

## PERCEPTUAL JUDGEMENT OF THE RELATIONSHIP BETWEEN MUSICAL AND VISUAL COMPONENTS IN FILM

Scott D. Lipscomb  
The University of Texas at San Antonio

Roger A. Kendall  
University of California, Los Angeles

In this study, the authors investigate the relationship between the musical soundtrack and visual images in the motion picture experience. Five scenes were selected from a commercial motion picture along with their composer-intended musical scores. Each soundtrack was paired with every visual excerpt, resulting in a total of 25 audiovisual composites. In Experiment 1, the 16 subjects selected the composite in which the pairing was considered the "best fit." Results indicated that the composer-intended musical score was identified as the best fit by the majority of subjects for all conditions. In Experiment 2, the 15 subjects rated all 25 composites on semantic differential scales. A significant interaction ( $p < .00005$ ) between audiovisual combination and the various semantic differential scales was found. Analysis of this interaction revealed that the composer-intended combination yielded higher mean scores in response to the 4 adjective pairs of the Evaluative dimension. Clustering the subject responses into 2 factor scores (Evaluative vs. a hybrid of Activity and Potency), confirmed these high Evaluative mean scores. In addition, the response contours of the Activity/Potency dimension remained relatively consistent, suggesting that music exercises a strong and consistent influence over the subject responses to an audiovisual composite, *regardless of visual stimulus*. The results corroborate previous research, indicating that a musical soundtrack can change the "meaning" of a film presentation. Comparison of the various soundtracks in music theoretical terms assisted in identifying musical elements that appeared to be relevant to specific subject ratings. These comparisons were utilized in the formulation of a model for music communication in the context of the motion picture experience.

Music has played an integral part in the motion picture experience almost since its inception.<sup>1</sup> Even so-called "silent films" were usually accompanied by musical performers. Considering the popularity of this artform and the fact that it has developed into a multi-billion dollar industry, it is quite surprising that there has been little empirical investigation into the role of film music. In the present study, the authors investigate the relationship between visual activity on-screen and the musical soundtrack. Two specific questions are of particular interest. First, can listeners reliably select the composer's intended soundtrack for a given visual scene from among several musical selections? Second, does a significant amount of variation occur in the perceptual response to a given scene when the visual stimulus remains constant and only the music is changed?

## Related Literature

There has been much speculation about the interaction of music and the visual element in motion pictures (Eisenstein, 1947; Evans, 1975; Gorbman, 1987; Kracauer, 1960; Prendergast, 1977; Thomas, 1973; Thomas, 1991; Weis & Belton, 1985; Zettl, 1990). A search for related literature in this area, however, has shown a paucity of empirical investigations dealing explicitly with this interaction.

In the field of perceptual psychology, interaction between the aural and visual sensory modalities is well-documented. Radeau and Bertelson (1974) found that when a series of lights and tones is presented at a 15 degree 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 and Donderi (1983) showed that introducing an aural stimulus lowered the interstimulus interval at which the 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; Massaro & Warner, 1977; Mershon, Desaulniers, Amerson, & Kiever, 1980; Regan & Spekreijse, 1977; and Ruff & Perret, 1976).

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 the bipolar adjectives used in semantic differential scaling,<sup>2</sup> 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 the Potency and Activity dimensions. The overall evaluation of the play did not change significantly. However, there are several problems concerning the musical aspect of his stimulus presentation. First, the musical selection was not composed for the specific scene that it accompanied. In a rather ambiguous explanation of the selection process, Tannenbaum explains that the piece was chosen by a person who "has considerable experience in this kind of work" and confirmed by a panel of four "experts" (Tannenbaum, 1956). 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, the synchronization of the dramatic action and the musical accompaniment was left largely to chance. This procedure is not at all representative of the relationship that occurs 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 "volume" (meaning loudness) was manually increased "for dramatic impact" (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 and Levenson (1983) recorded five different physiological measurements during exposure to a 12 min black and white industrial safety film depicting three accidents. These measures included the subject's interbeat interval of the heart, general somatic activity, skin conductance level (SCL), finger pulse transmission times, 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 musical scores were composed for presentation with the film for comparison with the responses to a control (i.e., no music) condition. The *documentary music* is 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* is 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 context of 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 s, the second accident by 10 s, and the final accident by 30 s. In each instance, the music ended approximately 10 s after the accident at a natural cadence point (both musically and visually). Although 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, although Thayer and Levenson may conclude from their data that the use of music caused either a heightened or reduced electrodermal response to the stressful stimuli, more evidence is 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 and 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 (which were described as *strong* and *weak*) varying on a number of musical dimensions, for example, major/minor mode, fast/slow tempo, high/low pitch, and single-/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 onto the videotape. The authors provide a brief description of the action occurring at the point when each of the themes is introduced. However, apart from the beginning point, no information is provided concerning the specific interaction of the aural and visual stimuli. A second problem with these musical compositions is that

