0005

CHAPTER SIX

EXPERIMENT THREE

The stimuli for this final experiment were selected from among currently available American popular motion pictures. One specific practical consideration was of paramount importance to stimulus preparation. A method of extracting the musical score from excerpts used in the experiment had to be determined so that alignment of the audio and visual tracks could be controlled by the investigator, as in the previous experiments.

In isolating the musical score, it was crucial that sound effects, dialogue, and any other ambient sound were removed as well. The author discovered a series of laserdisc recordings released by Pioneer Special Editions in which the musical score was isolated on one of the analog tracks. The author added necessary programming code to MEDS so that computer-control of the laserdisc player included the ability to switch between the various modes of audio playback. Therefore, it was possible to digitally sample the isolated musical excerpts and save them as WAVE files. Once these files were created, the same procedure utilized in Experiment Two could be followed in creating the various alignment conditions.

Exploratory Studies

Stimulus Selection. After viewing several laserdiscs in the Pioneer Special Editions series, "Obsession" (a Brian DePalma film with a musical score composed by Bernard Herrmann) was selected for use in Experiment Three. Since the scenes had to be

selected from a single side of the laserdisc—the Pioneer LD-4400 is a one-sided laserdisc player—the author decided to utilize the concluding side of the laserdisc, due to its high level of dramatic activity, leading to the climax of the film.

Thirteen 20-second excerpts were selected from side 2 of the laserdisc. The specific scenes were selected because they were considered to exhibit a high level of synchronization between the audio and visual accent structures. Four UCLA graduate students in the Department of Ethnomusicology and Systematic Musicology were instructed to rate each excerpt (visual image accompanied by only the isolated musical score) on a scale of “not synchronized–synchronized,” as determined by alignment of salient moments in the visual image with important events in the musical sound stream. The mean rating for each excerpt was used to determine which scenes presented the highest level of AV synchronization to the subjects in this exploratory study. The two scenes rated most synchronized were highly similar, both in musical and visual content. Both were set at an airport and focused on the same male character searching intently for another character. Likewise, the musical scores consisted of closely-related motivic ideas and instrumentation. In order to avoid using scenes that were too similar, the scene of this pair with the lower mean score was eliminated from consideration, and the two scenes with the next highest mean scores were selected for use in the experiment. These three excerpts, along with their Start Time and Stop Time values, are presented in Table 6.1.²⁸ In the following discussion, the audio and visual excerpts are referred to by the abbreviations provided in Table 6.2.

Table 6.1. Excerpts from side 2 of the Pioneer Special Edition laserdisc of “Obsession” (catalog number: PSE91-18).

<u>Chapter Title</u>	<u>Chapter #</u>	<u>Start Time (mm:ss)</u>	<u>End Time (mm:ss)</u>
Portrait of Elizabeth	3	4:47	5:07
Flashback	10	32:02	32:22
Reunion	13	40:26	40:46

Table 6.2. Abbreviations used to identify the various audio and visual components excerpted from “Obsession” for use in Experiments Three.

<u>Abbreviation</u>	<u>Description</u>
<i>Visual</i>	
V1	visual excerpt from “Portrait of Elizabeth”
V2	visual excerpt from “Flashback”
V3	visual excerpt from “Reunion”
<i>Audio</i>	
A1	audio track from “Portrait of Elizabeth”
A2	audio track from “Flashback”
A3	audio track from “Reunion”

Establishing Accent Structure. To determine accent structure of the audio and visual stimuli, four graduate students (including the author) studying in UCLA’s Department of Ethnomusicology and Systematic Musicology (all with a high level of musical training) participated in two tapping procedures similar to those used in the previous experiments. One procedure required the subjects to tap in response to visual excerpts, while the other required them to tap along with the audio tracks.

The four subjects were randomly assigned to two groups. One group responded to the visual images in the first procedure followed by the audio tapping procedure, while the other group responded to the audio tracks first. Due to the high level of complexity in these actual movie excerpts (in addition to expressive deviations such as rubato, crescendo/decrescendo, etc.), this task was substantially more difficult than the exploratory studies for the earlier experiments.

Figure 6.1. Notation representing excerpts from Bernard Herrmann's musical score for "Obsession."
Score reduction by author. Permission to use granted by Hal Leonard Corporation and George Litto
Productions, Inc.

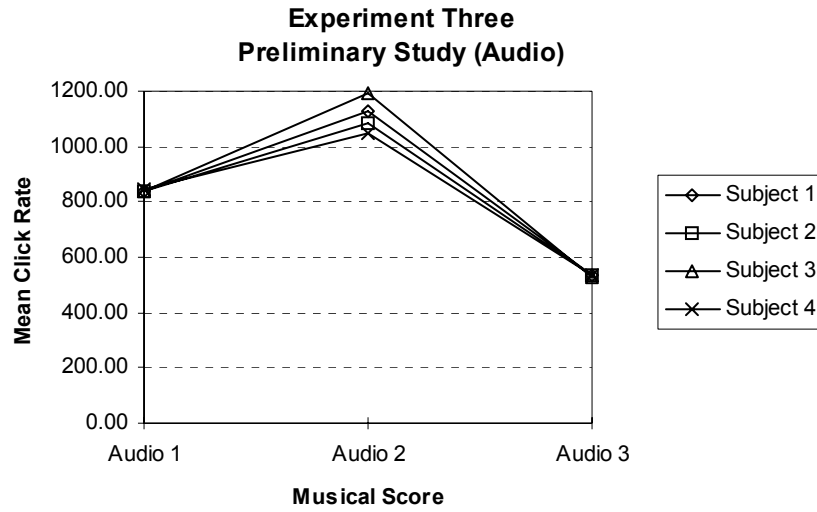


As in Experiment Two, subjects were allowed to practice tapping along with the stimuli as many times as they wished before recording their responses. After completing each example, they were allowed to either save the recorded IOIs or try again. As always, the stimuli were presented in a random order to every subject.

Subject tap rates in the audio exploratory study exhibited a high degree of agreement in the perceived accent periodicity. The IOIs between taps were equalized as in the previous exploratory studies to arrive at the mean tap rates plotted in Figure 6.2. Notice the high level of agreement in perceived accent periodicity. The slightly divergent mean tap rates observed for Audio Two probably resulted because musical excerpt exhibited the most prominent use of rubato of the three audio examples.

Subject tap rates to the visual image, however, did not reveal a consistent accent periodicity. Instead subjects simply tapped the spacebar whenever a significant visual event occurred, resulting in IOIs that were of widely varying durations. These moments in the visual image were observed by the investigator for between-subject consistencies. Events determined to be particularly salient in the visual image were sudden movements made by characters in the scene or abrupt changes of camera angle. Since the audio tap rates were once again more consistent than those in response to the isolated visual image, subject responses to the audio exploratory study were used as a basis for creating the various alignment conditions.

Figure 6.2. Equalized mean subject responses to the auditory portion of the exploratory study.



Alignment Conditions. Audio tracks of the three movie excerpts were once again saved as WAVE files (22.05 KHz; 16-bit mono) in order to allow the investigator control over synchronization of the AV composites. The same four UCLA graduate students that assisted in matching the intended alignment of animations in Experiment Two participated in the process of ensuring that the consonant alignment condition of the “Obsession” excerpts matched precisely the AV alignment as it was intended in the final edit of the motion picture. Using the same method described for creating alignment conditions in Experiment Two, the mean IOIs from the exploratory study above ($A1 = 840.39\text{ms}$; $A2 = 1113.73$; $A3 = 532.47$) were used to calculate the base amount of temporal shift (i.e., the mean tap rate divided by five) for each excerpt was as follows: $A1 = 168.08$, $A2 = 222.75$, $A3 = 106.49$. Since these excerpts were significantly longer than the animation excerpts used in Experiment Two (i.e., 20-seconds as opposed to 8-seconds), only five versions of each AV combination were created for use in the final exploratory study. In addition to the intended alignment, two combinations were made in which the audio accents preceded the visual accents (by 2 times the base temporal shift and 4 times the

base temporal shift; e.g., 336ms and 672ms for A1) and two combinations in which the audio accents were delayed by these same amounts. Each of the visual excerpts was combined with all five versions of every audio track so that subjects could rate all possible composites on a scale of synchronization, allowing determination not only of the least synchronized version of the intended AV combination, but also determining the least synchronized versions of all unintended AV combinations. Therefore, the total stimulus set consisted of 45 AV composites (3 visual excerpts x 3 musical tracks x 5 alignment conditions). The consonant condition used in Experiment Three was always the alignment intended by the composer, as it appeared in the motion picture. The out-of-phase and dissonant conditions were determined by the following exploratory study.

A group of ten UCLA undergraduates rated each of the 45 AV composites on a single scale of “not synchronized–synchronized.” The out-of-phase alignment condition was determined by the lowest mean synchronization rating given to any of the five alignment conditions of the intended AV combination, as described above. For V1 and V3, the out-of-phase alignment conditions selected were the combinations in which the audio track was delayed by 4 times the base temporal shift (672ms and 425ms, respectively). The out-of-phase alignment condition chosen for V2 was that in which the audio track was shifted in the opposite direction by 4 times its base temporal shift (890ms).

The dissonant alignment conditions were also determined from subject responses in this exploratory study. The author observed that the audio accent periodicity for A2 (532ms) came close to nesting within the accent periodicity for A3 (1113ms). Since it was important that accent periodicities in the dissonant AV alignment conditions were as divergent as possible, V2A3 and V3A2 were eliminated from consideration in order to lessen the likelihood of nesting accent structures. Therefore, V2A1 and V3A1 were used as dissonant alignment conditions, leaving only the determination of a dissonant combi-

nation for V1. Since either A2 or A3 could have been used, the final decision was based on the ratings of the ten combinations utilizing either A2 and A3 with V1 in the previous exploratory study. The lowest mean synchronization rating (33.50) for V1A2 was that in which the audio track was delayed by 891ms (4 times the base temporal shift), while the lowest mean synchronization rating for V1A3 (47.70) was that in which the audio track was delayed by 436ms (also 4 times the base temporal shift). As a result, the former combination will be used as the dissonant alignment condition for V1. The 9 AV composites selected for use in Experiment Three are referred to using the labels presented in Table 6.3.

Table 6.3. Labels used to identify each specific AV composite to be used in Experiment Three; based on the abbreviations presented in Table 6.2.

<u>Alignment Condition</u>	<u>AV Composite Label</u>
<i>Consonant</i>	V1A1_C
	V2A2_C
	V3A3_C
<i>Out-of-Phase</i>	V1A1_O
	V2A2_O
	V3A3_O
<i>Dissonant</i>	V1A2_D
	V2A1_D
	V3A1_D

Main Experiment

Research Design, Hypotheses, Subjects, and Equipment

The research design and hypotheses were identical to those of the previous experiments. Subjects for this experiment were 40 UCLA students (ages 18 to 28) taking a general education (G.E.) class in the Music Department—Development of Rock (Keyes, Spring 1995).²⁹ These students were randomly assigned to one of the two experimental procedures: verbal scaling task (Group One; N = 20) or similarity judgments (Group

Two; N = 20). The number of subjects falling into each of these categories is represented in Table 6.4. Equipment used for stimulus presentation in both the exploratory studies and Experiment Three was identical to that utilized in Experiment Two.

Table 6.4. Number of subjects falling into each cell of the between-subjects design (Experiment Three).

<u>Exp. Task</u>	<u>Musical Training</u>		
	<i>Untrained</i>	<i>Moderate</i>	<i>Trained</i>
VAME	8	7	5
Similarity	8	9	3

Experimental Tasks

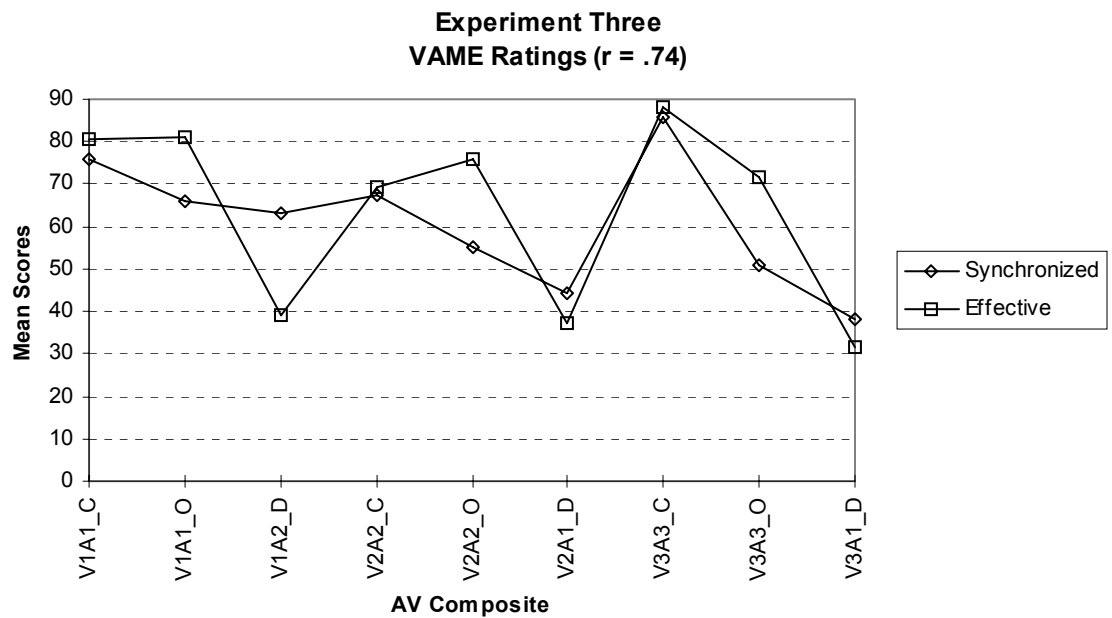
Both the VAME procedure and the similarity scaling task were identical to those described in Experiment Two, differing only in the stimulus materials presented (i.e., actual movie excerpts as opposed to McLaren’s experimental animations). The complexity and ecological validity of the audio-visual stimuli were increased significantly over the previous two experiments by utilizing actual motion picture excerpts.

Results

Group One Data Analysis and Interpretation. A repeated measures ANOVA was calculated using subject responses on each of the VAME ratings scales provided by subjects in Group One. As in Experiments One and Two, one between-groups variable (3 levels of musical training) and one within-groups variable (3 alignment conditions) were considered in the statistical analysis. Figure 6.3 provides a graphic representation of the mean subject responses for both VAME scales. Notice how the synchronization ratings and the effectiveness ratings diverge consistently in response to the out-of-phase alignment conditions. This is reflected in the substantial decrease in correlation between the two VAME scales ($r = .74$).

