Perceptual Coherence as an Analytic for Procedural Music and Audio Mappings in Virtual Space

Rob Hamilton*
Center for Computer Research in Music and Acoustics
Stanford University

ABSTRACT

Real-time data generated by virtual actors and their mediated interactions in simulated space can be repurposed to dynamically generate sound and music. Procedural audio and music systems afford interaction designers, composers and sound artists the opportunity to create tight couplings between the visual and auditory modalities. Designing procedural mapping schemata can become problematic when players or observers are presented with audio-visual events within novel environments wherein the validity of their own prior knowledge and learned expectations about sound, image and interactivity are put into question. This paper presents the results of a user-study measuring users' perceptions of audio-visual crossmodal correspondences between low-level attributes of motion and sound. Study results were analyzed using the Bradley-Terry statistical model, effectively calculating the relative contribution of each crossmodal attribute within each attribute pairing to the perceived coherence or 'fit' between audio and visual data.

Index Terms: H.5.5 [Sound and Music Computing]: Modeling—Signal analysis, synthesis, and processing; G.3. [Probability and Statistics]: Experimental design—

1 INTRODUCTION

Humans are natural viewers and listeners. As such we perceive countless sound-generating interactions each and every day. Our expectations for how any given interaction should sound as well as our abilities to perceive the causalities behind sonic events are shaped by a combination of our own embodied actions, our observations of the surrounding world and our understandings of the physics of our universe. Having acted and perceived interactions in the past we have learned to reasonably anticipate the audible result of not-yet perceived interactions between similar or previously understood objects and actions. Through our knowledge of how forces act upon physical bodies we can generalize and to a certain extent predict how different events and interactions will sound, at least within the confines of our known world. Put another way, we all know that we know that certain kinds of objects will create certain kinds of sounds when acted upon in certain ways.

When presented with sonic events in a rendered virtual environment - a potentially novel reality within which observers' *a priori* and *a posteriori* knowledge about sound, image and interactivity are no longer necessarily valid - observers must draw conclusions about perceived sonic interactions without full benefit of their own knowledge or past experiences. Their internal sonic lexicons, developed by everyday physical interactions during a lifetime of observation and experience, are intrinsically based in the laws of the familiar physical world. Within a generated environment where the rules governing interactions and "reality" itself are subject to the whim of a software designer or developer, observers are often forced to look outside of their own experiences when forming associations between visual and sonic action.

One increasingly common interaction model that exists outside of our physical reality is formed through the linking of action, motion and gesture with processes capable of generating sound that is musical in form and function. 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 background 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 [15]. And for games based around musical paradigms themselves, gesture and motion in both virtual and real-world environments are routinely mapped to dynamic musical generating and modification processes [6, 7, 17, 5].

2 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. Sonifications of each avatar motion example were audible to subjects while viewing each video. Stimuli consisted of video captures recorded alongside real-time data streams of avatar coordinate motion and state data. Each 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 then created by attaching and synchronizing the sonfications to each video example.

Subjects using the Mechanical Turk online tasking platform [11] were asked to watch short two-video example sets of these 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 were defined by a combination of motion and sound descriptors, allowing for the statistical analysis of correlated motion/sound pairs. For each modal pair a weighted fit value was calculated and used to calculate relative rankings 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. Additional analyses not detailed in this paper 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.

3 RESEARCH QUESTIONS

In the context of this study, analysis was conducted with three primary goals or questions in mind.

*e-mail: rob@ccrma.stanford.edu

1. Which examples exhibited the strongest fit across all participants?

2. Which attributes were the most significant predictors of fit?

3. Which crossmodal attribute pairs were the most significant predictors of fit?

Additional questions were also considered, including the influence of matched crossmodal contour, cross-example similarity and previous results found by Eitan and Granot that supported asymmetrical relationships between perceived parameters of audio stimuli and visual stimuli [2].

4 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 created, using UDKOSC [4] to control avatar motion and record parameter data. Attributes of motion including avatar speed, rotation and height were mapped to parameters of a series of synthesized instruments, including a simple sine wave, white noise and 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.