their extreme simplicity of content and excessively repetitive structures fail to provide an accurate representation of the highly developed craftsmanship evident in a typical movie score. Even using this limited musical vocabulary, the results produced 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, appeared to depend on a complex interaction of the musical and visual materials. However, in this particular investigation, the simplicity of the stimulus materials seriously limits generalizability of the results to motion pictures.

#### Significance of the Present Study

The significance of the present study lies in three main issues. First of all, unlike previous studies, this investigation uses "real" musical and visual examples. Hevner (1936) stated that

Since we are looking for elements of *music* we must be sure that the material provided for observation represents real *music* and not merely *elements* trimmed down for experimental purposes to such an extent that all *music* has been left out (p. 248).

To the knowledge of the authors, the present investigation is the first to use actual excerpts from a major motion picture as stimuli for an experimental procedure of this type.<sup>3</sup> By improving the validity of the stimulus materials, the conclusions may be more confidently generalized to the "real world" from which they were abstracted.

Secondly, this study provides a method for quantifying compatibility between the aural and visual modalities. In the measurement of this compatibility, the motion picture experience is considered as a system. Judgments made within this context are not based on either modality (aural or visual) in isolation, but are made in reference to an audiovisual composite, taking into account cross-modal interactions and interrelationships.

Finally, the quantitative scores used in formulating this scale of compatibility will be made using the perceptual aesthetic judgment of each individual subject. The analytical methods utilized in the present study consider perception as the critical frame of reference.

Quantization of perceptual response poses certain difficulties for empirical studies. In selecting semantic differential scaling, the purpose of the present investigation was *not to determine the connotative or denotative meaning* of various musical stimuli, but rather to observe indications of perceptual change based on the subjects' semantic differential responses to various audiovisual stimuli. Since musical communication does not require the transmission of an explicit connotative or denotative meaning, one might ask what information is gained from subject responses on a semantic differential scale to a musical stimulus. In this case, words must be assigned via the implicit perceptual/cognitive apparatus using past experience as a refer-

ent. Kendall and Carterette (1991) suggest that the process of assigning verbal attributes to musical sound involves:

the superposition of (at least) two multidimensional spaces. One space is comprised of musical images, relationships among sound schemas, the other comprises referential, semantic (verbal) meanings. The mapping of one space to another is almost certainly not linear, and the resulting composite space is hybrid, neither "musical" nor "verbal." . . . Words are useful in communicating musical ideas, but they are not the ideas themselves (p. 391).

Therefore, within the context of this investigation, the responses are considered to be an indication of the subjects' changing evaluation of the audiovisual composite scaled in a semantic space. This semantic space is determined by the identification of a relationship or the lack of such a relationship between the present stimulus and an implicit model formed from past experience. Therefore, the semantic differential is merely an indicator, using words as a communicative medium, of perceptual change.

#### Basic Hypotheses

In Experiment 1, it was hypothesized that the highest percentage of responses on the "best fit" categorizations would match the composer's intended combination. It was also hypothesized that the ratings given by subjects on the semantic differential scale in Experiment 2 would vary significantly when the visual stimulus is kept constant and only the musical soundtrack is changed.

In comparing the responses in the matching procedure (Experiment 1) with those on the semantic differential scale (Experiment 2), it was hypothesized that determinate factors in the subjective decision of selecting "best fit" may be identified. By revealing such influences on this decision-making process, it was also hypothesized that insight could be gained into elements contributing to a successful symbiosis of audio and visual stimuli within the context of a motion picture.

#### Basic Methodology

The present study employed a post-test only repeated measures factorial design in which the musical soundtrack was manipulated as the independent variable. Subject responses (either categorization or scores on the semantic differential scale) were recorded as the dependent variable.

