Measuring the Effectiveness of Sonified Crossmodal Attribute Pairings Using Contour Matching, Symmetry and Perceived Similarity

Rob Hamilton
Stanford University
Center for Computer Research in Music and Acoustics
rob@ccrma.stanford.edu

ABSTRACT

This paper presents the results of a user-study measuring users' perceptions of musically-sonified audio-visual crossmodal correspondences between low-level attributes of motion and sound in virtual space. Metrics including contour matching, perceived symmetry and crossmodal similarity are calculated and discussed with the goal of determining strong candidates to predict user preferences. Study results were analyzed using the Bradley-Terry statistical model, effectively calculating the relative contributions of crossmodal attributes within each attribute pairing to the perceived coherence or fit *between audio and visual data.*

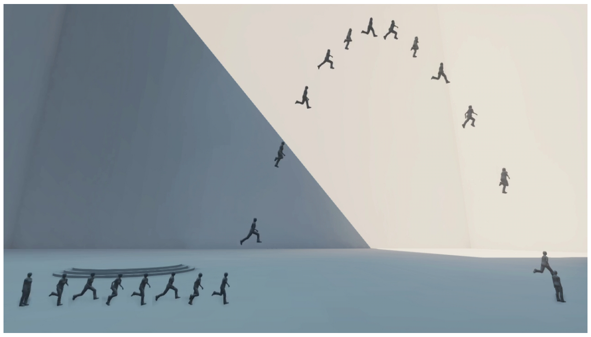

Figure 1: Study participants viewed video examples showing avatar motion in virtual space.

1. INTRODUCTION

Our relationships with sound and space are complex, bounded on one side by the inflexible laws of physics and on the other by human cognition and perception. Through experience, exposure and experimentation we each develop a personal cognitive understanding of our sound world. In doing so we are learning implicitly, continually building and revising internal models that describe how we expect sound - both environmental as well as musical sound - to accompany certain real-world interactions [1, 2, 3]. The strike of hammer on steel, the rumble of a passing train, or the cheering of a frenzied crowd all are familiar enough sounding events that most of us could agree that our internal representations of these sounds share a great number of commonalities. And while each one of us views and hears the world through different eyes and ears, our shared experiences within reality's relatively consistent sound worlds lead us to expect and predict certain sight-sound interactions in a similar and reasonably consistent way.

When perceiving and experiencing rendered immersive graphical computer environments, humans have the ability to completely reorient their visual and auditory systems, allowing a generated reality to take precedence over a physical one. In these created spaces, motion and gesture can act as direct extensions of our own physical actions or can be abstracted into forms which would be difficult if not impossible to recreate within the confines of the physical world. No commonalities of crossmodal interaction are guaranteed when stepping across the digital frontier and no limits exist to the potential mappings of sight to sound (and sound to sight) other than the creativity and whimsy of designer and developer. As such our personal internal representations must reorient themselves within each new virtual experience, often requiring us to relearn and readjust our expectations while continuously reforming our own predictive models.

If these interactions can be seen as gateways to new internal models of sonic representation, how should we consider music? Music, with its loose affiliation of time and frequency-based structures painted with varying degrees of rigor and haphazard freedom already poses distinct challenges to the idea of a commonly held perception and internal representation for all but its most basic elements. Musical form and function as well as method and meaning vary widely from composer to composer, not to mention from listener to listener, across years of history and miles of geography. In many cases music serves as an external representation of a composer or performer's internal sonic world, an abstraction of any number of ideas, influences and goals into an audible construct. Our personal internal representations of music and musical sound are thusly influenced not only by the sounds we hear in space but also by the proposed intentions of composer and performer, whether we consciously understand them or not. Adding another level of abstraction to such an

Copyright: © 2015 Rob Hamilton. This is an open-access article distributed under the terms of the Creative Commons Attribution License 3.0 Unported, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

already rich and personal set of representations is a daunting task, but one we must investigate when bringing together action and gesture from the visual modality with sound and musical expression in the auditory modality.