Synchronization ratings continued to exhibit the trend observed in Experiments One and Two, i.e., consonant alignments rated highest, dissonant alignments rated lowest, and out-of-phase alignments rated in between. However, a significant change occurred in the ratings of effectiveness. Two of the three out-of-phase composites (V1A1_O & V2A2_O) were rated higher than their related consonant composite (V1A1_C & V2A2_C)! The one instance in which the consonant alignment condition received the highest rating (V3A3) was the scene that was rated highest on a scale of synchronization in the initial exploratory study to the current experiment and, therefore, was perceived to exhibit the highest degree of synchronization between audio and visual accent structures.

Figure 6.3. Mean subject VAME responses to the AV composites in Experiment Three.



As in Experiment Two, the scores for each alignment condition were collapsed across the three AV composites prior to being submitted for statistical analysis. Neither the synchronization ratings nor the effectiveness ratings showed any significant effect of

musical training (synchronization— $F_{(2,17)} = .342, p < .715$; effectiveness— $F_{(2,17)} = 1.297, p < .2991$). Once again there were highly significant effects of alignment condition on both VAME scales (synchronization— $F_{(2,34)} = 12.243, p < .0001$; effectiveness— $F_{(2,34)} = 39.258, p < .0001$). The synchronization ratings showed no significant effect of the interaction between musical training and alignment condition ($F_{(4,34)} = .951, p < .4469$), while the effect of this interaction on the ratings of effectiveness was highly significant ($F_{(4,34)} = 4.582, p < .0046$). Therefore, instead of presenting a single line graph representing both VAME scores averaged across levels of musical training, Figure 6.4 provides the averaged synchronization ratings since all levels of musical training were considered to be equal on this VAME scale. In contrast, Figure 6.5 separates the mean effectiveness responses by level of musical training. The now-familiar descending trend from consonant to dissonant is immediately apparent in the ratings of synchronization, though the difference between the mean response for the consonant alignment condition and the mean response for the dissonant alignment condition is smaller than that observed in the previous two experiments. This suggests that, in observing these complex stimuli, alignment condition influenced subject synchronization ratings to a lesser degree than when considering the simpler AV composites in Experiments One and Two.

Figure 6.4. Mean synchronization ratings from subjects in Experiment Three, collapsed across alignment condition.

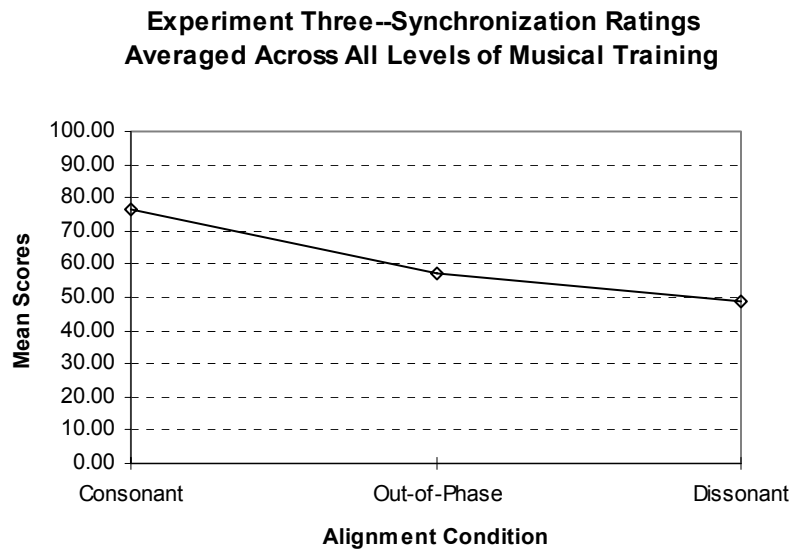
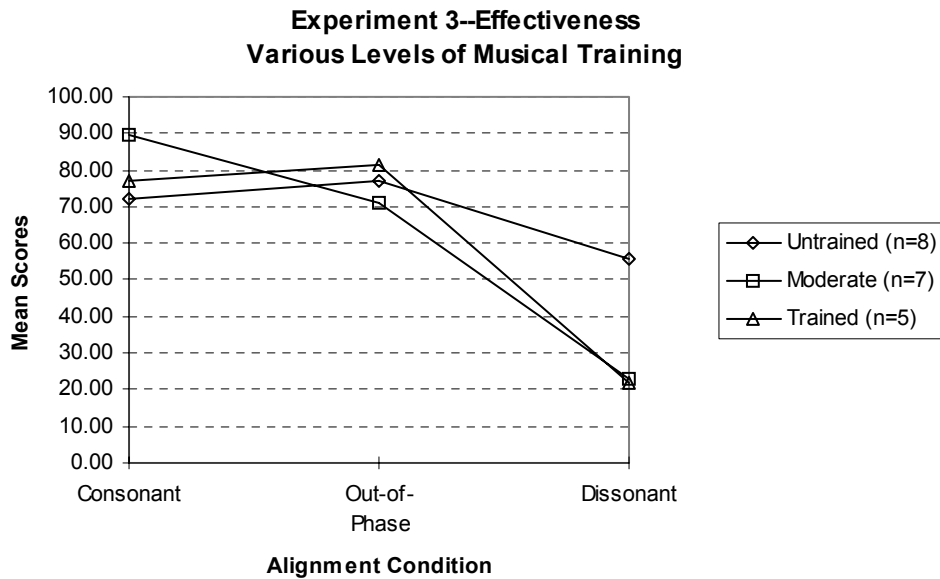


Figure 6.5. Mean effectiveness ratings for subjects of varying levels of musical training, collapsed across alignment condition (Experiment Three).



Recall that considering musical training as the grouping variable for the present series of studies was based on the theoretical assumption that musically-trained individu-

als would be more attentive to the audio track and, hence, better discern between the various alignment conditions. In the first two experiments, this prediction seems to have been incorrect, since there were no statistically-significant differences between the responses of individuals with less than two years of musical training, subjects with 2 to 7 years of musical training, and those with 8 or more years of formal musical instruction. The synchronization ratings for Experiment Three lend further supported this conclusion.

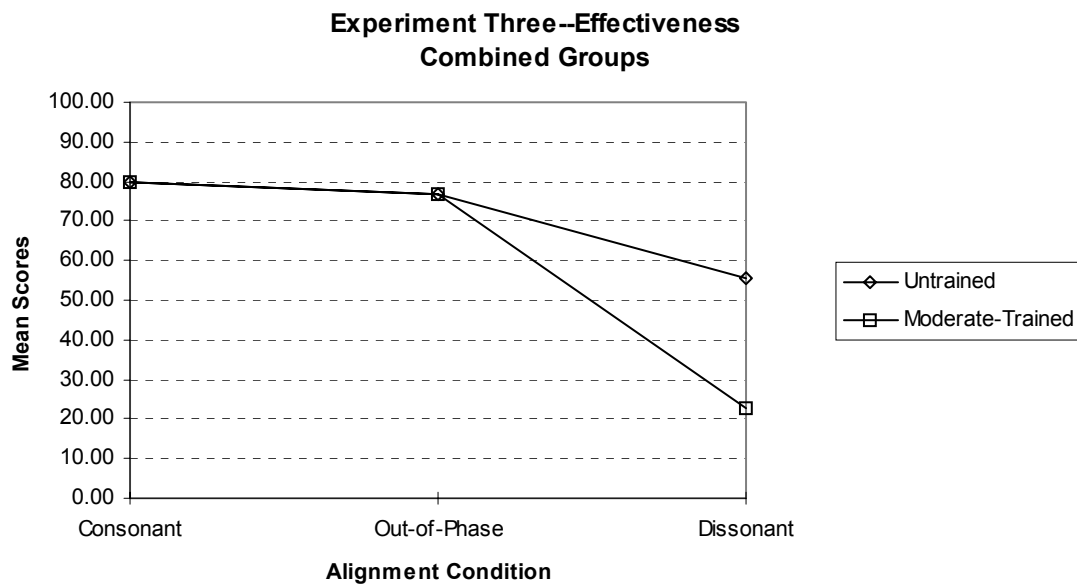
However, the ratings of effectiveness in this same experiment revealed for the first time a highly significant interaction between musical training and alignment condition. Therefore, a series of planned comparisons was run in an effort to determine specifically how ratings of effectiveness differed across alignment conditions as a function of musical training.³⁰

The first set of three comparisons simply determined whether the responses to each individual alignment condition were significantly different between the groups representing various levels of music training. Neither the consonant nor the out-of-phase ratings were shown to be significantly different on the basis of musical training (consonant— $F_{(2,17)} = 3.125, p < .0698$; out-of-phase— $F_{(2,17)} = .875, p < .4349$). However, there is a significant difference between responses to the dissonant alignment condition as a function of musical training ($F_{(2,17)} = 4.557, p < .0260$).

It is apparent from the mean ratings of the dissonant condition in Figure 6.5 that responses of individuals with a moderate amount of musical training did not vary significantly from responses of highly trained musicians. It is equally apparent that individuals with little or no musical training responded quite differently from the other two groups. The average between the moderate and highly trained groups was used as the best estimate of their combined mean score for comparison with the group of untrained individuals, providing the basis for the final planned comparison. A statistically significant dif-

ference was determined to exist between individuals with less than two years of musical training and those with two or more years of musical training ($F_{(1,17)} = 9.059, p < .0079$). This was the only statistically significant between-groups difference in the VAME procedure across the entire series of three experiments. Figure 6.6 appropriately represents the effectiveness ratings in Experiment Three across groups, i.e., all groups were identical in rating the consonant and out-of-phase alignment condition, but untrained individuals separate from those with moderate to high levels of musical training on the dissonant alignment condition.

Figure 6.6. Mean ratings of effectiveness from Experiment Three, combining groups as appropriate.



Group Two Data Analysis. A repeated measures ANOVA was performed on the responses provided by Group Two, considering a single between-groups variable (3 levels of musical training) and one within-groups variable (45 paired comparisons). There was no significant effect of either musical training ($F_{(2,17)} = 1.74, p < .206$) or the interaction between musical training and similarity ratings ($F_{(88,748)} = .60, p < .999$).

However, similarity ratings varied once again at a high level of significance ($F_{(44,748)} = 51.14, p < .0005$).

Multidimensional Scaling. The MDS solution for Experiment Three is provided in Figure 6.7. The 3-dimensions account for 99.99% of the variance in subject responses at a stress level of .00230. The three resulting dimensions, as labeled in the figure, are similar to those determined in Experiment Two. All composites utilizing V1 are located on the positive side of the “Visual” dimension, composites incorporating V3 are on the negative side, while V2 composites are near the zero-point. Composites utilizing A1 (except V2A1_D) are located on the positive side of the “Audio” dimension, composites with A2 are on the negative side, while A3 composites occupy the middle ground. Once again, the third dimension clusters the consonant and out-of-phase composites near the zero-point, while the dissonant combinations are found at the dimensional extremes. Notice how the consonant and out-of-phase alignment conditions for each AV combination lie almost on top of one another within the 3-dimensional space and that the dissonant combinations are located near the midpoint between the consonant and out-of-phase cluster representing their audio and visual components—e.g., V2A1_D is midway between the V1A1_C & V1A1_O cluster (audio component) and the V2A2_C & V2A2_O cluster (visual component) when considered across the Audio and Visual dimensions. The third dimension is labeled “Intent” rather than “AV Sync” in the figure, because a close examination of the triangular matrix of mean subject responses confirmed that, when comparing actual film excerpts, the subjects in the present study simply did not distinguish between the consonant and out-of-phase composites.

Table 6.5 provides the mean subject similarity judgment for each AV composite when comparing the consonant alignment to itself (identity) and the mean rating when comparing the consonant alignment to its related out-of-phase alignment (e.g., comparing

V1A1_C to V1A1_O). These values are illustrated graphically in Figure 6.8. The scale of this figure is left complete (i.e., 0 to 100) purposely to illustrate the closeness of these two similarity judgments when considered over the entire range of the possible responses. Subjects consistently considered both of these comparisons to be identical. Implications of this failure to discriminate between identities and consonant-to-out-of-phase comparisons when rating complex stimuli will be enumerated in the following chapter. As suggested in the discussion of the MDS solution for Experiment Two, the third dimension may actually separate AV composites on the basis of AV congruency, considering the associational (i.e., referential) meaning of the audio and visual components, as well as accent structure alignment.

Figure 6.7. MDS solution derived from mean similarity judgments in Experiment Three.