The visual excerpts were 35 to 40 s sequences of cinematic images excerpted from the motion picture *Star Trek IV: The Voyage Home* (see Appendix A for a brief description of the scenes used). All musical soundtracks were Leonard Rosenman's compositions, intended to accompany each of the visual scenes (reductions of the musical scores are provided in Appendix B). They were edited directly from a compact disc recording of the original soundtrack onto the video tape. An audiovisual composite is defined as any combination of a visual excerpt and a musical selection. There-

fore, there were composer-intended composites and composites consisting of musical and visual combinations not intended by the composer. In Experiment 1, subjects provided "best fit" categorization responses, while subjects in Experiment 2 responded by providing ratings on a semantic differential scaling procedure.

#### Stimuli

The first objective in selecting stimulus materials was to choose scenes that were intercategory, that is, no two scenes could be so similar that the content would be easily confused. Using excerpts which were selected from the same motion picture allowed control of the production variables (e.g., composer, production quality, etc.).

Initially, 10 scenes were selected for use in a pilot study. Six student composer volunteers from a class in Composition for Motion Pictures and Television at The University of California at Los Angeles (UCLA) individually viewed a video tape of the 10 selected scenes exactly as they appeared in the film (i.e., including cinematic image, musical soundtrack, and sound effects). One scene was designated as the standard to which all others were compared; subjects responded on a 9-point Likert scale (*same/different*). Using cluster analysis with complete linkage (farthest neighbor), responses to the nine scenes were statistically clustered into four groups. One scene was selected from each of these clusters. The four chosen scenes, along with the standard, were used as the five excerpts for the main investigation.

The second objective in stimulus selection was to determine the method of aligning each of the soundtracks with every visual example so that even the most extreme mismatch results in the "best fit" possible (i.e., a potentially reasonable combination). The assistance of three student film composers was utilized in the actual editing process. In each composite, an attempt was made to synchronize, as well as possible, the aural and visual components. By trying several variations and discussing the results among the participants, a general consensus was reached concerning appropriate synchronization.

Since the ambient (i.e., non-musical) sound included on the movie soundtrack could have provided unwanted cues for the appropriateness of its pairing with a given visual scene, the original soundtrack was erased. The musical score from the compact disc soundtrack recording was then dubbed directly onto the videotape, eliminating this extramusical noise (e.g., car horns, water splashing, dialogue, etc.). The resulting composite stimulus materials consisted of only music and the cinematic image.

For Experiment 2, bipolar adjectives were selected for use in the attitude scaling procedure. Initially, adjectives were selected on the basis of factor loadings provided by Osgood et al. (1957). In addition, a review of past literature assisted in compiling a list of adjectives which had been utilized in previous studies. The final selection criterion was based on the semantic appropriateness of the adjective pair to the present study. As a

result, the 10 adjective pairs listed in Table 1 (followed by factor loadings on the first three dimensions, where available) were chosen. The semantic differential scale utilizes bipolar adjectives from each of the three factors discussed by Osgood et al. (Evaluative, Potency, and Activity). A 10th adjective pair (effective/ineffective) was added as a means of providing a direct measure of the subjects' ratings on the audiovisual pairing.

Table 1  
*Bipolar Adjectives Used in the Semantic Differential Scaling Procedure*

	I	II	III
<i>Evaluative</i>			
good/bad	1.00	.00	.00
beautiful/ugly	.52	-.29	-.02
interesting/boring	.40	-.09	.22
effective/ineffective	—	—	—
<i>Potency</i>			
strong/weak	.30	.40	.10
heavy/light	-.20	.48	-.02
tense/relaxed	.39	-.54	-.57
<i>Activity</i>			
active/passive	.17	.12	.98
fast/slow	.01	.26	.35
agitated/calm	-.15	.03	.26

The subject response procedure was slightly different from that utilized in most semantic differential studies. Rather than using a 5-, 7-, or 9-point scale, a continuous line was drawn between the adjective pair. The subject was asked to make a mark at the point on the line that was considered to provide the most appropriate response. Such an undifferentiated line scale maximizes the potential variation of subject responses and has been determined to be an important procedural modification (Schiffman, Reynolds, & Young, 1981). In order to quantify the responses, a ruler was used to measure from the zero point of each line to the mark provided by the subject. Each line segment was divided into a zero point and 25 equal divisions which could, in turn, be subdivided into four equal parts, providing a 101-point response scale.

### Subjects

Using a random number table, 31 volunteers (both musicians and nonmusicians) from a Psychology of Music class at UCLA (Winter, 1989) were randomly assigned to one of two independent groups (Experiment 1,  $N = 16$ ; Experiment 2,  $N = 15$ ). The students were given the opportunity to participate in this investigation in exchange for three points of extra credit.