The understanding and analysis of musical sound provides us a low-level entry point from which we can engage this problematic. There exists a rich body of physics-based musical interaction and gesture in the history of instrument design and performance practice. From the drawing of bow on tuned string to the strike of hand on drum-head, to the arc of conductor's baton in space, musical gesture has over time evolved into a number of basic archetypes that have shaped many of our own internal representations about how musical sound is created and controlled. And while these archetypes vary from culture to culture and from age to age, their grounding in the physics of the real-world reinforces an inherent commonality in how humans perceive and internalize musical action and gesture. But when attempting to create novel sonic and musical events within rendered environments, there are no requirements as to how interaction and generated sound must relate. Musical sound and the gestures and interactions that create it can be explored and modified in rendered space by mapping data-generating process to aspect of sound, anywhere along the continuum from low-level parameter mapping all the way to high-level control over elements of musical structure or abstract musical process.

2. CROSSMODAL MAPPING

For composers and media artists seeking to present their own creative intentions and internal sonic representations to the outside world, the problem becomes one of crossmodal mapping: how to best marry sound generating processes and elements of musical form to visual occurrence within the rendered environment. In doing so we encounter a new problematic: when motion and action in space can directly create and control sound and music from low-level sounding events to high-level compositional structures, how does one decide upon the "right" cross modal mapping schema? Turning towards psychology and cognition, by better understanding mechanisms which drive our memory of and expectation for the sonic outputs of perceived interactions, physical or virtual, composers and designers can better generate creative and musical outputs that "make sense" to their audiences.

As cinema and interactive gaming experiences have grown more complex and integrated with technological processing, the interplays between visual action and musical sound have grown more pronounced and more tightly intertwined. Choreographies of camera angle and on-screen action are routinely synchronized to musical elements in musical presentations within motion pictures and music videos. In video game development it has become increasingly common to design sonic events generated within gameplay to seamlessly blend with the game's musical score [4]. And for games based around musical paradigms, gesture and motion in both virtual and real-world environments are routinely mapped to dynamic music generating and modification processes [5, 6, 7].

3. USER STUDY OVERVIEW

This research explores the perception of crossmodal relationships or correspondences between actions and gestures performed in virtual space and procedurally-generated sound processes. During the course of an exploratory user study, subjects were presented with a series of audio-visual stimuli in the form of short videos depicting humanoid avatar motion within a rendered three-dimensional environment. Musical sound, generated by mapping parameters of avatar motion to sound generating processes, is audible to subjects while viewing each video. Each stimulus consisted of video captures recorded alongside real-time data streams of avatar coordinate motion and state data. Each simple musical sonification was generated by mapping parameters from each example's multidimensional data stream to a set of parameters of a physically modeled instrument. Composite audio-visual examples were created by attaching and synchronizing the musical sonifications to each video example. A visual description of the study itself, as well as a description of techniques utilized for creating the audio-visual examples, data-validation methodologies and core results for the predictive power of paired crossmodal attributes can be found in [8].

3.1 Study Procedures

Subjects using the Mechanical Turk online tasking platform [9] were asked to watch short two-video example sets of these musically sonified avatar motions in a pairwise comparison task, choosing the example with the greatest perceived coherence or "fit" between visual and auditory events. During analysis, each visual and audio example was defined through a combination of motion and sound descriptors, allowing for the statistical analysis of correlated motion/sound pairs. For each of these modal pairs a weighted fit value was calculated and then used to calculate rankings for each example across the entire sample set, resulting in a measure of the perceived fit or coherence between individual component pairs across modalities. The perceived fit of examples exhibiting individual attributes of motion and sound were also calculated and ranked. Analyses were performed to gauge the influence of mapping direction and contour on perceived fit, as well as a separate analysis investigating the perceived similarity between examples.

Study participants were presented with a pairwise comparison task and asked to choose the audio-visual example which exhibited the strongest fit between elements in the visual modality and elements in the auditory modality. The examples were short video files showing humanoid avatar motion in a game-like rendered space (see Figure 1). For each example, one attribute of avatar motion was mapped to one attribute of sound and used to procedurally generate an audio track. The sound for each example was generated by sending a given parameter of motion using Open Sound Control output from UDKOSC [10] to a real-time synthesis process running in Supercollider. 480 examples across 861 randomly-ordered pairings were presented, representing each unique combination of 3 attributes of motion, 3 attributes of sound, 2 directional mapping schemata and 1 instrument type. The three attributes of motion tracked were actor speed, height in coordinate space and degree of rotation. The three attributes of sound modulated were *frequency* or *pitch*, *amplitude* or *volume* and *breath pressure*, presenting as a timbral shift in tone color for the physically-modeled clarinet instrument used for each example.

