Modeling Audiovisual Tension

Morwaread M. Farbood

Dept. of Music and Performing Arts Professions, New York University, USA

mfarbood@nyu.edu

ABSTRACT

An experiment was conducted to examine the perception of audiovisual tension. Subjects provided real-time tension judgments of 2'30" stimuli in one of three categories: visual animation with audio, visuals alone, or audio alone. The animations consisted of random-dot kinematograms and the audio consisted of three excerpts taken from electronic art music. The animations were described by changes in the features that were used to generate them: contrast, speed of motion, coherence of motion, and brightness. The musical excerpts were described in terms of loudness, pitch, onset frequency, and the timbre attributes inharmonicity, roughness, spectral centroid, spectral deviation, and spectral flatness. Initial analysis consisted of correlations between auditory/visual features and mean subject responses. A computational model based on trend salience was then used to predict subject judgments of tension given the visual and auditory descriptors. The model was able to predict judgments well, resulting in high correlations between predictions and mean tension responses.

I. INTRODUCTION

While many aspects of auditory and musical tension have been explored in prior work, there has been a relative dearth of empirical work on timbral features contributing to perception of tension and no work that has examined timbral tension in combination with visual tension. Based on prior work on timbre (Bailes & Dean, 2012; Dean & Bailes, 2010; Farbood & Price, 2014; Pressnitzer, McAdams, Winsberg, & Fineberg et al., 2000, Schubert, 2004) and methodological approaches in vision research (cf. Baker & Braddick, 1982; Schütz, Braun, Movshon, & Gegenfurtner, 2010), several auditory and visual features were used to model tension responses to audiovisual stimuli.

The purpose of the study was to examine contributions of specific auditory and visual features to audiovisual tension. The auditory features included the five timbre descriptors inharmonicity, roughness, spectral centroid, spectral deviation, spectral flatness as well as loudness, pitch, and onset frequency. The visual features included speed of motion, coherence of motion, visual contrast, and visual brightness. The goal was to better understand the relative contributions of these features and to empirically determine the weights of their individual contributions to perceived tension.

II. METHOD

An experiment was conducted in which 45 participants were asked to judge how they felt tension was changing by moving a slider while observing and listening to 15 audiovisual stimuli. Each stimulus paired one of four audio files with one of four visual animations. The audio consisted of three excerpts from electronic compositions by Nono, Stockhausen, and T. H. Park, or silence, and the visual animations consisted of different types of changes in the visual parameters, or no visuals (static black screen). The animations were designed to reflect three general types of changes: random changes, gradual ramps, and shorter ramps. These changes were visualized using a monochromatic random-dot kinematogram. The number and size of dots were generated to take up exactly half of the available screen space in order to allow for equal proportions of foreground and background brightness.

III. RESULTS

The first step in the analysis was to examine how the different auditory and visual features correlated with mean subject responses. Table 1 shows the correlation coefficients (Spearman's rho) for each of the 15 stimuli. Although the *r*-values vary considerably between stimuli, the means for each feature provide an initial suggestion for which features might significantly influence perceived tension; in particular, these include loudness, roughness, and onset frequency among the auditory features, and speed of motion for the visual features.

In order to better understand the contribution of the various features to tension judgments, a trend-salience model, developed for musical tension (Farbood, 2012), was used to predict the empirical data. The model integrates tension contributions from individual features by taking into account the cumulative slope of those features within a moving "attentional" window and adjusting the slope based on whether it is a directional continuation of what happened immediately before it (the "memory" window). The attentional window snapshots are then merged as overlapping windows in time.

The durations of the memory and attentional windows as well as the memory window weight are all variables and that can be adjusted to improve the predictive power of the model. The values used in this case were mostly derived from prior work (Farbood, 2012). However, deciding the feature weights (and whether they were even necessary) were obtained in the current study through a manner similar to stepwise regression: all features were given equal weight initially and then individually removed to see if the predictions improved. The decision to retain or eliminate features and assign relative weights were done using the audio-only and visual-only stimuli.

The optimized model for the audio-only stimuli retained loudness, spectral centroid, roughness, and onset frequency. All weights were equal, except for loudness, which was twice that of the other features. The mean correlation coefficient (Spearman's rho) between the audio-only stimuli and the predictions produced by the model was .81. The optimized model for the visual-only stimuli retained all four visual features—speed of motion, contrast, coherence, and brightness (respective weights 6, 3, -1, -1)—with speed having the greatest weight followed by contrast. The mean correlation coefficient value (Spearman's rho) between the visual-only stimuli and the predictions was .72.