### Apparatus

Stimulus presentation was made to one subject at a time using an Hitachi VT38EM Multi-System video cassette recorder and a 25-in Sony Trinitron color monitor. Since ratings on the combination of the aural and visual stimuli were the source of interest, the audio track was presented to the subjects without the use of headphones in an attempt to simulate a typical home video presentation. This environment avoids further separation of the two modes of perception under investigation. The entire procedure for an individual in either group required between 25 and 45 min to complete, depending on the length of time taken to review various combinations (Experiment 1) or to respond on the semantic differential scale (Experiment 2).

## EXPERIMENT 1

### Procedure

The subjects in Experiment 1 were assigned the task of selecting the musical score which provided the "best fit" for each visual scene. Subjects were presented with the first visual scene (randomly determined) in combination with each of the musical soundtracks, also randomly ordered. They selected the version that they considered to be the "best fit." After selecting a choice for the first visual scene, subjects moved on to the second visual scene in combination with each soundtrack, continuing until a match had been selected for each of the five visual examples. They were free to see each composite as many times as necessary in the decision-making process.<sup>4</sup>

### Results

The responses resulted in the frequency counts provided in Figure 1, wherein the cells on the diagonal represent the composer's intended combinations. Notice that the total of each vertical column equals 16 (the number of subjects). Immediately evident is the fact that the composer-intended combinations were selected by the majority as "best fit" in every case. In fact, no other combination was selected by more than 25% of the subjects. This may be interpreted as confirmation of the effectiveness of Rosenman's compositional technique. However, no single composite was selected as the most appropriate combination by every subject. A closer look at Figure 1 reveals that Visual 5 is the most convincing example of agreement between the subject pool and the composer with Visuals 1 and 2 also reflecting a high

level of agreement. Visuals 3 and 4 appear to have left the subjects less certain, as exemplified by the spread of their responses.

Considering these same results from the auditory dimension (horizontally, in Figure 1), notice that Audio 3 was not selected as an appropriate accompaniment to any scene other than that which was intended by Rosenman. Audio 4, however, seems to have been considered appropriate for a variety

		VISUAL				
		1	2	3	4	5
A	1	12	1	0	1	0
U	2	1	11	0	3	0
D	3	0	0	8	0	0
I	4	3	4	4	8	3
O	5	0	0	4	4	13

Figure 1. Data matrix showing the number of subjects who selected each cell as "best fit."

of visual scenes, being selected as best fit for each scene by at least three subjects.

Although, overall, the responses appear to be heavily weighted toward the composer-intended combinations, the spread of responses to Visual 4 across both the audio and visual domains warrants some explanation. Visual 4 is the only visual excerpt in which no human figure appears. Also, due to miscommunication during an early conversation with the composer, Audio 4 is similar to, but not identical with, the composer-intended musical score. One possible explanation for the diffusion of responses to both the audio and visual components of Visual 4 is that, in the determination of a categorization response pairing musical stimuli and cinematic excerpts, two variables interact: abstractness of the visual image (in this particular instance, presence or absence of human interaction) and availability of a soundtrack that was specifically composed for the given scene. Further investigation will be necessary to test the following hypothesis.

In a cinematic excerpt such as Visual 4, because of its abstract imagery, the music takes on a more prominent role in determining the perceived meaning of the audiovisual composite. As a result, subjects might not necessarily judge contrasting musical scores as *incompatible* with a given visual image, but may simply render a *different interpretation* of the action occurring on the screen as a result of the variation of audio stimuli. In reference to the

other possible interactive variable (i.e., presence of a composer-intended soundtrack among the choices), responses of the subjects in Experiment 1 suggest that categorization decisions made between a set of options including the composer's intended musical score are much more homogenous than selections made when the composer-intended music is not among the possible choices, although further research is necessary to confirm this possibility.