3.2 Study Participants

219 unique subjects participated in this study by selecting the study as an Amazon "Human Intelligence Task" (HIT) within the Mechanical Turk online interface. To present the study to a diverse group of participants while still retaining a high level of accuracy and validity, Workers were required to have achieved a previous HIT Approval Rate (or percentage of approved HITs) greater than or equal to 90% across all previously submitted tasks. To mitigate potential language issues, Workers were limited to those users whose locations (as verified through Amazon's billing and payments system) were determined to be in the United States. On average, participants completed approximately 28 HITs with a maximum per-participant HIT count of 392. The entire set of 6027 HITs was processed in less than eight hours with an average time of 1 minute 49 seconds for each assignment.

4. CROSSMODAL DATASET

To examine potential crossmodal correlations between parameters of motion and parameters of sound, an audio-visual dataset consisting of musically sonified avatar interactions was recorded, using UDKOSC to both control avatar motion and record parameter data. Attributes of motion including avatar speed, rotation and height were mapped to parameters of a physically-modeled clarinet. Motion attribute data was linearly scaled for each mapped instrument parameter, so that a noticeable change in the parameter would be experienced by subjects viewing and listening to the examples. The attributes of sound driven by motion data include Frequency, Breath Pressure and Amplitude.

Six examples of avatar motion were recorded depicting the same humanoid avatar running in various patterns across a simple room. The primary attributes of motion exhibited in these examples were speed, rotation, and coordinate height. The scene was lit in such a way as to show depth of field in an otherwise feature-sparse environment. In the center of the space was a simple pedestal construct, similarly used to establish depth of field. For all visual examples used in this study, a static camera position was chosen to frame the entire sequence of motion without changing a viewer's position or angle of perspective.

5. DATA ANALYSIS

By framing participants' subjective preference of crossmodal media examples as a discrete choice model, this study was designed to determine both the rank of preference for each example across the entire participant set as well as the relative effect of individual attributes and attribute pairs. Rank of preference and attribute contributions can be determined using a binomial choice model such at those proposed by Bradley and Terry [11, 12]. Data analysis was conducted using the R statistical programming language [13, 14] and the BradleyTerry2 package [15].

5.1 Bradley-Terry Model

The Bradley-Terry model (BTm) provides a method of extracting associative rankings from binomial choice datasets. Commonly used for the evaluation of multiple-participant binomial competitions such as baseball seasons or chess tournaments, Bradley-Terry has been used to model pairwise comparison tasks in fields ranging from genetics to marketing to election results [16]. By presenting examples to be compared as 'competitors' in a matched pairwise comparison task or 'contest', the Bradley-Terry model proposes a logit model for paired evaluations, capable of ranking examples based on their ability to 'win' a given comparison. One advantage of the Bradley-Terry model when compared to simpler averaging or mean comparisons is that the BTm factors the relative strength of competitors when calculating results, so a 'victory' in a pairwise comparison over a strong competitor counts more when calculating ranking scores than a victory over a weak competitor.

Essentially for any pairwise comparison or contest, the Bradley-Terry model assumes for any two paired 'players', i and j $(i, j \in \{1, \ldots, K\})$, the odds that player i beats player j can be represented as α_i/α_j, where α_i and α_j are positively-valued parameters representing 'ability'.

To express the Bradley-Terry model using a logit-linear form we can say

$$[(i \text{ beats } j)] = \lambda_i - \lambda_j, \tag{1}$$

where $\lambda_i = \log \alpha_i$ for all i. Therefore if we assume independence for all contests, maximum likelihood can estimate parameters $\{\lambda_i\}$[15].

The Bradley-Terry model can rank ability for explanatory variables or 'predictors' that can be found in each example, effectively allowing the algorithm to assess which component attributes of motion and sound in our dataset exhibit the most or least predictive power in participants' assessments of relative fit. And while there do exist extended techniques to factor 'ties' into the Bradley-Terry model [17], for this study ties were not allowed. Participants were allowed to choose 'Same' in their pairwise comparison task; these results were subsequently excluded from the Bradley-Terry calculation. In total, 1,487 results were marked as 'Same' and were not processed by the Bradley-Terry model.