Short motion sequences were scripted and streamed in real-time into UDKOSC. Visual output from UDKOSC was captured to file using a digital video capture card while OSC messages representing the avatar's motion and action were recorded as binary time-stamped packets embedded into YAML markup using OSCRecorder. Sonifications were recorded by playing back recorded YAML data with OSCPlayer and streaming that data into Supercollider. Audio files were recorded in Supercollider and subsequently combined with the recorded video footage using a FFMPEG batch process. Synchronization between video and audio examples was handled using an inline ChucK script to calculate the onset time of the recorded audio event using UAna [18] RMS analysis and then trimming the audio excerpt to begin at that onset time. A shell script running FFMPEG merged each audio and visual example and created both .mp4 and .webm formatted video files to accommodate HTML5 video playback across a variety of user web browsers.

Six examples of avatar motion were recorded depicting a 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. For each example detailed below, the units of speed used can be thought of as Unreal-Units/second. An Unreal-Unit (UU) is a constant unit of measurement in the game environment and represents approximately 0.02 meters in scale.

5 Study Procedures

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. For each example, one attribute of avatar motion was mapped to one attribute of sound and used to procedurally generate an audio track. In total, 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 for this study 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-modelled clarinet instrument used for each example. The clarinet instrument was chosen due to the perceptible timbral shifts created when manipulating that instrument's breath pressure parameter.

5.1 Amazon Mechanical Turk

The study was carried out online, within the Amazon Mechanical Turk online service, in which participants are paid to carry out Human Intelligence Tasks (HITs) [10, 11]. An html page was prepared using Amazon's web survey templates coupled with custom javascript tracking and video playback code. Each video example was rendered as both .mp4 and .webm compressed video files and stored on CCRMA's webserver. To build a sample set for each Turk experiment, a comma-separated .csv file of 861 unique video filename pairs was uploaded to the Amazon system with one pairing passed into the web-form for each session of the experiment. Response and session data - including not only subjects' selected form answer fields but also a time duration for the task and boolean tracking variables showing playback counts for each video example - were retrieved using Amazon's Turk web toolkit. In total 6027 HITs were processed, with each of the 861 pairings viewed by seven different participants.

5.2 Participants

219 unique subjects participated in this study for which each were paid $0.09 for the successful and valid completion of each HIT. 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 and an average pay rate of approximately $2.97 per hour.

5.3 Study Format

Participants who qualified for the study and who chose this study's HIT from the Mechanical Turk Available HIT listings were presented with an html page displaying two sets of video pairs for pairwise comparison and three forced-choice questions in the form of html radio-buttons. The videos were presented using a side-by-side layout, with their respective order of presentation randomly determined for each individual presentation. The first set of videos were the actual sonified motion examples. The second set of videos consisted of confound examples, designed to determine if participants were properly attending to each video watched or just clicking through the study as quickly as possible. Participants were instructed to watch and listen to each video and, using the radio buttons, select the video in which the visual and auditory content exhibited the best 'fit'. There also existed the option to choose "Same" if participants believed the fit of each video was approximately equal. One additional question accompanied the first set of videos asking participants to rank the similarity between excerpts on a scale from one to seven, where one represented no similarity and seven represented a high level of similarity.

Figure 1: Avatar motion sequences clockwise from top left: A) forward acceleration, B) continuous left turn, C) forward jump, D) full left circle, E) discrete left turn, F) forward deceleration.

5.4 Data Validation

The Amazon Mechanical Turk is a for-pay service and as such there exists the possibility that participants taking part in any HIT might prioritize speed of completion over accuracy. To identify participants who may have not correctly completed the requested tasks, a series of validation techniques were used in this study. Result data from each HIT was audited by the study coordinator and any results which did not meet specific validation requirements were flagged and subsequently discarded.

Click counting The primary task of these HITs required the complete viewing of each short video example. If the videos were not watched, the HIT could not be considered valid. One potential technique would make use of a simple form validation which would disable the form's "Submit" button until each video had been viewed. However, since participants could work on multiple HITs, those participants trying to game the system would quickly learn how to click each video without necessarily attending to its content or engaging the required task. Instead, by simply tracking the number of times each video's "Play" button was clicked by users and passing that value to the Mechanical Turk engine, that play count was displayed in the results for the given task, allowing the study coordinator to flag any cases where each video had not been watched.