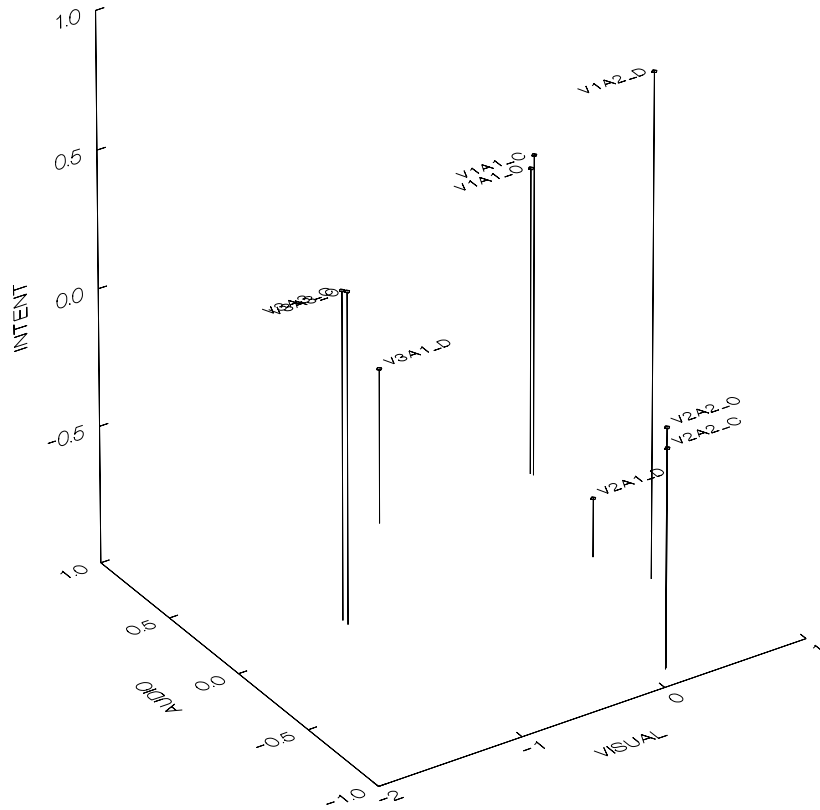
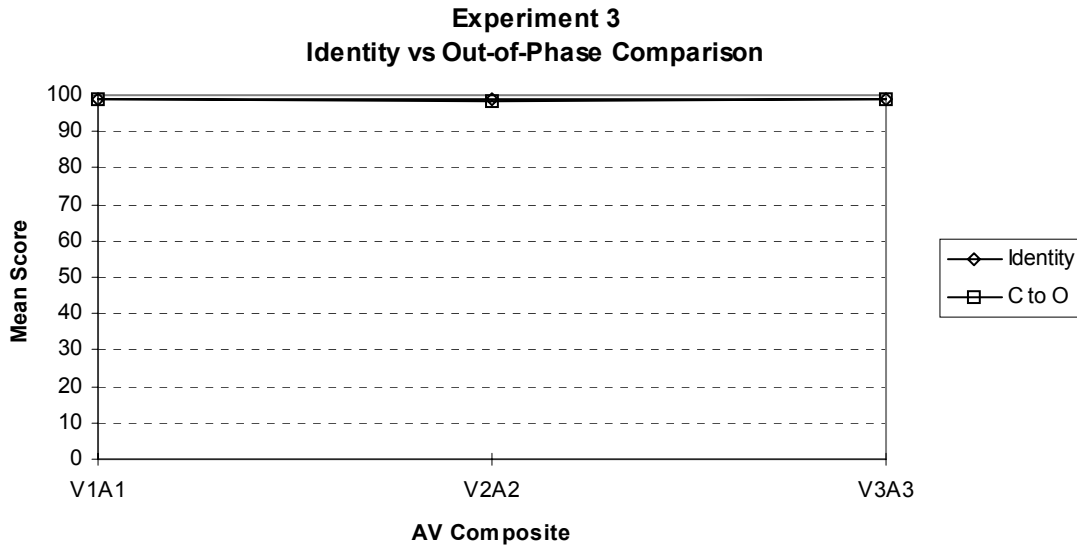


Table 6.5. Mean subject similarity judgments to identities (comparing a consonant alignment with itself) and the consonant-to-out-of-phase comparison for each of the three AV combinations.

Stimulus Comparison	AV Composite		
	<u>V1A1</u>	<u>V2A2</u>	<u>V3A3</u>
Consonant to Consonant (identity)	99.15	99.15	99.05
Consonant to Out-of-Phase	98.75	98.35	99.05

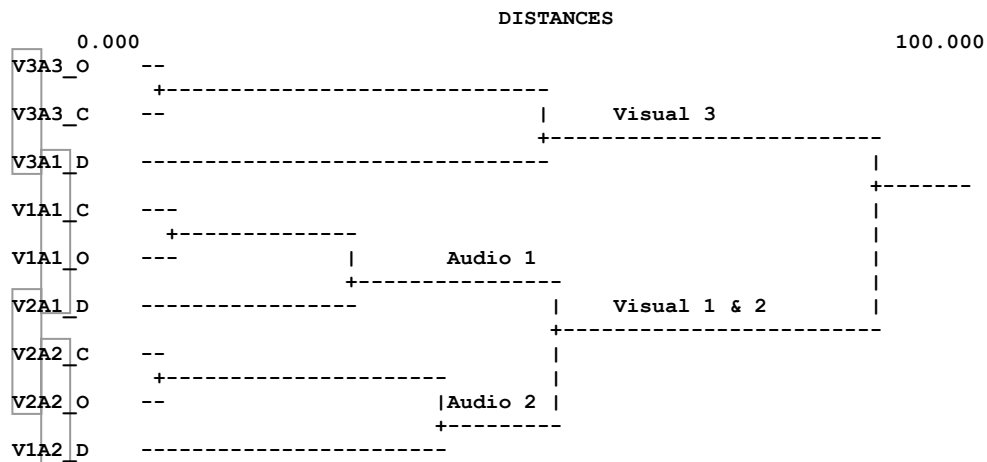
Figure 6.8. Line graph representing the mean responses in Table 6.5.



Cluster Analysis. The tree diagram presented in Figure 6.9 represents the results of a cluster analysis (complete linkage, furthest neighbor) for similarity judgments provided by Group Two in response to the AV composites in Experiment Three. The branching structure is so similar to that derived for Experiment Two that it must be stated explicitly that there is no relationship between the various audio (A1, A2, and A3) and visual (V1, V2, and V3) stimuli in Experiment Two and Experiment Three. The configurational similarity (i.e., the fact that the main branch divides V3 from V1 and V2, the latter of which is then separated according to audio component) is purely coincidental. The revelation gleaned from this particular tree diagram is the fact that similarity ratings in response to these actual film excerpts exhibited no apparent effect of alignment condition upon the resulting clusters. The branching structure can be accounted for *entirely* by audio and visual components, as illustrated by the overlapping groups enumerated below and highlighted by rectangles in Figure 6.9. Proceeding from top to bottom of the dia-

gram, V3A3_O, V3A3_C, and V3A1_D all share the same visual component (V3). V3A1_D, V1A1_C, V1A1_O, and V2A1_D share the same audio component (A1). V2A1_D, V2A2_C, V2A2_O share the same visual component (V2), while V2A2_C, V2A2_O, and V1A2_D all share the same audio component (A2). The neighboring composites between the main V3 branch and the V1 & V2 branch share the same audio component, while the neighboring composites between the A1 and A2 subbranches share the same visual component. Notice that, in no case, do two neighboring composites incorporate the same alignment condition. This rules out completely the possibility of a mirroring relationship as observed in the tree diagrams for both Experiment One and Two.

Figure 6.9. Cluster Analysis tree diagram—complete linkage (farthest neighbor)—for similarity ratings provided by Group Two subjects in Experiment Three.



Conclusions

Subject VAME responses to the movie excerpts in Experiment Three began to show a divergence between ratings of synchronization and effectiveness. Synchronization ratings continue to exhibit the response pattern observed in the two previous experiments (i.e., from highest to lowest: consonant, out-of-phase, then dissonant conditions), though the difference between consonant, out-of-phase, and dissonant alignment condi-

tions was much smaller than when considering simple or moderate AV stimuli. Ratings of effectiveness given in response to the consonant and out-of-phase alignment conditions, however, were essentially equal, suggesting that the difference between these two alignment conditions was less discernible. In addition, individuals with 2 or more years of musical training rated the dissonant alignment conditions of these movie excerpts significantly lower than those persons with no formal musical training. Therefore, the hypothesized relationship between musical training and alignment condition held only for these ratings of effectiveness given in response to dissonant combinations utilizing the most complex AV stimuli. In all other cases, level of musical training was not shown to be a significant factor in the subject VAME ratings.

The mean subject similarity ratings produced an MDS solution of three dimensions similar to those derived in Experiment Two. The first two dimensions were based on the visual and audio component of the AV composite, respectively. Cluster analysis revealed that alignment condition appeared to have no influence on the resulting tree diagram. As a result, the third dimension of the MDS solution is suggested to be a dimension of “Intent” (i.e., “appropriateness” of the AV combination, including both stylistic congruency and accent structure alignment), separating composer-intended combinations (V1A1, V2A2, & V3A3) from those AV combinations that were not intended (V1A2, V2A1, & V3A1).

Confirmation of Perceived Difference Between Consonant and Out-of-Phase Composites

The MDS and cluster analysis on subject similarity judgments seemed to suggest that the effect of alignment condition was negligible at best. We must once again reconsider whether the stimuli actually exhibited the characteristics of “alignment condition” expounded by the investigator, or if this third dimension was simply a measure of audiovisual congruency. One of the advantages of convergent methods (Kendall & Carterette,

1992a), in addition to providing multiple confirmations or disconfirmations of the hypotheses, is that it may provide a system of checks and balances in situations such as this one. If subjects were incapable of explicitly perceiving a difference between the various alignment conditions on a scale of synchronization, there would be no reason to assume that it would influence judgments in the similarity scaling task. However, if the subjects did distinguish between the consonant and out-of-phase composites—those sharing the same audio and visual component, but a different alignment condition (e.g., V1A1_C & V1A1_O)—we can be confident that a differentiation between the two stimuli existed based on a scale of synchronization.

Another series of same-sample pairwise *t*-tests was run to determine whether there was a perceived difference between the various alignment conditions. As shown in Table 6.6a, explicit ratings of synchronization discriminated between consonant and out-of-phase alignment conditions at a high level of significance ($t = 3.81, p < .001$). Therefore the quality alignment condition *was* being perceived. The difference between synchronization ratings of consonant and dissonant composites was also highly significant ($t = 4.92, p < .0005$). The ratings for out-of-phase and dissonant combinations, however, were not significant.

Turning now to the subject ratings of effectiveness (Table 6.6b), though subjects clearly distinguished between consonant and dissonant pairs and between out-of-phase and dissonant pairs, there was *no significant difference between ratings of consonant and out-of-phase composites*. Taken together, these results suggested that subjects were *capable* of discriminating between alignment conditions even when the audio and visual components were identical (i.e., compare the mean synchronization ratings of consonant composites to out-of-phase composites in Figure 6.4). However, when asked to rate the composites on a scale of effectiveness, the statistical difference between consonant and

out-of-phase composites disappeared (i.e., compare the mean effectiveness ratings of consonant composites to out-of-phase composites in Figure 6.5).

Table 6.6a. Paired *t*-test values for mean synchronization ratings in Experiment Three (graphically represented in Figure 6.4).

Comparison	<i>t</i>-value	df	2-tail significance
<i>Consonant to Out-of-phase</i>	3.81	19	.001
<i>Consonant to Dissonant</i>	4.92	19	<.0005
<i>Out-of-phase to Dissonant</i>	1.33	19	.198

Table 6.6b. Paired *t*-test values for mean effectiveness ratings in Experiment Three (graphically represented in Figure 6.5).

Comparison	<i>t</i>-value	df	2-tail significance
<i>Consonant to Out-of-phase</i>	.75	19	.463
<i>Consonant to Dissonant</i>	5.34	19	<.0005
<i>Out-of-phase to Dissonant</i>	5.83	19	<.0005

Interpretation of Data Analysis Across All Three Experiments

VAME Ratings

Collapsing subject VAME responses across alignment conditions allowed the entire data set across the three experiments (N = 60) to be analyzed as a single ANOVA, adding stimulus complexity as a second between-subjects variable. As in the independent analyses of each experiment, the resulting design consisted of two between-groups factors: 3 levels of stimulus complexity (simple, moderate, and complex—relating to Experiment One, Experiment Two, & Experiment Three, respectively) and 3 levels of

musical training (untrained, moderate, and highly trained). Alignment condition was the single within-groups factor (3 levels—consonant, out-of-phase, and dissonant). As mentioned previously, this was the second statistical analysis of the experimental data sets.³¹ In order to maintain an acceptably high level of confidence in the statistical results (i.e., greater than 95%) the alpha level was set *a priori* to .025. In reviewing the following results, recall that the number of Ns in each cell of the design is unequal. As in the data analysis for Experiment One, transformed F- and *p*- values (Wilks' *lambda*— F_{λ}) will be reported when appropriate, since the complete data set across all three experiments failed the likelihood-ratio test for compound symmetry. The transformed values do not assume compound symmetry. These factors should be considered in assessing the reliability of the following results. The total number of subjects falling into each category are represented in Table 6.7.

Table 6.7. Number of subjects in each cell of the between-subjects design across all three experiments.

<u>Experiment</u>	<u>Musical Training</u>		
	<i>Untrained</i>	<i>Moderate</i>	<i>Trained</i>
Experiment One	10	7	3
Experiment Two	8	7	5
Experiment Three	8	7	5

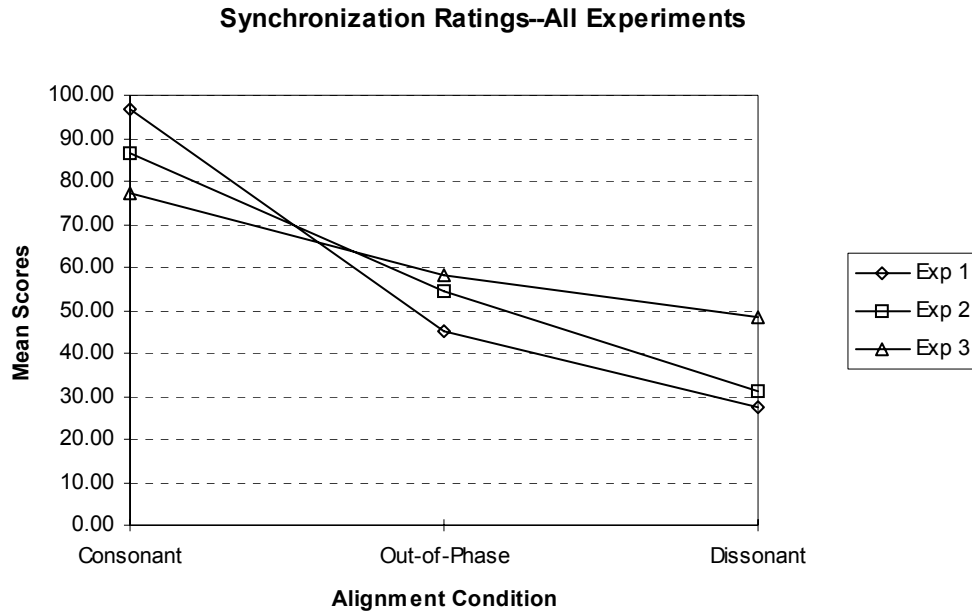
Synchronization. Considering subject responses to the VAME scale of synchronization, none of the following factors or interactions attained the level of significance: stimulus complexity ($F_{(2,51)} = .901, p < .4123$), musical training ($F_{(2,51)} = .629, p < .537$), nor the interaction between stimulus complexity and musical training ($F_{(4,51)} = .423, p < .791$), the interaction between musical training and alignment condition ($F_{\lambda(4,100)} = 1.241, p < .2987$), nor the interaction between complexity, musical training, and alignment condition ($F_{\lambda(8,100)} = 1.549; p < .1496$). However, both the alignment condition ($F_{\lambda(2,50)} =$

196.671; $p < .0001$) and the interaction between stimulus complexity and alignment condition ($F_{\lambda(4,100)} = 10.385$, $p < .0001$) were highly significant.