6. CROSSMODAL ATTRIBUTE DESCRIPTORS

At the heart of this study are the crossmodal relationships and measurable perceptual coherence between attributes of motion in virtual space and attributes of sound. Attributes

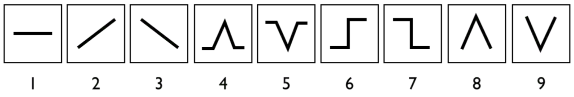

Figure 2: Contour shapes exhibited in the full dataset. Contours 2-8 were sonified in this user study.

from each modality mirror the parameters used when creating the crossmodal dataset and were defined as simple directional binomial descriptors of each parameter of motion and each parameter of sound. Directional attributes marking the type and direction of a generalized trait (such as ”increase in speed”) were used rather than continuous values of attributes over time. Composite attribute descriptors, showing the coexistence of attributes from both modalities were used to search for trends related to the pairing of crossmodal parameters. During analysis, the impact of each individual attribute and crossmodal attribute pair on the aggregate perceived fit was determined across all examples in the set. For instance, video excerpts showing an increase in speed mapped directly to frequency would exhibit the composite attribute “Positive Speed, Positive Frequency” while an excerpt showing an increase in rotation mapped inversely to breath pressure would exhibit the composite attribute “Positive Rotation, Negative Breath Pressure”. In this manner the directionality of each mapping can be examined both in the context of the direction of its original source motion as well as in the context of the direct or inverse mapping schema.

7. RESULTS

7.1 Directional attribute pairs as predictors of fit

Each crossmodal pairing represented by the examples used in this study can be described as a pair of attributes from both the visual and auditory modality with an associated parameter-data direction. The directionality of parameter data from each modality, or whether a given parameter increases or decreases, gives us four paired states to consider, i.e. an increase in both attributes, a decrease in both attributes, or one increase paired with a decrease. By looking at these pairings as attributes themselves, we can plot the mean perceived fit for each state, for each crossmodal attribute pairing. Results detailing the perceived fit of directional attribute pairs can be found in [8].

7.2 Contour Matching

One particularly interesting way of looking at the parameter data used in this study involves the reduction of each recorded motion attribute and generated sound attribute to simple parameter contours based upon a hypothesis that examples exhibiting matched contours between motion and sound would exhibit a greater perceived fit. Figure 2 shows nine simple contour shapes that are exhibited in the full dataset, with contours 2-9 exhibited in parameters sonified in this user study. For example, contour #2 shows a linearly increasing parameter value, such as would be exhibited by the speed parameter during a linear acceleration. During a direct mapping of acceleration to frequency, the contour exhibited by frequency would also be #2, while if an inverse mapping were to be used, the contour exhibited would instead be #3.

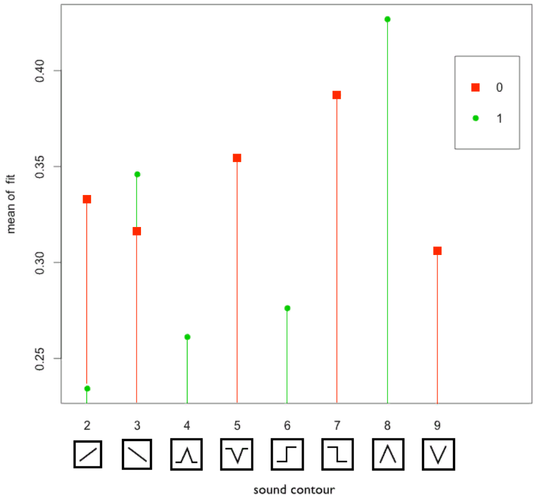

Figure 3: For examples exhibiting sound contours 2-9, the mean of fit for both matched (1/green circle) and unmatched (0/red square) contours are displayed.

7.2.1 *Mean of Fit Grouped by Contour*

Looking at the calculated mean of fit for each contour shape in Figure 3 we can see the following patterns of behavior:

- For contours 2 and 3, linear increase and decrease, we can see a discrepancy between the mean fit for matched contour vs. unmatched contour. For linear increases, there is a much higher mean of fit for unmatched contours than for matched contours. For linear decreases, the mean fit values are approximately equal for matched and unmatched contours. Note that examples exhibiting contours 2 and 3 can display both linear increases and decreases such as direct and inverse mappings of speed for the acceleration and deceleration examples, and direct and inverse mappings of rotation for the continuous curve example.