Table 1. Correlations between features and responses for each stimulus. Abbreviations: Aud = musical excerpt; Vis = animation; Lou = loudness; Cen = spectral centroid; Dev = spectral deviation; Inh = inharmonicity; Rou = roughness; Fla = spectral flatness; Pit = pitch; Ons = onset frequency; Spe = speed of motion; Coh = coherence of motion; Con = contrast; Bri = brightness.

Stimulus		Features											
		Audio								Visual			
Aud	Vis	Lou	Cen	Dev	Inh	Rou	Fla	Pit	Ons	Spe	Coh	Con	Bri
A1	-	.71	.04	.20	.02	.21	.29	-.01	.68	-	-	-	-
A2	-	.53	-.44	-.63	.04	.58	-.61	.28	.57	-	-	-	-
A3	-	.69	.65	.63	.40	.28	.53	.18	.12	-	-	-	-
A1	V1	.55	.17	.32	.09	-.01	.41	-.10	.69	.26	.20	.42	-.33
A1	V2	.71	.10	.24	-.04	.10	.34	.02	.73	.41	.39	-.06	.22
A1	V3	.75	.10	.24	-.01	.21	.26	.04	.62	.26	.17	-.28	-.23
A2	V1	.22	-.22	-.25	.07	.30	-.28	.39	.44	.37	-.16	.27	-.28
A2	V2	.30	-.31	-.38	.10	.38	-.40	.40	.48	.56	.52	.03	-.07
A2	V3	.36	-.42	-.42	.03	.33	-.43	.25	.46	.42	.22	-.32	-.14
A3	V1	.41	.32	.23	.21	.39	.20	.23	.17	.33	-.21	.04	-.03
A3	V2	.26	.22	.23	.17	.12	.18	.17	.36	.41	-.16	.30	-.42
A3	V3	.37	.35	.39	.31	.18	.29	.09	-.08	.39	.06	.17	.08
-	V1	-	-	-	-	-	-	-	-	.38	-.22	.39	.06
-	V2	-	-	-	-	-	-	-	-	.58	.29	.39	-.13
-	V3	-	-	-	-	-	-	-	-	.59	-.06	-.16	-.30
Mean		.49	.05	.07	.12	.26	.07	.16	.44	.41	.09	.10	-.13

When tested on stimuli with both audio and visual components, the best results were obtained when audio features were given twice the weight (or more) of the visual features. The optimized model for the audio + visual stimuli resulted in a mean correlation value of .67 (min of .31, max of .87). In general the model did quite well, particularly in predicting more local tension changes.

IV. CONCLUSION

This paper reports a preliminary analysis of an experiment that explored the perception of audiovisual tension. Subjects were asked to judge perceived tension when watching/listening to stimuli featuring three musical excerpts taken from electronic compositions paired with random-dot animations. A trend-salience model was utilized as a way of determining which auditory and visual features were influencing tension perception and the relative contributions of those features. When the model was optimized to fit the mean responses, the feature with the greatest weight was found to be loudness, followed by spectral centroid, roughness, onset frequency, and speed of motion (all equal in weight), then by contrast, and finally with much smaller negative contributions from coherence of motion and brightness. In general, the auditory features contributed significantly more to perceived tension than the visual features.

REFERENCES

Bailes, F., & Dean, R. T. (2012). Comparative time series analysis of perceptual responses to electroacoustic music. Music Perception, 29(4), 359–375.

Baker, C. L., & Braddick, O. J. (1982). The basis of area and dot number effects in random dot motion perception. Vision Research, 22(10), 1253–1259.

Dean, R. T., & Bailes, F. (2010). Time series analysis as a method to examine acoustical influences on real-time perception of music. *Empirical Musicology Review, 5*(4), 152–175.

Farbood, M. M. (2012). A parametric, temporal model of musical tension. Music Perception, 29(4), 387–428.

Farbood, M. M., & Price, K. (2014). Timbral features contributing to perceived auditory and musical tension. In *Proceedings of the 13th International Conference of Music Perception and Cognition*, Seoul, Korea.

Paraskeva, S., & McAdams, S. (1997). Influence of timbre, presence/absence of tonal hierarchy and musical training on the perception of musical tension and relaxation schemas. In *Proceedings of the 1997 International Computer Music Conference*, 438-441.

Pressnitzer, D., McAdams, S., Winsberg, S., & Fineberg, J. (2000). Perception of music tension for nontonal orchestral timbres and its relation to psychoacoustic roughness. *Perception & Psychophysics, 62*(1), 66–80.

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. Music Perception, 21(4), 561–585.

Schütz, A. C., Braun, D. I., Movshon, J. A., & Gegenfurtner, K. R. (2010). Does the noise matter? Effects of different kinematogram types on smooth pursuit eye movements and perception. *Journal of Vision, 10*(13), 26–26.