The consistent trend across alignment condition was that consonant alignments received the highest ratings, dissonant alignments received the lowest rating, and out-of-phase alignments received a rating in-between the other two, as clearly illustrated in Figure 6.10. Also immediately apparent is the cause of the “complexity by alignment condition” interaction. The simple stimuli in Experiment One received the highest mean rating for the consonant alignment condition followed by a drastic plunge to the lowest mean rating on the out-of-phase and dissonant alignment conditions. In contrast, subject mean responses to the stimuli in Experiment Three received the lowest rating on the consonant alignment condition, sloping mildly downward to points representing the highest rating on both the out-of-phase and dissonant alignment conditions. Response means for Experiment Two followed the same general trend, consistently maintaining a middle position between means representing the other two experiments. In addition, the mean responses for Experiment Three exhibited a much smaller range across the alignment conditions than those for Experiments One or Two (i.e., ratings of synchronization were more similar across alignment conditions when responding to the more complex motion picture excerpts).

In conclusion, it appears that the dissonant conditions were consistently rated slightly lower than the out-of-phase condition. However, the real difference occurred in the discrimination between the consonant and out-of-phase conditions. These distinctions are quite apparent when observing simple AV stimuli (Experiment One), but became increasingly less conspicuous as stimulus complexity is increased.

Figure 6.10. Mean synchronization ratings for each alignment condition across all three experiments.



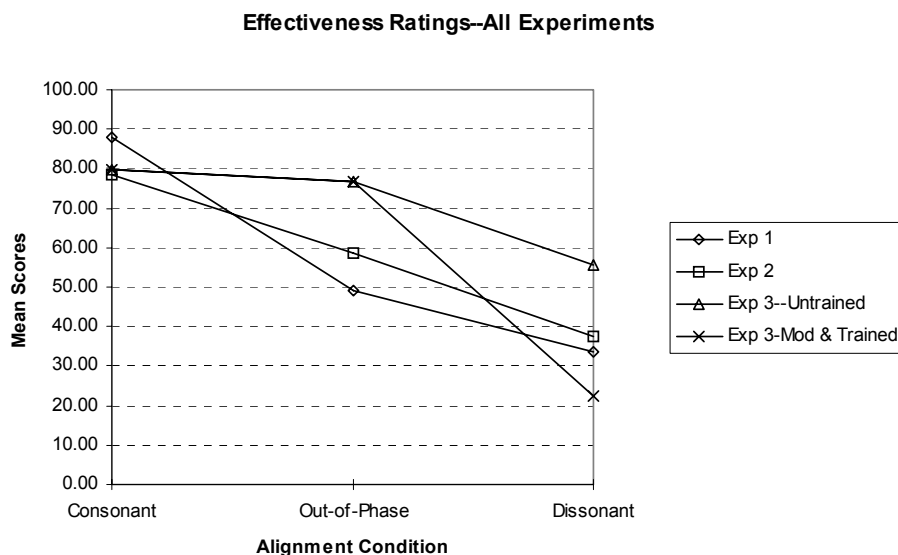
Effectiveness. Subject responses on the VAME scale of effectiveness revealed no significant difference in stimulus complexity ($F_{(2,51)} = 1.503, p < .2321$), musical training ($F_{(2,51)} = 1.495, p < .2340$), nor in the interaction between complexity and musical training ($F_{(4,51)} = .247, p < .910$). Significant differences were observed between alignment conditions ($F_{\lambda(2,50)} = 102.245, p < .0001$) and in the interactions between stimulus complexity and alignment condition ($F_{\lambda(4,100)} = 10.160, p < .0001$) and between musical training and alignment condition ($F_{\lambda(4,100)} = 3.819, p < .0062$).³²

Effectiveness ratings for Experiment One and Two were very closely related to their respective ratings of synchronization. The significant difference came in the mean subject responses to Experiment Three, as shown in Figure 6.11. There was very little difference between the mean rating for consonant and out-of-phase alignment conditions, further confirming that—when responding to real movie excerpts—subjects did not appear to be bothered by what the exploratory studies determined to be perceptible mis-

alignment of the musical score and visual images. Rather, this aspect of the AV composite appeared to be masked by the degree of appropriateness exhibited by the audio-visual pair. The effectiveness ratings collected in Experiment Three revealed that this appropriateness was related positively to the composer-intended AV combinations (whether consonant or out-of-phase) and negatively related to the unintended combinations (dissonant alignment condition).

Considering the responses across all three experiments, the significant interaction between musical training and alignment condition on the scale of effectiveness appears to be solely related to the large divergence between individuals with no musical training (untrained) and those with 2 or more years of formal music instruction (moderate and highly trained) in response to the dissonant alignment condition in Experiment Three. The latter group proved to be much more sensitive to audio-visual “dissonance.” Future investigations will be required to determine whether this sensitivity is based upon appropriateness of the AV pairing, accent structure alignment, or a combination of the two.

Figure 6.11. Mean effectiveness ratings for each alignment condition across all three experiments.



Similarity Ratings

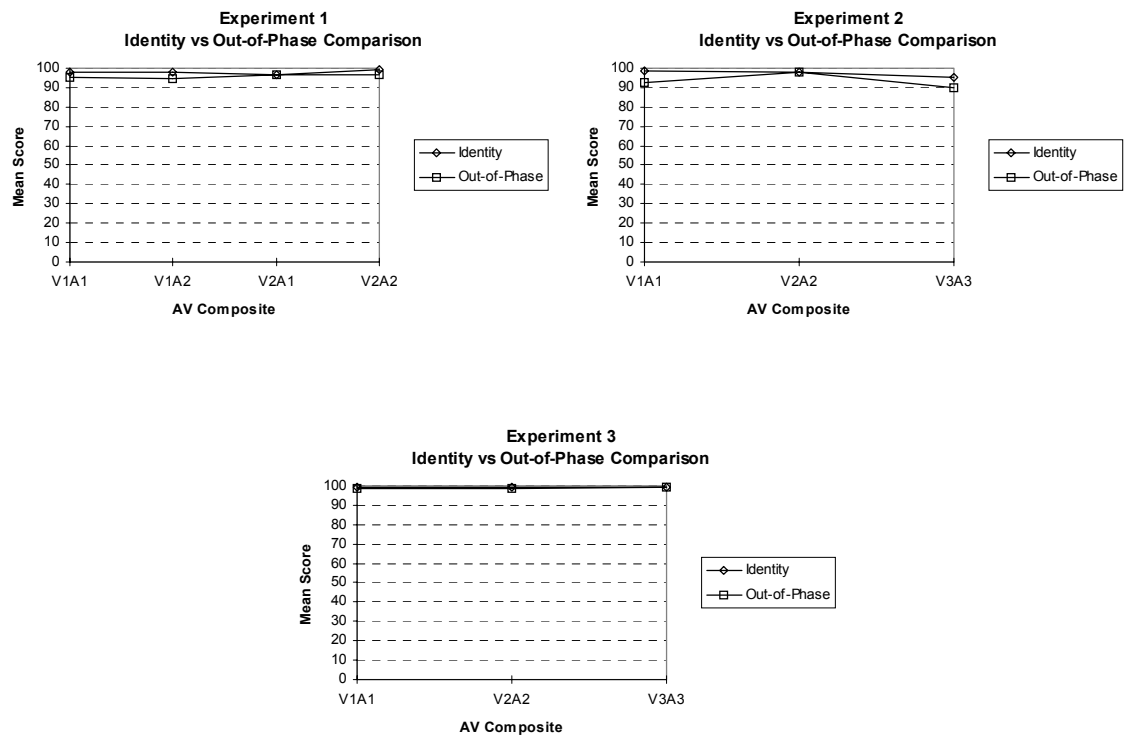
The MDS and cluster analyses for Experiment One clearly divided into three dimensions, including a dimension separating the dissonant alignment condition from the consonant and out-of-phase alignment conditions. In this case, the IOIs of the various stimulus patterns were manipulated by the investigator and combined in all possible AV combinations. Therefore, since the audio and visual components of a given AV combination were kept constant between the consonant, out-of-phase, and dissonant alignment conditions, it was possible to state without a doubt that the resulting third dimension (separating dissonant composites from the cluster of consonant and out-of-phase composites) was related to AV alignment. However, in the analysis of data from Experiments Two and Three, the fact that dissonant composites utilized a different audio track than their related consonant and out-of-phase composites removed this certainty. The decision to use identical musical scores for the consonant and out-of-phase alignment conditions—ensuring that these accent structures were identical, but offset—and a different musical score for the dissonant alignment condition forced the realization that, rather than a dimension of AV synchronization, this could simply be a dimension separating intended (i.e., appropriate) combinations from unintended (i.e., inappropriate) ones.

The subjects' raw data from all Three Experiments revealed a startling conclusion. In Experiments One and Two, subjects were able to discriminate between most consonant and out-of-phase alignment conditions. However, in Experiment Three, the mean subject responses to identities (i.e., comparisons of a consonant alignment conditions with itself—the same visual component, the same audio component, *and* the same alignment condition) and “consonant to out-of-phase” comparisons (i.e., comparing a consonant alignment condition to its related out-of-phase alignment condition—same visual component, same audio component, but *different* alignment conditions) clearly

showed that subjects simply were not distinguishing between the consonant and out-of-phase alignment conditions sharing the same audio and visual components.

Figures 6.12a to 6.12c provide graphic representations of these mean similarity ratings in response to identities (i.e., consonant to consonant comparisons) and out-of-phase comparisons (i.e., consonant to out-of-phase) for all three experiments. Once again, the entire scale (0 to 100) was used in the graphs to place mean ratings within a context of the entire range of possible responses. Even in Experiments One and Two (Figures 6.12a and 6.12b), the identities and the “consonant to out-of-phase comparisons” are always shown to be closely related, because the composites being rated didi share the same audio and visual component, regardless of alignment condition. However, in responses to Experiment Three (Figure 6.11c), the distinction between these two judgments completely disappeared—in both cases, the pairs of stimuli were perceived to be identical. This result, in addition to the previously-discussed cluster analysis, caused the investigator to conclude that similarity judgments to the actual motion picture excerpts showed no influence of alignment condition whatsoever, though the post-hoc analyses on VAME ratings for this experiment confirmed that subjects were *capable* of discriminating these same stimuli on the basis of alignment condition. For this reason, the third dimension of the MDS solution must be related to appropriateness of AV combination (stylistic and/or associational congruency), rather than accent structure alignment.

Figure 6.12. Relationship of mean similarity judgments for comparisons of identical stimulus pairs (i.e., consonant to consonant) and consonant-to-out-of-phase stimulus pairs in a) Experiment One, b) Experiment Two, and c) Experiment Three.



CHAPTER SEVEN

DISCUSSION, SUMMARY, AND CONCLUSIONS

Discussion

The following discussion will relate results of the three experiments described in the foregoing chapters to previous research. The first research question addressed identification of potential sources of accent. A review of relevant literature provided a list of potential sources in both the auditory and visual modalities. Several were selected by the author and used as a basis for creating animations and isochronous pitch sequences in order to produce periodic accents in listener/observer perception for use as stimuli in Experiment One.

Of all the potential sources of musical accent incorporated into an exploratory tapping procedure, four patterns proved extremely reliable in producing subject taps at the hypothesized rate³³ (see Figure 4.1): pitch contour direction change only (Pattern #1), change in interval size combined with pitch contour direction change (Pattern #2), dynamic accent combined with pitch contour direction change (Pattern #3) and timbre change combined with contour direction change (Pattern #5). In a similar procedure, the most reliable visual stimuli were determined to be (see Figure 4.2) three types of translation in the plane—left-bottom-right-top (Pattern #2), top-to-bottom (Pattern #3), & side-to-side (Pattern #4)—and translation in depth (motion from back-to-front along an apparent z-axis (Pattern #5). Results of the first exploratory study, therefore, revealed that

changes in audio or visual characteristics (or motion vectors) identified in the literature did, in most cases, result in reliable subject perception of points of accent.

Findings of the three main experiments confirmed that individuals were capable of discriminating between the three alignment conditions on a verbal scale of synchronization. This result remained constant, regardless of stimulus complexity and subjects' level of musical training. Likewise, at lower levels of stimulus complexity (Experiments One and Two), subject ratings of effectiveness also distinguished the three levels of alignment.

However, results of Experiment Three revealed that, at the highest level of complexity (i.e., actual movie excerpts), post-hoc analysis confirmed that mean ratings of effectiveness failed to discriminate between consonant and out-of-phase alignment conditions. A similar weakening in the influence of accent alignment on subject similarity ratings (Group Two) was observed both in Experiment Two and Experiment Three. The method used to incorporate ecologically valid AV stimuli (i.e., real movie excerpts with their accompanying musical score) into the experimental design made it impossible ultimately to distinguish whether the resulting relationships were caused by the appropriateness (i.e., congruence) of a specific AV combination, accent structure alignment, or a combination of the two.

By definition, both the consonant and out-of-phase composites had to employ the same accent structure. In the latter case, either the audio or visual component was offset temporally, resulting in accent structure misalignment. In order to ensure that consonant and out-of-phase conditions did, in fact, share the same accent structure, both were combined with the same (intended) soundtrack. In contrast, the dissonant alignment condition, by definition, had to combine different auditory and visual accent structures, so the dissonant combination always used a different (unintended) soundtrack than the conso-

nant and out-of-phase alignment conditions. Recommendations for remedying this confound are made below, in proposing future investigations.

The second result of the present study was the change in both VAME ratings as a function of stimulus complexity. Consonant alignment conditions were rated extremely high in response to the simple animations created by the author, while dissonant pairs were rated extremely low. In contrast, the responses to Experiment Two (i.e., moderately complex stimuli) revealed a slightly lower mean rating in response to consonant AV composites and a slightly higher mean rating of dissonant combinations.