- In contours 4 and 5, a sharp parameter increase followed by a decrease shows a strong mean fit preference for the unmatched contour. This is exhibited for inverse mappings of the height parameter on the jump event motion example. Similarly, contours 6 and 7, found when mapping rotation on the discrete turn example, show a preference for the unmatched contour.

- Contours 8 and 9 show a mean fit preference for a continuous increase followed by a decrease, as exhibited when mapping rotation for a circle event. It should be noted that in these cases, rotation was judged to be “positive” when the actor turned left, or away from the

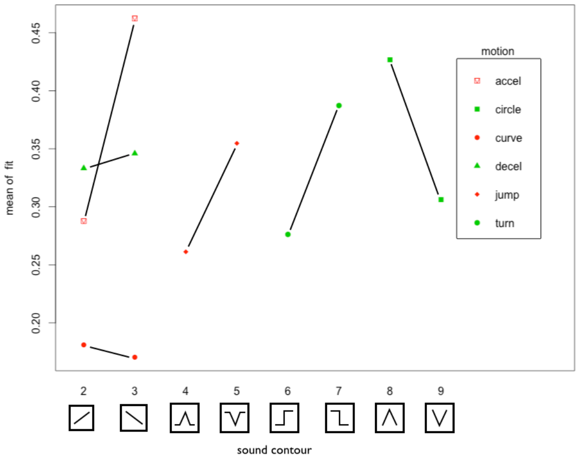

Figure 4: Mean of fit for each sound contour grouped by motion type is displayed.

camera location. If that choice had been reversed, then the pattern exhibited in contours 8 and 9 would match the same patterns evident when looking at contour pairs 4 and 5 or 6 and 7.

- Looking across contours 2 and 3, if the matched contour value for 2 is viewed as paired with the unmatched contour value in 3, we again see the same mean preference for unmatched contours as is evident in contours 4-9. In this case, the similar fit for unmatched contour 2 and matched contour 3 can then be seen as anomalous, in that there is no clear preference for unmatched contour.

7.2.2 *Mean of Fit Grouped by Motion Type*

Figure 4 groups means of fit for each contour shape by motion type. A few key points are summarized below:

- While acceleration, or an increase in speed, shows a marked preference for an inverse contour (3) over a direct contour (2), both deceleration and curve motions show little variation in their means of fit between contours 2 and 3.

- Jump and discrete turn motions (respectively 4,5 and 6,7) display stronger mean of fits for inverse mappings (5,7) than for direct mappings (4,6).

- Circle motions display stronger mean of fit for direct mappings (8) than for inverse mappings (9).

7.3 Symmetrical Pairings

Following Eitan and Granot, symmetrical example pairings can be described as example pairs in which both the directional attributes of motion and the directional attributes of sound are inverted or diametrically opposed. Figure 7 shows each possible symmetrical example pairing exhibiting changes for the motion attribute of speed. As this dataset contains motion examples exhibiting increasing and decreasing speed (acceleration and deceleration) symmetrical pairings can be examined for mappings to frequency, rotation and breath pressure. The difference in rank for each member of the pair can be seen in column Δi while the difference in perceived fit from column Estimate can be seen in column Δsym

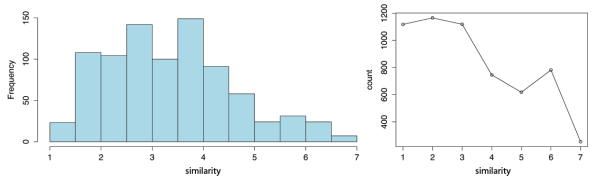

Figure 5: Similarity histogram and plots of example counts.On the left, a similarity histogram displays the average score distribution across each of 861 unique example pairings. On the right, a plot showing total example counts at each similarity level

Two sets of pairings shown in Figure 7 exhibit symmetrical tendencies, that is, their Δi and Δsym values are both extremely low. However the perceived fit (as seen in the Estimate column) is fairly low for both pairings with only one pairing showing significance. For these pairings to exhibit true symmetrical tendencies, not only should their perceived fits be approximately the same but they should also be fairly high.

7.4 Perceived Similarity

The perceived similarity between crossmodal examples presented to study participants was recorded as a user-chosen integer value. Participants were presented with the following question: ‘Please select a value from 1-7 to rank how similar the audio and video in the above videos are (where “1” means the two examples are completely different and “7” means they seem exactly the same).’