Considering the results in Figure 1, two outcomes were possible other than the formation of a general consensus among the subjects, agreeing with the composer's intent. The majority of responses could have clustered into a cell other than that representing the composite utilizing the composer's intended score, or the responses could have been spread more equally across the cells of the matrix. The former would suggest that, though the composer's intent was not matched by the subject responses, there was general agreement upon which of the five soundtracks was considered most appropriate. In the latter case (which occurred in both the audio and visual dimensions for Visual and Audio 4), the fact that the responses are more spread out among the cells of the design implies that the visual image and the musical score are considered to be more ambiguous and, as a result, could be combined successfully with several stimuli of the other modality. Since the responses to Visual 4 and Audio 4 were spread across the cells (both vertically and horizontally) rather than clustered, ambiguity appears to have played a strong role in the subject responses to the composites utilizing either of these components.<sup>5</sup>

## EXPERIMENT 2

### Procedure

The subjects in Experiment 2 rated each of the 25 audiovisual combinations on a continuous-line semantic differential scale consisting of 10 pairs of bipolar adjectives. In an attempt to negate ordering effect, two independent presentation orders were randomly generated. The only restriction imposed upon ordering the stimuli was that no selection could follow a previous composite containing either the same musical or visual component. The bipolar adjectives were also randomly arranged into three configurations with respect to both presentation order and polar position (e.g., good/bad as opposed to bad/good). Using a random number table, each subject was assigned to one of the six subgroups (i.e., two composite presentation orders and three bipolar adjective configurations).

### Results

Using only two within-subjects variables (audiovisual combinations and the semantic differential bipolar adjectives), an ANOVA was run as a separate statistical procedure on each visual scene in combination with all of the musical soundtracks (e.g., Visual 1 with Audios 1-5, Visual Two with Audios 1-5, etc.) and then keeping the audio constant and changing the visual

image (e.g., Audio 1 with Visuals 1-5, Audio 2 with Visuals 1-5, etc.).<sup>6</sup> Therefore, in order to keep the overall confidence level well above 95%, the alpha level for each of the 10 individual ANOVAs was set *a priori* to  $p < .005$ . The ANOVA results are provided in Tables 2a and 2b.

Table 2a  
Analysis of Variance for the Condition when Visual Stimulus is Constant and Only the Musical Soundtrack Changes

Source	df	SS	MSS	F
SCENE ONE:				
Within Subjects	735	18994.85		
Audio	4	201.73	50.43	1.10
Semantic Differential	9	395.19	43.91	1.63
Audio/Differential	36	4560.79	126.69	8.10**
SCENE TWO:				
Within Subjects	735	18146.46		
Audio	4	1348.54	337.14	8.06**
Semantic Differential	9	1778.25	197.58	6.77**
Audio/Differential	36	2564.34	71.23	5.58**
SCENE THREE:				
Within Subjects	735	21308.42		
Audio	4	2428.97	607.24	8.08**
Semantic Differential	9	1839.43	204.38	7.40**
Audio/Differential	36	3253.40	90.37	7.47**
SCENE FOUR:				
Within Subjects	735	18707.47		
Audio	4	822.55	205.64	3.08
Semantic Differential	9	467.90	51.99	2.46
Audio/Differential	36	3625.82	100.72	6.87**
SCENE FIVE:				
Within Subjects	735	20513.23		
Audio	4	2017.20	504.30	7.36*
Semantic Differential	9	202.66	22.52	0.84
Audio/Differential	36	4249.57	118.04	8.72**

\* $p < .0001$ . \*\* $p < .00005$ .

Table 2b  
Analysis of Variance for Condition When the Musical Soundtrack is Kept Constant and Only the Visual Stimulus Changes

Source	df	SS	MSS	F
AUDIO ONE:				
Within Subjects	735	20623.41		
Scene	4	434.97	108.74	1.65
Semantic Differential	9	1755.31	195.03	6.64**
Scene/Differential	36	3245.96	90.17	5.83**
AUDIO TWO:				
Within Subjects	735	16460.46		
Scene	4	1568.25	392.06	8.40**
Semantic Differential	9	915.27	101.70	2.71
Scene/Differential	36	1323.13	36.75	3.49**
AUDIO THREE:				
Within Subjects	735	23421.17		
Scene	4	1628.25	407.06	7.51*
Semantic Differential	9	5696.19	632.91	15.50**
Scene/Differential	36	2385.10	66.25	6.04**
AUDIO FOUR:				
Within Subjects	735	15766.76		
Scene	4	261.40	65.35	1.40
Semantic Differential	9	3329.92	369.99	11.02**
Scene/Differential	36	769.38	21.37	2.35**
AUDIO FIVE:				
Within Subjects	735	17831.11		
Scene	4	670.81	167.70	2.73
Semantic Differential	9	2587.35	287.48	7.37**
Scene/Differential	36	929.73	25.83	2.46**

\* $p < .0001$ . \*\* $p < .00005$ .