This trend continued in Experiment Three, in which mean ratings of consonant AV combinations were even lower, while mean ratings of dissonant composites increased again. The most convincing result in support of decreasing salience of accent structure alignment was the post-hoc comparison presented in the data analysis portion of Experiment Three. Effectiveness ratings for the consonant and out-of-phase alignment conditions were not significantly different from one another. The hypothesized affect of accent structure alignment on effectiveness ratings disappeared when reaching a level of complexity equivalent to that of a typical Hollywood film. This failure to discriminate between consonant and out-of-phase alignment conditions on ratings of effectiveness occurred even though the same subjects—responding in the same experimental context—rated these consonant and out-of-phase composites significantly different on a scale of synchronization (i.e., explicitly addressing the stratification of accent structures). Essentially, this illustrated that, though the subjects *were capable* of discriminating between these two alignment conditions (i.e., the ratings of synchronization), they did not appear to consider consonant alignment a necessary criterion in their ratings of effectiveness.

There was also no significant difference between synchronization ratings taken from individuals with varying degrees of musical training at any level of stimulus com-

plexity (i.e., in any of the three experiments). But, although effectiveness ratings in Experiments One and Two exhibited no significant difference between levels of musical training, individuals with two or more years of formal instruction appeared to be more sensitive to dissonant combinations of music and image in Experiment Three. It was not possible, due to experimental design, to determine if this sensitivity was due to accent structure alignment or a judgment of AV congruency as proposed by Marshall & Cohen (1988).

In the Film Music Perception Paradigm (Lipscomb & Kendall, in press), interaction between stimulus complexity and perceived effectiveness, as well as the influence of musical training, were not accounted for. The following modification of the earlier model proposes a weighted relationship between the factors of accent structure alignment and association judgment, both of which are subcomponents of an individual's overall determination of AV congruency. In this model, the importance of accent alignment is considered appropriately high at low levels of stimulus complexity, but decreases in salience as AV complexity increases. The dynamic aspect of these two interrelated components is illustrated in Figure 7.1 by the dotted boxes. This implies that the salience of each of these factors changes in association with the type of audio-visual experience. Determination of specific weightings demand further research.

The attentional aspect of the modified model has not yet been explicated. Due to limitations of our perceptual system, it is impossible to pay attention to everything going on in our immediate environment. A necessary part of the cognitive process is data reduction (i.e., selecting elements around us that warrant attention).

Figure 7.1. Revised version of the Film Music Perception Paradigm.

Jones and Yee (1993) elaborated on the varied definitions of attention that have been suggested within the field of psychology, lamenting the fact that there is no general agreement on a precise definition. Jones and Yee suggested that

this state of affairs exists because attention itself is an inferred construct.

Any meaningful definition of it quickly becomes ‘theory bound,’ virtually compelling a commitment to ‘what’ attention is. Ultimately, definitions of attention become theories of attention (1993, p. 70).

However, Matlin (1994) provided a general definition that suffices for the present model. She defined attention as simply “a concentration of mental activity” (p. 43).

The modified Film Music Perception Paradigm suggests that, if the associational (i.e., referential) aspect of the music is congruent with the visual image and if (for simple to moderately complex AV combinations) the accent structures are aligned, then attention will remain on the composite as a whole. It appears that accent structure alignment plays little role in the effectiveness of combining audio and visual components in contexts representing high complexity, i.e., the “accent structure relationship” portion of the model collapses. However, if the audio and visual components are of a lower complexity, then misalignment of the accent strata may produce a problem-solving strategy to determine the source of the asynchrony. The human mind does, of necessity, have a tolerance for slight misalignment. Perhaps this is due to the fact that stimuli reaching our eyes and ears travel at different velocities, e.g., if we observe someone at a distance of 500 ft. bouncing a ball, we see the ball bounce before we hear the associated bouncing sound. In the context of an AV composite, the amount of temporal misalignment by which the audio and visual must be out of sync in order to be considered out-of-phase is undefined and requires further research.

Progressing from Experiment One to Experiment Three represented change in another factor, as well. The stimuli in the first experiment (i.e., a circle moving around the screen on a black background) had very little inherent meaning. Therefore, the relationship between this figure and the isochronous pitch sequences could be described as syntactical, rather than semantic. In this case, the “meaning” of the composite was determined primarily by the relationship of one musical tone to the next, the motion pattern of the ball, and the interrelationship between visual motion and musical sound. This relationship was illustrative of a symbolic relationship, i.e., one of the classifications proposed by Pierce (1931-35). As the stimuli progressed toward actual motion picture excerpts, an aspect of referentialism (i.e., semantic meaning) was introduced into the AV relationship. In these visual scenes, there were human characters interacting within the context of a storyline. This shift has been referred to thus far in the present study as “simple-to-complex” on the basis that there are more events per unit time than in the simple animations. However, this same progression may just as well be described as a move from a symbolic relationship between the audio and visual components to indexical or iconic (i.e., from syntactic to semantic).

By developing a method of quantification for determining the referential (or semantic) content of a visual scene and the referential content of a musical passage, it would be possible to supply two of the three values necessary to place an AV combination within the 3-dimensional space of Lipscomb & Kendall’s model of film classification (see Figure 2.3). The third dimension of this model should be labeled AV Congruence instead of AV Synchronization. This judgment incorporates both consideration of associational aspects and accent structure alignment, as well as their dynamic interrelationship. Since subject ratings of effectiveness seem to incorporate both of these elements, this verbal metric may provide a means of quantifying the third dimension in the

model of film classification. Both of the modified models presented above will require future research in order to determine their reliability and validity.

Limitations

In all experimental research, the ability to generalize results is limited by the number and type of stimuli employed. 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.

It was also necessary to keep the excerpts brief so that subjects could either respond to a single AV composite on two verbal scales or compare two AV composites on a scale of similarity without suffering response degradation due to memory retention errors. Therefore, stimuli for the three experiments were kept to lengths of between 5 and 20 seconds. Even the most complex of these brief excerpts was not completely representative of real world motion picture experience due to its brevity.

Another limitation of the present study is that all animation excerpts employed in Experiment Two were taken from the work of a single animator and all of the excerpts used in Experiment Three were taken from a single movie as an attempt to control for variation in production technique, composer style, direction, etc. Also, subjects observing these stimulus pairs experienced them as *only* visual images and musical sound (i.e., there were no sound effects, dialogue, or other ambient sound) in order to eliminate unwanted auditory cues. Though these elements of control were important to the internal design of the experiment, the ability to generalize results outside the context of the present investigation was restricted.

It is also likely that the subjects used in this study may not have been representative of the population of university students from 18 to 31 years of age. All subjects were volunteers taking general elective courses in UCLA's Department of Ethnomusicology and Systematic Musicology. Therefore, these individuals (even those classified as having two years or less of musical training) probably had an interest in music elevated over the general nonmusician population.

Several weaknesses in the present study must be addressed, so that future investigations may improve upon techniques and procedures described herein. First, there was an unequal number of subjects in the various between-subject cells of the research design (i.e., the number untrained, moderate, and highly trained musicians). Though the amount of difference was not too large in either Experiment Two or Three, the highly trained musicians cell in Experiment One contained only three subjects. Ideally, these cells should be equal in order to increase statistical reliability. Second, between the three experiments, length of the AV excerpts increased as well as stimulus complexity. If stimulus length had been controlled within the design of the present experiment, there would be greater confidence in knowing that the manipulated independent variables (i.e., audio component, visual component, and alignment condition) caused the observed change in the dependent variables (i.e., verbal responses or similarity judgments).

The most obvious limitation of the present series of experiments was due to the utilization of ecologically valid stimulus materials in the investigation for the purpose of creating three operationally-defined alignment conditions. In Experiments Two and Three, because the consonant and out-of-phase alignment conditions shared the same (intended) music and the dissonant condition had a different (unintended) musical score, two variables (alignment condition and composer intent) were confounded. Future investigations are proposed below, offering specific suggestions on how to separate these two

aspects of the musical sound and still retain at least a significant amount of the desired ecological validity.

Summary

The purpose of the present study was to investigate the role of audio-visual (AV) alignment in motion pictures and animation. Film music was considered to be one of the most effective and prominent uses of music for manipulating (i.e., augmenting) a person's response to visual images. Several research questions were posed. First, what determines an accent in both the visual or aural sensory modalities? Second, in order for an AV combination to be considered effective, is it *necessary* that accents in the musical sound stream align precisely with salient moments in the visual scene?

A review of related literature, determined that there has been relatively little past research utilizing AV materials that have a high degree of ecological validity (i.e., are representative of real world experience). Perceptual psychologists have shown an interaction between the auditory and visual modalities when using sine tones and light flashes (e.g., Radeau & Bertelson, 1974; Staal & Donderi, 1983). However, because of the reductionism inherent in these stimulus materials, it is impossible to generalize the results to a complex AV experience like motion pictures. Three past studies of music accompanying dramatic action (film, video, or theater) were discussed at length (Tannenbaum, 1956; Thayer & Levenson, 1984; Marshall & Cohen, 1988). Consistent weaknesses in these previous studies were identified. There was a general lack of attention to *the way in which music was added* to a visual experience. Secondly, the musical stimuli did not exhibit the craftsmanship typical of a motion picture score. Both of these weaknesses have been corrected in the current series of experiments.

The Film Music Perception Paradigm proposed by Lipscomb & Kendall (in press) proposed two implicit decisions made by a motion picture observer. “Association judgment” concerns the perceived appropriateness of a particular combination of music and image, while “accent structure relationship” suggests a comparison between accents in the music and those moments considered particularly salient in the visual image. The concepts of musical and visual periodicity were then explained, referencing scenes from extant films as examples.

An underlying assumption of this model—and, hence, the present study—is music’s ability to communicate. Two theories of meaning were briefly presented (Pierce, 1931-35; Meyer, 1956) with appropriate references to several motion pictures. Based on these theories, a second model of film music was proposed; this time, based on the degree of musical referentiality, visual referentiality (i.e., representationalism), and AV synchronization. A hypothetical example placed an extant film, an experimental animation, and a cartoon within this 3-dimensional space for illustration purposes.

Various investigators and theorists have proposed—stating the obvious, perhaps—that the sensation of accent occurs when a change is perceived in an incoming stream of sensory information. In order to address any of the research questions restated above or to explicate either model of film music further, it was necessary to determine what aspects of musical sound and/or visual images were capable of producing a perceived accent. A review of music perception literature revealed the following potential musical parameters: contour (i.e., interval size and melodic direction), a note’s position in the tonal system, pitch height, serial position, articulation, loudness, timbre, duration, rhythm, meter, and delayed onset. A similar review of visual perception literature revealed the following potential sources of visual accent: shape, size, orientation, location, color, lighting, pattern, texture, and motion (in any of several specific forms).

The term *vector* was proposed as a possible basis for a “common language” of both musical sound and visual images, since both perceptions are perceived only within the inexorable flow of time. As such, these aspects of both vision and audition may be described with a specific direction and magnitude as a function of time. Once a vector is established, expectations are generated for continuation. Proceeding from Meyer’s (1956) theory of emotion in music, it was proposed that accents occur at points of change due to a blocking (or inhibition) of this continuation in any of the musical or visual parameters listed above. From the list of potential sources of accent, a subset was selected and used as a theoretical basis for creating stimuli for the first exploratory study.

Previous film music research has focused almost exclusively on the referential and associational aspects of music when combined with visual images. No previous investigation has addressed specifically the relationship of accent structures between the two domains. The author proposed that, in the motion picture experience, there exists a stratification (i.e., layering) of accent structures, such that the musical accent structure and the visual accent structure share a specific relationship (or alignment condition). Three alignment conditions (consonant, out-of-phase, and dissonant) were operationally defined for use in a series of experiments.

In order to determine whether accent structure alignment had any affect on subject perception of an AV composite, two experimental tasks were utilized: verbal attribute magnitude estimation (VAME) and similarity scaling. The VAME task required subject responses on two verbal scales: “not synchronized–synchronized” and “ineffective–effective.” Subjects responded on both VAME scales to a series of AV composites in which the visual component, musical component, and alignment condition were manipulated by the investigator as independent variables. Subjects in the similarity scaling task were asked to rate all possible pairs of AV stimuli on a scale of “not same–same.” This method of converging methods (i.e., using both verbal scales and a

This method of converging methods (i.e., using both verbal scales and a similarity judgment) allowed the investigator to answer the research question in two different ways, comparing results of the two methods for either confirmation or disconfirmation.

Experiment One

Exploratory Studies. Seven isochronous pitch sequences and seven animations were created by the author, using the musical and visual parameters determined from the literature review. Three versions of each pitch sequence and each animation were created, such that they exhibited the following inter-onset intervals (IOIs) between accent points: 500ms, 800ms, or 1000ms. An exploratory study utilizing a subject tapping procedure was carried out in order to determine which examples produced the most reliable sense of accent periodicity.³⁴ Immediately apparent was the fact that some subjects tapped at an IOI either subdividing or nesting the hypothesized accent periodicity. These tap rates were equalized by mathematically transforming all mean responses to the range representing the responses given by a majority of subjects.³⁵ Subject responses confirmed that the hypothesized IOIs (or their relative nesting or subdividing durations) were perceived by subjects, as predicted. The two most reliable pitch sequences and the two most reliable animations were selected for use in the first main experiment. In Experiment One, subjects viewed the four AV composites (2 animations x 2 pitch sequences) in all possible alignment combinations—5 AV accent structure relationships (consonant, nested consonant, out-of-phase, out-of-phase [nested], and dissonant) x 3 IOIs.³⁶

VAME Procedure. Verbal responses to these combined stimuli revealed no significant response difference between subjects with varying levels of musical training. However, there was a highly significant difference between ratings to the various alignment conditions. In general, subject ratings on both the synchronization and effective-

ness scales followed a consistent pattern of responses from highest to lowest: identical consonant alignments, nested consonant alignments, out-of-phase (identical) alignments, followed by either out-of-phase (nested) or dissonant alignments. Analysis of variance also revealed a statistically significant interaction between AV composite and alignment condition, ascribed mainly to the spread of mean subject ratings in response to the out-of-phase consonant alignment condition across the four different AV combinations.

Because of this variation in subject responses across the nested pattern, the data set was reduced to only three alignment conditions (consonant, out-of-phase, and dissonant), by eliminating both the nested consonant and out-of-phase (nested) conditions. Plotting mean responses for this reduced data set clearly established the response pattern that would be observed in all mean subject synchronization ratings across the series of experiments described herein, i.e., consonant alignments are rated highest, dissonant alignments are rated lowest, and out-of-phase alignments receive a rating in between the other two. The consistency of this pattern across various AV combinations was statistically confirmed. Effectiveness ratings follow the same pattern in subject responses collected in Experiment One.