The majority of rated pairs were judged to be relatively low scoring or not similar. Looking at the top twenty-five most similar pairings from the example set, the following key points can be seen:

- 20 of the top 25 example pairings ranked for similarity were motion similar, meaning they shared the same motion sequence with a different sonification or sonification mapping direction.

- 2 of the top 25 example pairings were exactly sound similar, meaning the generated sound result came from the same mapping contour and parameter range.

- 4 of the top 25 example pairings were inversely sound similar, meaning the examples were generated from inverse mappings of the same contour and parameter range.

- 8 of the top 25 were acceleration/deceleration pairings sharing similar or inverse contours.

- Breath pressure examples comprise 9 of the top 10 examples, while breath pressure paired with gain make up 29 of the top 50 examples.

To get an overall feel for the influence of individual attributes of both sound and motion on perceived similarity, Figure 6 displays the average similarity rating for each attribute. Breath pressure and amplitude both show high relative similarity averages from the sound modality while attributes of motion acceleration and deceleration show high relative similarity. The lowest average similarity values can be seen from the circle motion and frequency sound attribute, both exhibiting a significantly lower average similarity rating than all other attributes.

Attribute	Avg. Similarity
Breath Pressure	3.568717354
Amplitude	3.56033717
Deceleration	3.470830762
Acceleration	3.455576169
Speed	3.392429793
Height	3.300762732
Turn	3.28860029
Curve	3.282622143
Jump	3.179633362
Rotation	3.167114366
Circle	2.971243035
Frequency	2.777136944

Figure 6: Average similarity for each attribute of motion and sound.

8. DISCUSSION

These analyses of subjects' perceived fit of image and sound components provide some possible insight into the types of crossmodal correlations we as humans tend to feel are more coherent. By making use of the Amazon Mechanical Turk service, the subject pool for this study was extremely large and avoided common limitations found in academic user studies that draw small subject counts from extremely homogeneous environments. The speed and scale of the Mechanical Turk service also allowed for incredibly quick turn-around and validation of user data, affording the study coordinator the ability to rapidly iterate modifications to the formatting of the study presentation scripts and dataset. The use of the Bradley-Terry model for the statistical analysis of user results reframed the ranking of perceived fit from a scalar ranking issue to a more manageable pairwise comparison task. The Bradley-Terry model also presented a methodology for assessing the relative impact of individual and paired attributes of motion and sound on subjects perceived fit across the entire study sample set.

8.1 Assessing Perceived Coherence and Fit

While the results answering the first primary goal of this exploratory study, namely "Which examples exhibited the strongest fit across all participants?" as well as ranked example results for the entire study set are detailed in [8], for context it is worth briefly discussing them here as well.

The study's top ranked result with a ranking of 0.64579 was a discrete left turn inversely mapping rotation to frequency, effectively causing a decrease in frequency during the turn event. Its inverse mapping - that is a discrete left turn directly mapping rotation to frequency - exhibits an increase in frequency and is ranked quite lowly (rank 31). Only one other turn event was ranked in the top 50% of examples (rank 6, a direct mapping of rotation to breath pressure) suggesting that the turn event itself wasn't a strong predictor for the high rank. With regards to specific directional attributes, the decrease in frequency exhibited by this example does correspond with the relatively strong predictive ability of the frequency_decrease attribute.

8.2 Symmetrical Mappings

The role of symmetry in the perception of crossmodal relationships, or more specifically the conclusion that "musical-spatial analogies are often asymmetrical, as a musical change in one direction evokes a significantly stronger spatial analogy than its opposite" was explored by Eitan and Granot [18]. Their study was based in analogy, with subjects visualizing and describing attributes of multi-dimensional motion when prompted by musical auditory stimuli.

In their initial hypothesis of "Symmetry of associative space", Eitan and Granot define crossmodal symmetry:

> Other things being equal, diametrically opposed musical processes <m,−m> would evoke diametrically opposed kinetic processes <k, −k>. In experimental terms: a listener who associates a musical stimulus m (e.g., a crescendo) with a kinetic quality k (e.g., a spatial ascent) would associate the inverse stimulus −m (e.g., diminuendo) with the opposite kinetic quality −k (e.g., descent).