Task Duration If a minimum time threshold for the duration of the entire task was not exceeded, then the assumption was made that participants could not have watched the required videos in their entirety. Submitted HITs that failed to exceed the minimum time threshold had their results flagged and subsequently discarded. Across the entire 6057 HIT response set, the average work time duration was 109.32 seconds. A work time duration of 23 seconds or less was used to filter bad results from the subject set.

Confound Videos A set of confound videos was viewed by each participant as a second pairwise ranking task with the intention of determining whether or not participants were in fact attending to the video content or instead simply clicking through the task as quickly as possible. One confound was designed to clearly exhibit a stronger fit than the other example so that users who selected the weaker fit choice could be flagged as potentially not paying attention. This was accomplished by delaying the sonification in the "poor" fit confound by approximately three seconds, creating one video that clearly exhibited a weaker fit between audio and visual modalities.

Each of the two confound examples consisted of the same visual recording of an avatar moving forward while rotating to the right, using a static camera angle positioned above and behind the avatar's location of origin. This motion example as well as this particular camera angle were not used for any example in the primary set. For the confound videos' sonification, the avatar's rotation was mapped directly to frequency.

Disqualified Trials During the manual validation stage, in which results submitted by users were examined by the study coordinator, 32 HITs were rejected based on null click count values for the two primary videos. As this stage of validation took place before approval and payment of each HIT, all of these 32 disqualified videos were re-introduced to the pool of Turk users and were all subsequently re-processed by new subjects.

357 HITs were flagged for not exceeding the minimum task duration time threshold of 23 seconds or less. These results were removed from the study after approval and payment had already been made and were subsequently not re-processed. Upon closer inspection, the majority of these 357 disqualified trials were submitted by a small group of repeat offenders who fit the profile of Turk users who were intentionally not properly carrying out each HIT.

The use of confound videos as automatic flags for the invalidation of a subject's results proved to be less clear than originally intended. Since the confound videos were exactly the same for each HIT (though their order was randomly generated), users who processed multiple HITs soon realized that they only needed to view one HIT to determine which was "correct". This conclusion was supported by comments submitted by subjects. Therefore missing confound click counts was not a factor used to automatically disqualify any trials. Instead, confound results were used during manual review of potentially invalid results as an additional determining factor by the study coordinator.

5.5 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 from each modality mirror the parameters used when creating the crossmodal dataset. For this study, attributes 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.

Directional Attributes An analysis of the impact of whether a crossmodal example exhibits a single attribute of motion or sound is not a particularly useful one, as each low level attribute can be said to exist within every example. For instance, all motions exhibit some level of speed, rotation and height at any given moment, as do all sonifications created using a physically-modeled clarinet exhibit some level of frequency, amplitude and breath pressure. Instead it is the directional delta or change of any one of these base attributes that can then be tracked across examples and across the entire study.

Composite Attribute Pairings Of primary interest in this study was the furthering of our understanding about which *pairings* of crossmodal attributes significantly contribute to participants' perceived fit of crossmodal media examples. These pairings represent

basic mapping schemata that themselves form the basis of more complex and artistic musical sonifications. Each video example in the sample set was tagged with composite attribute descriptors which allowed the ranking of example fits based upon not only the single attributes which were exhibited but also on which attribute pairings were being exhibited. 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.

6 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 [13, 1]. Data analysis was conducted using the R statistical programming language [8, 12] using the BradleyTerry2 package [16].

6.1 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 [9]. 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 $\{\lambda_i\}$[16].

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 [14], 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.

7 Results

This section details the most noteworthy results derived from the Bradley-Terry analysis with primary attention given to the ranked perceived fit or coherence for the crossmodal video examples, core attributes and attribute pairings.

	Estimate	Std. Error	z value	Pr(>|z|)	
turn_rotation_freq_inverse	0.6458	0.1940	3.33	0.0009	***
circle_rotation_breathpressure	0.5960	0.1935	3.08	0.0021	**
accel_speed_freq_inverse	0.5321	0.1867	2.85	0.0044	**
jump_speed_freq_inverse	0.5027	0.1909	2.63	0.0085	**
jump_speed_amplitude	0.4727	0.1930	2.45	0.0143	*

Table 1: Top five examples ordered by ranked fit.