In order to increase reliability of the response measure on the concept of interest (i.e., AV alignment), subject responses were collapsed across alignment condition. In Experiment One, this provided 12 separate measures of each alignment condition (i.e., 4 AV composites x 3 IOIs). An ANOVA on the collapsed data set revealed no significant difference between levels of musical training or in the interaction between musical training and alignment condition on either VAME scale. However, the difference between ratings to the various alignment conditions was highly significant for both synchronization and effectiveness. Therefore, it was concluded that the subjects did, in fact, clearly

distinguish between the three alignment conditions on both VAME scales, regardless of level of musical training.

A noticeable trend was observed in the relationship between synchronization and effectiveness ratings. Though the correlation between the two was extremely high in Experiment One ($r = .96$), the effectiveness ratings were always less extreme than the synchronization ratings. For example, when synchronization ratings were high (e.g., consonant alignment conditions), the effectiveness rating tended to be slightly lower. In contrast, when synchronization ratings were low (out-of-phase or dissonant alignment conditions), effectiveness rating tended to be slightly higher. The author suggests that, while synchronization ratings mapped precisely onto the varying alignment condition, effectiveness ratings were likely to be influenced by other factors inherent in the AV composite, such as associational congruency.

Similarity Judgment Procedure. Similarity judgments were collected in response to a subset of the in Experiment One stimulus set. Statistical analysis revealed no significant difference on the basis of musical training and no interaction between musical training and similarity ratings. The difference between similarity ratings to the various AV composites, however, were determined to be highly significant. Multidimensional scaling of the mean response matrix resulted in three clearly interpretable dimensions: audio component, visual component, and alignment condition. These dimensions were confirmed by cluster analysis.

In response to simple stimuli (isochronous pitch sequences and simple single-object animations), we may conclude that subjects clearly distinguished between various alignment conditions (i.e., consonant, out-of-phase, and dissonant) on both VAME scales. In addition, similarity judgments to all possible pairs of AV composites revealed

that subject responses were clearly based on three factors, one of which was alignment condition of the audio and visual components.

Experiment Two

Exploratory Studies. Three excerpts were selected from the experimental animations of Norman McLaren for use in the second experiment. These works were considered appropriate for the present investigation because of the attention that McLaren devoted to audio-visual alignment. Visual stimuli were played from a laserdisc on the computer monitor and synchronized with a digitally-sampled sound file produced by an internal sound card. This configuration allowed the sound file to be edited for the purpose of moving the musical events forward and backward in time in relation to the visual images. An exploratory study was conducted using both the audio tracks and visual images in isolation to confirm perceived accent periodicity. The audio procedure was extremely successful at uncovering underlying accent structure in the music track. However, the visual procedure was highly variant between subjects since some tended to tap along with important events as they occurred, rather than establishing a steady pulse. Therefore, the IOIs determined in the audio exploratory studies were used as a basis for creating a variety of alignment conditions. Eleven different versions of each AV combination were created by offsetting the audio and visual relationship by a series of incremental multiples related to the perceived IOI.

A second exploratory study was then run in order to determine two things: 1) the AV alignment using the intended audio and visual that was perceived as the worst possible synchronization and 2) the AV combination using an audio track other than that intended by McLaren that resulted in the worst possible AV synchronization. The former combination was utilized as the out-of-phase alignment condition for use in Experiment Two and the latter was selected as the dissonant alignment condition. McLaren's original

version was used, in every case, as the consonant alignment condition. Therefore, each of the three animation excerpts was presented in 3 different alignment conditions: consonant, out-of-phase, and dissonant. Experimental procedures for Experiment Two were identical to Experiment One, varying only in degree of complexity of the stimulus materials.

VAME Procedure. Once again, there was no effect of musical training on subject responses and no interaction between alignment condition and musical training. The difference in ratings between the three alignment conditions was highly significant. As in Experiment One, subjects clearly rated the consonant alignment conditions higher on both VAME scales than the out-of-phase and dissonant conditions, rating the latter lowest. Mean ratings of effectiveness are still less extreme—as in Experiment One—than the associated rating of synchronization, though the correlation between effectiveness and synchronization ratings remained very high ($r = .94$). The range between the mean consonant rating and the mean dissonant rating for both VAME scales was not as large as that observed in Experiment One. Also, the out-of-phase condition was not rated as low on either VAME scale as in the responses to those simple animations. It appeared, therefore, that subjects continued to distinguish clearly between all three alignment conditions on both VAME scales, though the consonant relationships were not rated as high—nor are the dissonant composites rated as low—as in the previous experiment.

Similarity Judgment Procedure. Subject responses on a scale of similarity did not differ significantly on the basis of musical training or in the interaction between musical training and similarity ratings. There was, however, a high degree of significance in the difference between ratings of similarity. The MDS solution of these subject responses revealed a separation into three dimensions that are once again identified as visual component, audio component, and alignment condition, though the latter dimension was

unique in the way that consonant and out-of-phase alignments were drawn toward the zero-point, while dissonant combinations were placed at the extremes (both positive and negative). Cluster analysis confirmed that the effect of the audio and visual components remained strong, but the influence of alignment condition appeared weakened. The author proposes, therefore, that the third dimension may be a dimension based on “appropriateness” of the AV combination, *including* accent structure alignment.

In general, results of Experiment Two continued to substantiate the role of AV synchronization in subject VAME responses to moderately complex stimuli. The high-to-low relationship between alignment conditions (i.e., consonant to out-of-phase to dissonant) was clearly evident in the subject responses on both VAME scales. Even in these responses, however, the degree of differentiation was lessened. A series of post-hoc paired samples *t*-tests confirmed that subject synchronization ratings for consonant and out-of-phase alignment conditions—composites sharing the same audio and visual component—were, in fact, significantly different. Subjects in the similarity judgment task appeared to have used the alignment cues less consistently, making their determination based on audio and visual component and the appropriateness of the AV pairing. The author insists that, if this third dimension does represent a judgment of AV appropriateness, it must include the relationship of accent structures in addition to any associational or referential criteria. This claim is supported by the significant differences discovered in the post-hoc analysis.

Experiment Three

Exploratory Studies. Stimuli for the third experiment were selected from the movie “Obsession” with a musical score composed by Bernard Herrmann. The author added programming code to MEDS, allowing access to the various audio tracks of the laserdisc, since the score was isolated on one of the analog tracks. The same tapping

procedure was used with both the audio tracks and visual images in order to determine accent structure. Since this study used actual movie excerpts (ecologically valid, but highly complex), this exploratory procedure was much more difficult than previous tapping procedures. Therefore, graduate music students were used instead of students from the general undergraduate population. The audio exploratory study resulted in highly reliable IOIs, while subjects responding to the visual task once again simply identified significant points in the visual image by tapping the spacebar, rather than providing a consistent underlying pulse. The various alignment conditions were created using a system identical to that described above in the discussion of Experiment Two.

VAME Procedure. With the VAME responses to Experiment Three, we began to see a divergence between responses of synchronization and effectiveness. The correlation coefficient dropped substantially ($r = .74$), because subjects rated out-of-phase and consonant alignment conditions at essentially the same level of effectiveness. An ANOVA revealed no significant difference in level of musical training for either VAME scale. There was, once again, a highly significant difference between responses to the various alignment conditions. Finally, though the analysis revealed no significant difference in the interaction between musical training and alignment condition for ratings of synchronization, there was a highly significant difference in this same interaction for the effectiveness ratings.

Three planned comparisons were run testing the relationship between levels of musical training and effectiveness ratings to the various alignment conditions. These analyses revealed that the significant difference was due solely to the manner in which individuals with less than two years of musical training (untrained) differed from those with 2 or more years of training (moderate and highly trained). This was the only statis-

tically significant difference related to musical training discovered across the entire set of three experiments.

Similarity Judgment Procedure. Subject similarity responses to the complex stimuli in Experiment Three revealed no significant effect of musical training, nor in the interaction between musical training and similarity rating. However, differences between the similarity ratings of the various pairs of stimuli were highly significant. The MDS solution for these responses once again clearly separated the composites according to audio component and visual component. However, the effect of alignment condition appeared disappear from these similarity judgments completely. Instead, subjects were judging the “consonant to out-of-phase” comparisons (i.e., comparing a consonant alignment condition to the composite with the same audio & visual components, but incorporating an out-of-phase alignment condition) as identical. In fact, the similarity judgments in response to these pairs appeared no different from those comparing consonant identities (i.e. seeing a consonant AV composite compared to itself). Cluster analysis confirmed that the effect of alignment condition vanished from the grouping structure. In the resulting tree diagram, no two neighboring composites share the same alignment condition.

Results of Experiment Three lead the author to conclude that subjects *could* discriminate between the various alignment conditions when asked on a verbal rating scale specifically addressing synchronization. This ability existed even when responding to actual movie excerpts (i.e., highly complex stimuli). However, responses on a scale of effectiveness and subject similarity judgments between pairs of stimuli revealed that AV alignment was not appear one of the primary criteria in either of these determinations. Therefore, the third dimension in the MDS solution for Experiment 3 was considered likely to be related instead to general audio-visual congruency (i.e., appropriateness).

Data Analysis Across All Experiments

Synchronization. By combining all three experiments into a single data set, it was possible to consider responses to all alignment conditions across increasing levels of stimulus complexity. First, considering synchronization ratings, one final ANOVA revealed no significant main effect of either level of stimulus complexity or musical training and no significant differences were found in the interactions between stimulus complexity and musical training, between musical training and alignment condition, or between stimulus complexity, musical training, and alignment condition.

Two statistically significant differences were discovered. As always, there was a highly significant difference between responses to the various alignment conditions. The consistent trend was for consonant composites to receive the highest rating, dissonant composites to receive the lowest rating, and out-of-phase composites to receive a median rating. In addition, the interaction between stimulus complexity and alignment condition was also highly significant. This latter interaction confirmed that subject synchronization ratings were modified significantly as AV stimulus complexity increased. Composites exhibiting consonant AV relationships were rated higher on a scale of synchronization when given in response to simple stimuli, than when given in response to actual movie excerpts (i.e., complex stimuli). Conversely, synchronization ratings in response to dissonant composites were lower when judging simple stimuli than when responding to actual movie excerpts. Mean ratings for stimuli of moderate complexity were always found in the response range between the simple and complex mean ratings.

The range of responses between consonant and dissonant was much smaller for the complex stimulus situation (Experiment Three) than for the simple stimuli (Experiment One) or even the moderately complex stimuli (Experiment Two). It appears, therefore, that the *perceived level of synchronization* between consonant and dissonant com-

posites is less for the complex stimuli than for the simple animations. In other words, the consonant composites using complex stimuli (i.e., the actual motion picture excerpts used in Experiment Three) are rated less synchronized and the dissonant combinations are perceived as more synchronized than their related simple animations (Experiment One) sharing the same alignment condition.

Effectiveness. Subject responses on the VAME scale of effectiveness were also analyzed across the entire data set. The effects were identical to those of the synchronization ratings except for one additional significant interaction between musical training and alignment condition. Therefore, all that was stated above concerning interpretation of the results for synchronization ratings applies ratings of effectiveness, as well. The significant interaction between musical training and alignment condition must have resulted exclusively from the divergence of responses to the dissonant alignment condition in Experiment Three, since no other experiment revealed this interaction.

The present investigation showed that individuals representing the three levels of musical training, responded in the same way, except when observing the most complex AV stimuli. In this case, those with more than two years of musical training (moderate and highly trained) were more sensitive to the dissonant alignment condition than those individuals with less than two years of musical training.

Conclusions

Confirmation or Disconfirmation of Null Hypotheses

The following ten null hypotheses were stated in the introduction to the present study.

1. There will be no significant difference between subjects' verbal ratings of *synchronization* on the basis of accent alignment of the AV stimuli.
2. There will be no significant difference between the subjects' verbal ratings of *effectiveness* on the basis of accent alignment of the AV stimuli.
3. There will be no significant difference between the subjects' verbal ratings of *synchronization* on the basis of level of musical training.
4. There will be no significant difference between the subjects' verbal ratings of *effectiveness* on the basis of level of musical training.
5. There will be no significant interaction between accent alignment condition and level of musical training in subject ratings of *synchronization*.
6. There will be no significant interaction between accent alignment condition and level of musical training in subject ratings of *effectiveness*.
7. When considering the entire data set—adding the level of complexity as a between-subjects factor—there will be no significant interaction between level of complexity, alignment condition and musical training in the ratings of *synchronization*.
8. When considering the entire data set—adding the level of complexity as a between-subjects factor—there will be no

significant interaction between level of complexity, alignment condition and musical training in the ratings of *effectiveness*.

9. There will be no significant difference between subject *similarity judgments* as a result of the various AV composites.
10. There will be no significant difference in subject *similarity judgments* as a function of level of musical training.

Statistical confirmation or disconfirmation may now be stated for each of these hypotheses based on results of the three experiments described in this study and the analysis across the combined data set. For each hypothesis, a statement of rejection or inability to reject is made for each experiment. In cases where the hypothesis is either rejected or not rejected for multiple experiments, a single sentence suffices.

The first null hypothesis was rejected in all three experiments on the basis that subject ratings of synchronization were shown to be significantly different between alignment conditions. This null hypothesis was also rejected for the analysis across the entire data set on the same basis.

The second null hypothesis was rejected in all three experiments on the basis that subject ratings of effectiveness were shown to be significantly different between alignment conditions. This null hypothesis was also rejected for the analysis across the entire data set on the same basis.

The third null hypothesis was not rejected in all three experiments on the basis that subject ratings of synchronization were not shown to be significantly different between alignment conditions as a function of musical training. This null hypothesis was also not rejected for the analysis across the entire data set on the same basis.

The fourth null hypothesis was not rejected in all three experiments on the basis that subject ratings of effectiveness were not shown to be significantly different between alignment conditions as a function of musical training. This null hypothesis was also not rejected for the analysis across the entire data set on the same basis.

The fifth null hypothesis was not rejected in all experiments on the basis that subject ratings of synchronization were not shown to be significantly different as a function of the interaction between alignment condition and level of musical training. This null hypothesis was also not rejected for the analysis across the entire data set on the same basis.

The sixth null hypothesis was not rejected for Experiments One & Two on the basis that subject ratings of effectiveness were not shown to be significantly different as a result of the interaction between alignment condition and level of musical training. This null hypothesis was rejected for Experiment Three and for the analysis across the entire data set on the basis that a significant difference was shown to exist between subject ratings of effectiveness as a function of the interaction between alignment condition and level of musical training.