Putting this hypothesis into terms that better relate to the experiment presented in this work, where one example exhibits a high-level of coherence or fit for a directional motion attribute (e.g. an increase in speed) when directly mapped to a directional sound attribute (e.g. an increase in frequency), the example exhibiting inverse directions for both motion and sound attributes (e.g. a decrease in speed mapped to a decrease in frequency) would also exhibit a high-level of coherence or fit. Results from asymmetrical pairings of speed examples show only two example pairs exhibiting symmetrical tendencies for which the perceived coherence for both pairings is low. The other four example pairs demonstrate weak asymmetrical tendencies but the perceived fit difference or Δsym for each is not strong.

i	Motion	Estimate	Std. Error	z value	Pr(>|z|)	∆sym	∆i	
27	acceleration_speed_frequency	0.23975	0.18791	1.276	0.201997	0.00103	1	
28	deceleration_speed_frequency	0.23872	0.18816	1.269	0.204544	0.00103	1	
18	acceleration_speed_amplitude	0.36257	0.19695	1.841	0.065632	0.01437	1	.
19	deceleration_speed_amplitude	0.3482	0.19436	1.792	0.073209	0.01437	1	.
7	acceleration_speed_breathpressure_inverse	0.46355	0.19417	2.387	0.016968	0.069	8	*
15	deceleration_speed_breathpressure_inverse	0.39455	0.19375	2.036	0.041716	0.069	8	*
3	acceleration_speed_frequency_inverse	0.53207	0.18672	2.85	0.004378	0.15891	14	**
17	deceleration_speed_frequency_inverse	0.37316	0.18708	1.995	0.046075	0.15891	14	*
16	acceleration_speed_amplitude_inverse	0.39134	0.19084	2.051	0.040307	0.15941	14	*
30	deceleration_speed_amplitude_inverse	0.23193	0.19346	1.199	0.230595	0.15941	14	
8	deceleration_speed_breathpressure	0.45092	0.19677	2.292	0.021926	0.19024	17	*
25	acceleration_speed_breathpressure	0.26068	0.18795	1.387	0.165453	0.19024	17	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	Estimate	Std. Error	z value	Pr(>|z|)
Std. Dev.	0.05218	0.04974	1.049	0.294

Figure 7: Bradley-Terry model results showing groups of paired symmetries for speed and each sound attribute. Column ∆i represents the difference in index for each paired symmetry and ∆sym shows the difference in Estimate for each member of the pair.

9. FUTURE CONSIDERATIONS

With the approaches and methodologies detailed in this study there are a number of considerations that should be mentioned and considered for future work.

- To fit this experiment into a simple Mechanical Turk project template, there was no attempt to engage users in a minimum (or maximum) number of HITs, rather they were permitted to submit as many or as few as possible. This introduces the possibility of a small group of subjects who processed many HITs exerting more influence over the results than subjects who processed less. Due to the scale of unique combinations presented, a straight-forward within-subjects design would not have been possible without inducing significant fatigue and likely carryover effects. Similar issues would complicate a standard between-subjects design. A middle-ground approach consisting of grouped stimulus pairs and limited group sizes could be investigated in the future to limit such individual impacts.

- While some attributes exhibited in the crossmodal dataset such as speed were addressed by multiple examples (acceleration, deceleration, jump event), others such as height were only exhibited by one example. Additional examples in the dataset should be created to gauge the influence of these attributes from multiple directions and sources. For height, simple examples showing an avatar walking up and down a slope or jumping up to a ledge would be useful additions.

- For the sake of consistency in the subject's viewpoint, each example in the dataset was created with avatars moving from screen-left to screen-right. The inverse direction showing motion from screen-right to screen-left should be added to take into account the perceived differences in general directional movement. Similarly all turn events showed the avatar turning left; examples showing turns to the right can also be added.

- Rotation examples were all generated using a polar mapping where parameter data increases linearly until rotation hits 180 degrees, then decreases until it reaches 360 or 0 degrees. While this mapping schema takes the human understanding of "forwards" and "backwards" into consideration, it would be interesting to also explore a simple linear mapping for rotation.

- Rotation examples mapped turns to the left as increases in rotation. The mapping of a left turn to an "increase" in rotation was purely arbitrary and could have easily mapped a left turn to a decrease in rotation. In studies where multiple camera views are explored, one possible mapping of interest would cause rotations towards the camera to cause parameter increases, while rotations away from the camera would cause corresponding decreases.

- The use of a human-like avatar was intended to mimick similar avatars commonly used in commercial computer games. While the humanoid paradigm is indeed common, the motion of limbs and the inherent animation of the skeletal mesh could potentially be a distraction when tracking gross motion contours in the environment. One solution would be to create the same motion examples using generic block shapes without extraneous limb motion.

- As mentioned previously, the use of confound videos as markers signifying a subject's attention to the task at hand proved to be less successful than intended. This was in part due to the limited set of two confound videos and the ease at which workers could select the "correct" HIT without even watching both videos. To make the confound videos more accurate predictors of user attention, a larger set of confound videos should be used to prevent this behavior.

- When subjects were permitted to choose 'Same' in the primary example comparison task, 1,487 results were marked in this way and subsequently excluded from the Bradley-Terry model calculations. If users had been presented with a forced-choice between example 1 or example 2 these results would have contributed to the BTm results. Davidson did however propose an extension to the Bradley-Terry model that can accomodate the existence of 'tie' results [19]. One future task will be to compare the current BTm results with results using the Davidson extension.

10. REFERENCES

[1] C. Francois and D. Schn, "Musical Expertise Boosts Implicit Learning of Both Musical and Linguistic Structures," *Cerebral Cortex*, 2011. [Online]. Available: http://cercor.oxfordjournals.org/content/early/2011/05/13/cercor.bhr022.abstract

[2] B. Tillmann, "Music and Language Perception: Expectations, Structural Integration, and Cognitive Sequencing," *Topics in Cognitive Science*, pp. 1–17, 2012.

[3] C. L. Krumhansl, *Cognitive foundations of musical pitch*. New York: Oxford University Press, 1990.

[4] Thatgamecompany, "Journey," [Digital Download], 2012.

[5] Harmonix, "Rock Band," 2007. [Online]. Available: http://www.rockband.com

[6] G. Wang, "Designing Smule's iPhone Ocarina," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, Pittsburgh, June 2009.

[7] R. Hamilton, J. Smith, and G. Wang, "Social Composition: Musical Data Systems for Expressive Mobile Music," *Leonardo Music Journal*, vol. 21, 2011.

[8] R. Hamilton, "Perceptual Coherence as an Analytic for Procedural Music and Audio Mappings in Virtual Space," in *Proceedings of 2015 IEEE Virtual Reality Conference*, Arles, France, 2015.

[9] J. Oh and G. Wang, "Evaluating crowdsourcing through amazon mechanical turk as a technique for conducting music perception experiments," in *...Conference on Music Perception ...*, E. Cambouropoulos, C. Tsougras, P. Mavromatis, and K. Pastiadis, Eds., Thessaloniki, Greece, 2012, pp. 738–743. [Online]. Available: http://icmpc-escom2012.web.auth.gr/sites/default/files/papers/738_Proc.pdf

[10] R. Hamilton, "UDKOSC: An immersive musical environment," in *Proceedings of the International Computer Music Conference*, Huddersfield, UK, August 2011, pp. 717–720.

[11] B. R.A. and T. M.E., "Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons," *Biometrika*, vol. 39, pp. 324–45, 1952.

[12] A. Agresti, *Categorical Data Analysis. 2nd edition*. San Francisco: John Wiley & Sons, 2002.

[13] R. Ihaka and R. Gentleman, "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, 1996.

[14] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org

[15] H. Turner and D. Firth, "Bradley-Terry Models in R: The BradleyTerry2 Package," *Journal of Statistical Software*, vol. 48, no. 9, pp. 1–21, 2012. [Online]. Available: http://www.jstatsoft.org/v48/i09/

[16] P. J. Loewen, D. Rubenson, and A. Spirling, "Testing the power of arguments in referendums: A Bradley-Terry approach," *Electoral Studies*, vol. 31, no. 1, pp. 212–221, March 2012.

[17] P. Rao and L. Kupper, "Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model," *Journal of the American Statistical Association*, vol. 62, pp. 194–204, 1967.

[18] Z. Eitan and R. Y. Granot, "How Music Moves: Musical Parameters and Listeners' Images of Motion," *Music Perception: An Interdisciplinary Journal*, vol. 23, no. 3, pp. 221–248, February 2006. [Online]. Available: http://www.jstor.org/stable/10.1525/mp.2006.23.3.221

[19] R. R. Davidson, "On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments," Kentucky, March 1970.