The seventh null hypothesis was rejected on the basis that there was a significant difference as a function of the interaction between level of complexity, alignment condition, and musical training in the synchronization ratings across the entire data set.

The eighth null hypothesis was rejected on the basis that there was a significant difference as a function of the interaction between level of complexity, alignment condition, and musical training in the effectiveness ratings across the entire data set.

The ninth null hypothesis was rejected because there was a significant difference between subject similarity judgments to the various AV composites.

The tenth, and final, null hypothesis was not rejected because there was not a significant difference between subject similarity judgments as a function of level of musical training.

Research Questions Answered

The first question posed was “What are the determinants of ‘accent?’” A review of the relevant literature revealed a plethora of potential sources of accent (i.e., visual and musical parameters). Several researchers and theorists proposed that introducing a change into a stimulus streams results in added salience (i.e., accent). The preliminary study to Experiment One determined that hypothesized accent points, using parameters (both aural and visual) gleaned from this literature review, were 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 (left-bottom-right-top, top-to-bottom, & side to side) and translation in depth (back-to-front). These results were obtained for the purpose of incorporating the most reliable stimuli into Experiment One, rather than generalizing these findings to situations outside the current experimental context.

The second research question—and the main source of interest in the present study—asked 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 in the VAME responses of Group One across all three experiments. For the simple and moderately complex AV combinations, subject ratings of synchronization and effectiveness shared a strong positive relationship ($r = .96$ and $r = .94$, respectively). Therefore, AV combina-

tions that were rated high in synchronization also tended to be rated high on effectiveness and vice versa. Interestingly, however, a divergence appeared in subject responses to the more complex stimuli in Experiment Three. Though subjects clearly distinguished between consonant and out-of-phase alignment conditions in their ratings of synchronization, their ratings of effectiveness did not discriminate between these two alignment conditions and the correlation between the two dropped substantially ($r = .74$). In other words, subjects were rating the alignment condition in which the audio and visual accent structures were purposely placed out of synchronization equally as high on a scale of effectiveness as when the accent structures were perfectly aligned! The lessening role of alignment condition is confirmed in the cluster analysis for similarity judgments.

We may conclude that, when the audio and visual components were simple or moderately complex, accent structure alignment did appear to be a necessary condition in order for an AV combination to be considered effective. However, when the audio and visual stimuli attained the complexity of a typical motion picture, accent alignment was no longer a necessary condition.

Suggestions for Further Research

The most important issue to be addressed in a series of future investigations is the resolution of the confound between the three alignment conditions in the present study. In effect the dissonant combination differed from the other two not only in alignment condition, but also in musical component. It was not possible to determine whether these judgments were based on accent structure alignment or on audio-visual congruence, since both the consonant and out-of-phase alignment conditions shared the intended soundtrack, while the music accompanying the dissonant combination was not composer-intended. There are several ways of resolving this problem, each with its own inherent problems and/or difficulties.

It would be possible theoretically to create an out-of-phase alignment condition that does not use the intended music. As a result, the confound between dissonant combinations and unintended music is eliminated. However, creating an out-of-phase alignment condition with unintended music brings up other issues. In the present study, the purpose of using the same music for both the consonant and out-of-phase alignment conditions was to control for differences in accent structure. In essence, by simply shifting the same music temporally, the accent structures remained constant but were set intentionally out of synchronization.

However, in creating an out-of-phase alignment condition with an unintended musical score, the investigator must establish with certainty that the accent structures are identical between the consonant and out-of-phase conditions, otherwise the conditions are no longer definitionally consistent. Creating such a stimulus set could be accomplished, using a theory-driven algorithm. However, incorporating real (i.e., ecologically valid) music into this paradigm would be no small feat due to the need to establish that the two musical scores share an identical accent structure.

A second potential resolution would be to incorporate a MIDI version of the intended musical score. Using a state-of-the-art sound card with WAVE table synthesis, sampled orchestral sounds might even be utilized as the auditory track. Using a less-expensive FM synthesis sound card would be unacceptable in terms of validity. A second issue in this potential method of replication is the difficulty in recreating a composer-intended version to accompany the consonant alignment condition. The rubati and other expressive characteristics of most orchestral scores would be difficult, if not impossible, to reproduce on a synthesizer keyboard. However, once the original version was sequenced and ready to play, it would be a simple to adjust the tempo or make other temporal changes to create an out-of-phase and dissonant condition. In this paradigm one could

simply use the same motivic ideas with different tempos (i.e., IOI periodicities) as in Experiment One of the present study. In this solution, however, there is certain to be a significant loss of ecological validity.

The best possible solution would be to utilize music that was specially-composed for a given scene with the specific intent of creating a consonant, out-of-phase, and dissonant combination, as operationally defined in the present study. Ideally, an orchestral recording of the musical score would be made and incorporated into the experimental design. However, a more cost-effective means of production would be to compose an electronic music score. In the latter case, the audio and visual tracks could be linked via SMPTE (a standard of the Society of Motion Picture & Television Engineers) time code, ensuring high precision of the AV synchronization.

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. Results of the present study suggested that the relationship of these two factors may change as complexity of the AV combination increases. These results would help further develop the modified version of the Film Music Paradigm.

The model could also be expanded 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 the subjects taps with the audio, the video, some underlying common pulse, or a complex combinatory rhythmic pattern.

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 in the first study (225ms) and amounts in the later studies were determined by exploratory studies (between 100ms and 890ms) 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 both the Film Music Paradigm and the 3-D Model of Film Classification.

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 literacy³⁷ as a potential grouping variable. In fact, incorporating both musi-

cal training and visual literacy would allow consideration of the musical training by visual literacy interaction, which might prove very interesting.

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 a paucity of research literature available addressing issues involved in high-level 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. As mentioned previously, the long-term goal of this line of research is to determine fundamental principles governing the interaction between auditory and visual components in the motion picture experience, leading eventually to a better understanding of human cognitive processing in general.

NOTES

¹ Developed during research on synesthesia, the semantic differential scale is one method employed to identify patterns in the stimulus/response chain of human perception and cognition for the purpose of inferring a cause-and-effect relationship. Semantic differentiation was defined by Osgood as "the successive allocation of a concept to a point in the multidimensional semantic space by selection from among a set of given scaled semantic alternatives." In reference to the concept being evaluated (music, in the present study), he identified two properties of this point in semantic space: direction from the origin (determined by the bipolar adjectives selected) and distance from the origin (i.e. extremeness of the subject response on the rating scale). These two properties of the semantic differential were assumed to be related to conceptual mediators evoked in the subject and the intensity level of this evocation, respectively (Osgood, 1957, pp. 26-30). Osgood, et al divided the semantic differential scale into three factors: Evaluative, Potency and Activity.

² In his theory of categorization, Jerome Bruner (1966) referred to specific characteristics of objects as "attributes." Those attributes necessary to define an object (i.e., distinguish it from other objects) are said to be "criterial," while those that do not are "irrelevant." For a discussion of Bruner's theory see LeFrancois (1982, p. 163).

³ The latter was the stated aesthetic ideal striven for by many early film sound theorists and realized in René Clair's Le Million (1931) and A nous la liberté (1932) as an alternative to that characterized by "talkies" (Weis & Belton, 1985, p. 77).

⁴ Treisman and Julesz discovered that, when elements are distinctively different (as in Figure 2.4), the boundaries are perceived almost automatically within 50 ms. However, if the elementary forms are similar, the recognition of a boundary takes longer (see Kandel, 1991 for a discussion).

⁵ It is also possible to use spherical coordinates, $P(\theta, \phi)$.

⁶ See Braunstein (1973) for a detailed discussion of motion perception.

⁷ The specific relationships are: 1000ms = 20 frames; 800ms = 16 frames, and 500ms = 10 frames

⁸ 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.

⁹ If there is any doubt that visual motion is capable of creating perceived accent structures as presented in this study, see Fischinger's "Radio Dynamics" (1943). The artist subtitled the piece "A color-music composition" and confirmed this sentiment by presenting a message in the opening titles which explicated his desire that no musical accompaniment be added.

¹⁰ More information about MEDS is available from the author or by contacting Roger Kendall directly at: UCLA Dept. of Ethno- & Systematic Musicology, Schoenberg Hall, 405 Hilgard Ave., Los Angeles, CA 90024. In addition to the KeyPress module, the author also incorporated commands into MEDS, allowing selection of any of the digital or analog tracks of the disk. These capabilities were necessary for both Experiments Two and Three. These capabilities have been retained in the latest version of MEDS (4.0a) in a modified form.

¹¹ Documentation accompanying the MicroHelp's Muscle subroutine utilized in the Keypress module claims a "PrecisionTick" resolution of 838 nanosecond ticks (.000838ms) as opposed to the system tick of 1/18 sec.

¹² Kendall & Carterette (1993 & 1992b) have shown that the ability of subjects to distinguish musical timbres on verbal scales can be improved and response variance decreased by using strict opposites (e.g. "good" and "not good") rather than familiar opposing terms (e.g. "good" and "bad"). They refer to the former type of scaling procedure as VAME (Verbal Attribute Magnitude Estimation), as opposed to

semantic differentiation, as exemplified by the latter pair of bipolar opposites. In their study of simultaneously-sounding wind instrument timbres, Kendall & Carterette (1993) were unable able to attain significant subject differentiation between 10 timbres using 8 factorially pure semantic differentials from von Bismarck's (1974) experiments. However, when the verbal scaling procedure was changed to VAME, the subject responses differentiated more clearly between the variety of timbres.

¹³ For a detailed description of how these images were created in 3D Studio using the 3D Loftter and Keyframer, contact the author.

¹⁴ The words in brackets represent those portions of the instructions that were altered between the audio and visual task. The word(s) appearing before the slash ("/") were used with the auditory task, while the words after the slash were used for the visual task.

¹⁵ Streaming, based on one of the Gestalt principles of organization, is normally considered to be the "phenomenon of organizing musical events into streams based on pitch proximity" (Lipscomb, in press). However, in this case the separation of sound streams occurred due to changing timbre.

¹⁶ This subject pool consisted of 24 males and 16 females.

¹⁷ The author discovered that, when the MIDI files created in Finale were combined with the visual stimuli and played back using the Autodesk Animation Player™ for Windows, there was a delay of approximately 40ms before the onset of the audio portion. Because it was necessary to maintain strict control over the onset of both the auditory and visual portions of the composite, other methods of saving the audio stimuli were attempted. The MIDI files were recreated using CakeWalk Pro for Windows 1.03. When these new files were combined with the visual images, there was no apparent delay in onset of the audio track. Therefore, the latter were used in Experiment One. In addition, since previous research has shown that when sources of accent occur simultaneously the resulting perceived accent takes on added significance (Monahan, Kendall, & Carterette, 1987; Drake, Dowling, & Palmer, 1991), a dynamic accent (veloc-

ity value of 110 in comparison to the normal level of 80) was placed at each hypothesized accent point to amplify the salience of those events.

¹⁸ Remember that there is no *nested consonant* composite for the 800ms stimuli.

¹⁹ 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 were perceiving 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.

²⁰ 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. The Shepard diagram (plotting distances in the final configuration against similarities) was clearly unacceptable, resulting in a stepwise function and the 2-dimensional solution claimed to account for 100% of the variance at a stress level of .00028. It appears that subjects may have been responding on a discrete metric (e.g., synchronized vs unsynchronized) instead of a continuous one (e.g., degrees of synchronization).

In cases resulting in a degenerate solution, Kruskal (1978) suggests randomly dividing the group into subgroups, as a test of the robustness of the dimensions in the MDS solution. The 20 subjects in Group Two were randomly assigned to one of two subgroups and an MDS solution was calculated. The solution for both subgroups resulted in the same three dimensions as the previous MDS solution, incorporating all 20 subjects. This supports the notion that the dimensions themselves are quite robust. However, the fact remains that the MDS solution is degenerate, so an alternate means of quantitative analysis was required. Therefore, throughout this study, results of the MDS solution will be supported by consideration of cluster analyses as well.

²¹ Cluster analyses throughout the present study were calculated using SYSTAT's Cluster module, because it uses a *seriation algorithm* in determining group branching structure for the tree diagrams. In

addition to determining group membership, the between-cluster neighbors were placed such that objects that were most similar are placed closest to each other (Wilkinson, et al., 1992, p. 29).

²² Code was written in Visual Basic 3.0 to allow computer control of both the laserdisc player and WAVE files stored on the computer hard drive. Essentially, using the appropriate MCI command strings, the computer prepared for each AV composite using the following sequence of events. Using Start and Stop values provided by the experimenter, the laserdisc player searched for the appropriate location on the disk and was placed in Pause mode. Then the WAVE file (also selected by the experimenter) was loaded into its memory buffer and placed into pause mode. After both the laserdisc player and sound card were thus prepared, a Play command was used to produce the AV composite. Therefore, because of the coprocessor overhead in a typical multimedia PC, there was potential for variability in the amount of time required for both the audio and visual stimuli to be placed into Pause mode. However, once both systems were paused, the Play commands required little effort on the part of the coprocessor, so alignment of the audio and visual was extremely accurate (i.e., unlikely to be off more than 1ms).

²³ The mean standard deviation across subjects in the auditory tapping exploratory study was 42.85 for “Canon” and 102.53 for “Synchrony,” suggesting higher reliability in the perception of accent structure for the former over the latter.

²⁴ This subject pool consisted of 24 males and 16 females.

²⁵ This dimension might be more accurately labeled “Audio Complexity,” since A1 exhibits a higher level of rhythmic complexity than the other two audio tracks. However, if this were truly the case, then A2 should lie somewhere in the middle ground between the extreme simplicity of A1 and A3.

²⁶ Recall that these animations were selected because of the high degree of synchronization between the audio and visual components.

²⁷ This procedure is sometimes referred to as paired-samples *t*-tests.

28 The Norman McLaren laserdisc used in Experiment Two was CAV format, allowing access by individual frames. The Obsession laserdisc, however, was in extended CLV format so the various excerpts were located using the precise Start Time and Stop Time (in *minutes:seconds* format).

29 This subject pool consisted of 15 males and 25 females.

30 These planned comparisons were performed using the preliminary professional version of GANOVA, created by M.L. Brecht, J.A. Woodward, & D.G. Bonett.

31 As discussed in the analysis of subject data in Experiment One, this was actually the third analysis of that particular data set.

32 The interaction between stimulus complexity, musical training, and alignment condition ($F_{\lambda(8,100)} = 2.060, p < .0469$) would have been considered significant at a less conservative level of alpha.

33 Recall that nested and subdivided responses were also considered to represent the hypothesized rate, because they either double or halve the IOI produced by the majority of subjects.

34 The author programmed a module for inclusion in MEDS to record these inter-keypress intervals.

35 This type of data transformation was utilized in the tapping study preceding each experiment.

36 Since there were no sound or animation files using an IOI of 1600ms or 400ms, it was impossible to create nested versions of the 800ms accent periodicities. However, there were two potential dissonant combinations (500ms & 1000ms). Therefore, there were a total of 14 alignment conditions.

37 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.

REFERENCES

- Adelson, E.H. & Bergen, J.R. (1991). The plenoptic function and elements of early vision. In M.S. Landy & J.A. Movshon (Eds.) Computational Models of Visual Processing. Cambridge, MA: MIT Press.
- Arnheim, R. (1985). A new laocoön: Artistic composites and the talking film. In E. Weis & J. Belton (Eds.) Film Sound: Theory and Practice. New York: Columbia University Press. Originally published in Film as Art, 1938.
- 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.
- Attneave, F. & Arnoult, M.D. (1956). The quantitative study of shape and pattern perception. *Psychological Bulletin*, 53, 452-471.
- 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.
- Bismarck, G. von (1974). Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acustica*, 30, 146-159.
- Bolton, T.L. (1894). Rhythm. *American Journal of Psychology*, 6, 145-238.
- Boltz, M. & Jones, M.R. (1986). Does rule recursion make melodies easier to reproduce? If not, what does? *Cognitive Psychology*, 18, 389-431.
- Börjesson, E., & von Hofsten, C. (1972). Spatial determinants of depth perception in two-dot motion patterns. *Perception & Psychophysics*, 11, 263-268.
- Börjesson, E., & von Hofsten, C. (1973). Visual perception of motion in depth: Application of a vector model to three-dot motion patterns. *Perception & Psychophysics*, 13, 169-179.
- Bregman, A.S. (1990). Auditory scene analysis. Cambridge, MA: MIT Press.
- Bregman, A.S. & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, 89, 244-249.
- Brown, R.W. (1981). Music and language. In *Documentary report of the Ann Arbor Symposium*. Reston, VA: pp. 233-265.
- Bruner, J.S. (1966). Toward a theory of instruction. Cambridge, MA: Harvard University Press.
- 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.
- Campbell, W. & Heller, J. (1980). An orientation for considering models of musical behavior. In D. Hodges' (Ed.) Handbook of music psychology, pp. 29-35. Lawrence, KA: National Association for Music Therapy.
- Cardinell, R.L. & Burris-Meyer, H. (1949). Music in industry today. *Journal of the Acoustical Society of America*, 19, 547-548.
- Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.
- Chomsky, N. (1975). Reflections on language. New York: Pantheon.
- Collins, M. (1990). The Universe of Norman McLaren. Liner notes for the Visual Pathfinders laser disk "The world of Norman McLaren: Pioneer of innovative animation" (PSI-90-018).
- Cooper, G. & Meyer, L.G. (1960). The rhythmic structure of music. Chicago: University of Chicago Press.
- 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.
- Davies, J.B. (1978). The psychology of music. Stanford, CA: Stanford University Press.
- Davies, J.B. & Jennings, J. (1977). Reproduction of familiar melodies and the perception of tonal sequences. *Journal of the Acoustical Society of America*, 61(2), 534-541.
- Deliege, I. (1987). Grouping conditions in listening to music: An approach to Lerdahl & Jackendoff's Grouping Preference Rules. *Music Perception*, 4(4), 325-360.
- Deutsch, D. (1982). Grouping mechanisms in music. In D. Deutsch (Ed.) The psychology of music, pp. 99-134.
- Dowling, W.J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, 85, 341-354.
- Dowling, W.J. & Fujitani, D.S. (1971). Contour, interval, and pitch recognition in memory for melodies. *Journal of the Acoustical Society of America*, 49, 524-531.
- Dowling, W.J. & Harwood, D.L. (1986). Music cognition. New York: Academic Press.
- Drake, C., Dowling, W.J., & Palmer, C. (1991). Accent structures in the reproduction of simple tunes by children and adult pianists. *Music Perception*, 8(3), 315-334.

- 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.
- Eisenstein, S.M., Pudovkin, V.I., & Alexandrov, G.V. (1985). A statement [on the sound film]. In E. Weis & J. Belton (Eds.) Film sound: Theory and practice. New York: Columbia University Press. Originally published in 1928.
- Evans, R.M. (1974). The perception of color. New York: John Wiley & Sons.
- 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.
- Gomery, D. (1985). The coming of sound: Technological change in the American film industry. In E. Weis & J. Belton (Ed.) Film Sound: Theory and Practice. New York: Columbia University Press.
- Gorbman, C. (1987). Unheard melodies: Narrative film music. Bloomington: Indiana University Press.
- Halpin, D.D. (1943-4). Industrial music and morale. *Journal of the Acoustical Society of America*, 15, 116-123.
- Harrell, J. G. (1986). Soundtracks: A study of auditory perception, memory and valuation. Buffalo, NY: Prometheus Books.
- Heinlein, C.P. (1928). The affective characters of major and minor modes in music. *Journal of Comparative Psychology*, 8, 101-142.
- Hevner, K. (1935). Expression in music: A discussion of experimental studies and theories. Psychological Review, 42(2), 186-204.
- Hevner, K. (1936). Experimental studies of the elements of expression in music. American Journal of Psychology, 48, 246-269.
- Hough, E. (1943). Music as a safety factor. *Journal of the Acoustical Society of America*, 15, 124.
- Hubel, D.H. & Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160, 106-154.

- 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).
- Hutchinson, W. & Kuhn, L. (1992). I musernas sällskap [The complementarity of music and the visual arts], pp. 534-564. In *Festschrift für Ulla-Britta Lageroth*. Wiken, Sweden.
- Jackendoff, R. (1987). Consciousness and the computational mind. Cambridge, MA: MIT Press.
- Jones, M.R. & Yee, W. (1993). Attending to auditory events: The role of temporal organization. In S. McAdams & E. Bigand (Eds.) Thinking in sound, pp. 69-112. Oxford: Clarendon Press.
- Julesz, B. (1984). Toward an axiomatic theory of preattentive vision. In B.M. Edelman, W.E. Gall, and W.M. Cowan (Eds.), Dynamic aspects of neocortical function. New York: Wiley, pp. 585-612.
- Kandel, E.R. (1991). Perception of motion, depth, and form. In E.R. Kandel, J.H. Schwartz, & T.M. Jessell (Eds.) Principles of Neural Science, 3rd ed. New York: Elsevier.
- Kendall, R.A. (1987). Model-building in music cognition and artificial intelligence. Proceedings of the First Annual Artificial Intelligence and Advanced Computing Conference. East Wheaton, Ill: Tower Conference, Inc.
- 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.
- Kendall, R.A. & Carterette, E.C. (1990). The communication of musical expression. *Music Perception*, 8(2), 129-164.
- Kerlinger, F.N. (1965). Foundations of behavioral research, (2nd ed.). NY: Holt, Rinehart, & Winston, Inc.
- Kerr, W.A. (1945). Effects of music on factory production. *Applied Psychology Monographs*, no. 5. California: Stanford University.
- Kilpatrick, F.P. (Ed.) (1952). Human behavior from the transactional point of view. Hanover, NH: Institute for Associated Research.
- Koffka, K. (1935). Principles of Gestalt psychology. New York: Harcourt, Brace.

- Köhler, W. (1929). Gestalt Psychology. New York: Liveright.
- 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.
- Kuhn, L.D. (1986). Film as *Gesamtkunstwerk*: An argument for the realization of a romantic ideal in the visual age. Unpublished doctoral dissertation, University of California, Los Angeles.
- Langer, S. K. (1942). Philosophy in a new key: A study in the symbolism of reason, rite, and art. Cambridge, MA: Harvard University Press.
- LeFrancois, G.R. (1982). Psychological theories and human learning, 2nd ed. Monterey, CA: Brooks/Cole Publishing.
- Lerdahl, F. & Jackendoff, R. (1983). A generative theory of Tonal Music. Cambridge, MA: MIT Press.
- Lipscomb, S.D. (in press). Cognitive organization of musical sound. In D. Hodges' Handbook of Music Psychology. San Antonio, TX: Institute for Music Research.
- 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. (in press). Perceptual judgment of the relationship between musical and visual components in film. *Psychomusicology*, 13(1).
- Livingstone, M.S. & Hubel, D.H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, 7, 3416-3468.
- 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.

- Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco: Freeman.
- Marshall, S.K. & Cohen, A.J. (1988). Effects of musical soundtracks on attitudes toward animated geometric figures. Music Perception, 6, 95-112.
- Martin, J.G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. Psychological Review, 79, 487-509.
- Massaro, D.W. & Warner, D.S. (1977). Dividing attention between auditory and visual perception. Perception & Psychophysics, 21, 569-574.
- Matlin, M.W. (1994). Cognition, 3rd ed. New York: Harcourt Brace Publishers.
- McGehee, W. & Gardner, J.E. (1949). Music in a complex industrial job. Personnel Psychology, 2, 405-417.
- McMullen, P.T. (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.
- Mishkin, M. (1972). Cortical visual areas and their interactions. In A.G. Karczmar & J.C. Eccles' (Eds.) Brain and human behavior. Berlin: Springer-Verlag.
- Mishkin, M., Lewis, M.E., & Ungerleider, L.G. (1982). Equivalence of parieto-preoccipital subareas for visuospatial ability for monkeys. Behavioural Brain Research, 6, 41-55.
- Mishkin, M. & Ungerleider, L.G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. Behavioural Brain Research, 6, 57-77.
- 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.
- Munsell, A.H. (1942). The Munsell book of color. Baltimore, MD: Munsell Color Co.

- 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.
- Ostwald, W. (1931/1933). Color science, Pt. I & Pt. II. London: Windsor & Newton.
- Ortmann, O. (1926). On the melodic relativity of tones. *Psychological Monographs*, 35(162), 1-47.
- Peirce, C.S. (1931-35). Collected Papers (Vols. 1-6) (C. Hartshorne & P. Weiss, Eds.). Cambridge, MA: Harvard University Press.
- Povel, D.-J. (1981). The internal representation of simple temporal patterns. *Journal of Experimental Psychology: Human Perception & Performance*, 7, 3-18.
- Povel, D.-J. & Okkerman, H. (1982). Accents in equitone sequences. *Perception & Psychophysics*, 30(6), 565-572.
- 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.
- Ruff, R.M. & Perret, E. (1976). Auditory spatial pattern perception aided by visual choices. Psychological Research, 38, 369-377.
- Rule, S.J. (1969). Equal discriminability scale of number. *Journal of Experimental Psychology*, 79, 35-38.
- Seashore, C.E. (1919). Seashore measures of musical talent. Chicago: C.H. Stoelting & Co.
- Seashore, C.E. (1938). Psychology of music. New York: Dover Publications.
- Senju, M. & Ohgushi, K. (1987). How are the player's ideas conveyed to the audience? *Music Perception*, 4(4), 311-24.
- Shannon, C.E. & Weaver, W. (1949). The mathematical theory of communication. Urbana: University of Illinois Press.
- Staal, H.E. & Donderi, D.C. (1983). The effect of sound on visual apparent movement. American Journal of Psychology, 96, 95-105.
- Stevens, S.S. (1956). The direct estimation of sensory magnitudes--loudness. *American Journal of Psychology*, 69, 1-25.

- Stevens, S.S. (1959). Cross-modality validation of subjective scales for loudness, vibration, and electric shock. *Journal of Experimental Psychology*, 57, 201-209.
- Sutherland, N.S. (1973). Object recognition. In E.C. Carterette & M.P. Friedman (Eds.) Handbook of Perception, vol. 3. New York: Academic Press.
- 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.
- Thomas, J. (1986). Basic sensory processes: Overview. In K.R. Boff, L. Kauffman, & J.P. Thomas (Eds.) Handbook of perception and human performance, vol. I, pp. II3-4. New York: John Wiley & Sons.
- Thomassen, J.M. (1982). Melodic accent: Experiments and a tentative model. *Journal of the Acoustical Society of America*, 71, 1596-1605.
- Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 194-214.
- Treisman, A. (1988). Features and objects: The fourteenth Bartlett Memorial Lecture. *Journal of Experimental Psychology*, 40A(2), 201-237.
- Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12, 107-41.
- Treisman, A. & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95, 15-48.
- Uhrbock, R.S. (1961). Music on the job: Its influence on worker morale and production. *Personnel Psychology*, 14, 9-38.
- Ungerleider, L.G. & Mishkin, M. (1982). Two cortical visual systems. In D.J. Ingle, R.J.W. Mansfield, & M.A. Goodale (Eds.) The analysis of visual behavior. Cambridge, MA: MIT Press.
- van Noorden, L.P.A.S. (1975). Temporal coherence in the perception of tone sequences. Unpublished doctoral dissertation. Technische Hogeschool Eindhoven, The Netherlands.
- von Ehrenfels, C. (1890). Über Gestaltqualitäten Vierteljahrschrift für wissenschaftliche Philosophie, 14, 249-292.
- Vos, P.G. (1977). Temporal duration factors in the perception of auditory rhythmic patterns. *Scientific Aesthetics*, 1, 183-199.

- Wedin, L. (1972). A multidimensional study of perceptual-emotional qualities in music. *Scandinavian Journal of Psychology*, 13, 1-17.
- Weis, E. & Belton, J. (1985). Film sound: Theory and practice. New York: Columbia University Press.
- Wertheimer, M. (1925). Über Gestalttheorie. Erlangen: Weltkreis-Verlag.
- Wilkinson, L., Hill, M.A., Welna, J.P. & Birkenbeuel, G.K. (1992). SYSTAT for Windows: Statistics, version 5 edition. Evanston, IL: SYSTAT, Inc.
- Wright, B. & Braun, B.V. (1985). Manifesto: Dialogue on sound. In E. Weis & J. Belton (Eds.) Film Sound: Theory and Practice. New York: Columbia University Press. Originally published in 1929.
- 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.
- Zusne, L. (1970). Visual perception of form. New York: Academic Press
- Zwislocki, J.J. (1969). Temporal summations of loudness. *Journal of the Acoustical Society of America*, 46, 431-441.