7.1 Ranked Fit

42 unique crossmodal examples were used in this study. Table 1 displays the five examples exhibiting both the highest perceived fit as well as the most significant results from the Bradley-Terry model for pairwise comparison across the entire dataset. These five examples represent:

1. A discrete left turn with rotation inversely mapped to frequency, generating a sharp decrease in frequency corresponding to the left turn event.
2. A circular path with rotation mapped directly to breath pressure, increasing then subsequently decreasing pressure around the axis of 180 degrees.
3. An acceleration with speed inversely mapped to frequency, resulting in a decrease in frequency.
4. A jump event with speed inversely mapped to frequency, causing frequency to decrease then subsequently increase.
5. A jump event with speed directly mapped to amplitude, causing an initial decrease in amplitude, followed by an increase.

A summary of these results include the following observations:

- Five of the top fifteen ranked examples were jump events
- Decreases in breath pressure resulting from mappings to increases or decreases in speed performed approximately equally well regardless of mapping direction.
- Accelerations mapped inversely outperformed all direct mappings of acceleration regardless of which paired sound attribute was exhibited.

7.2 Attributes as predictors of fit

When 'predictors' or explanatory variables are specified for each example being analyzed using a Bradley-Terry model, the estimated worth or predictive power of each variable can be determined. For this study, the predictors used are speed, rotation and height (motion), and frequency, amplitude and breath pressure (sound). Table 2 shows the estimated worth and significance of each directional attribute.

The results exhibiting the most significance and the strongest positive estimated worth are the directional motion attributes *frequency_decrease* and *speed_increase*, with estimates of 0.21220 and 0.11877 respectively. *height_increase* also shows a significant negative estimate of -0.26053. The relatively low estimates may suggest that single-modality attributes have a limited predictive ability

without also considering their paired crossmodal attribute counterparts. Results that contain estimates of 'NA' typically contain at least one parameter that has been set to zero.

	Estimate	Std. Error	z value	Pr(>|z|)	
frequency_decrease	0.21220	0.08963	2.367	0.0179	*
amplitude_decrease	0.12784	0.09039	1.414	0.1573	
breathpressure_increase	0.12253	0.09056	1.353	0.1761	
speed_increase	0.11877	0.07139	1.664	0.0962	.
breathpressure_decrease	0.09702	0.09082	1.068	0.2854	
amplitude_increase	0.09239	0.09114	1.014	0.3107	
speed_decrease	0.08322	0.07154	1.163	0.2447	
frequency_increase	0.02592	0.08944	0.290	0.7720	
height_increase	-0.26053	0.13625	-1.912	0.0559	.
height_decrease	*NA*	*NA*	*NA*	*NA*	
rotation_increase	*-0.08322*	*0.07154*	*-1.163*	*0.2447*	
rotation_decrease	*NA*	*NA*	*NA*	*NA*	

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

Table 2: Single-modality attributes and their respective predictive abilities from a Bradley-Terry analysis.

7.3 Directional attribute pairs as predictors of fit

Each crossmodal pairing represented by the examples used in this study can be described as a pairing 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 grouping these pairings as attributes themselves, we can plot the mean perceived fit for each state, for each crossmodal attribute pairing.

The interaction plots presented in Figure **??** display the relative variance of each crossmodal attribute pairing, displaying mean fit values (y-axis) against directional attribute pairings (x-axis). For each attribute pairing, "n" refers to a negative or "decreasing" parameter change-direction, while "p" refers to a positive parameter change-direction. Examples in which the increase in a motion parameter is in the same direction as the sound parameter can be said to have a correlated mapping direction. When attributes are not correlated, the mapping direction can be said to be inversely correlated.

7.3.1 Mean fit for Speed

A symmetrical relationship is seen between correlated attribute pairs where speed is mapped to frequency. Relatively low mean values for the **n_n** and **p_p** mappings are displayed, contrasting with higher mean values for the **p_n** and **n_p** mapping directions. The dotted line shows a relatively flat mid-range mean fit for all four examples when speed is mapped to breath pressure. No strong pattern or mean values are seen for mappings between speed and amplitude.

7.3.2 Mean Fit for Rotation

Examples mapping rotation to breath pressure show a symmetrical increase in mean fit for direct mappings and low mean fit values for each of the inversely correlated pairings. Relatively flat mean fit values can be seen for both frequency and amplitude mappings.

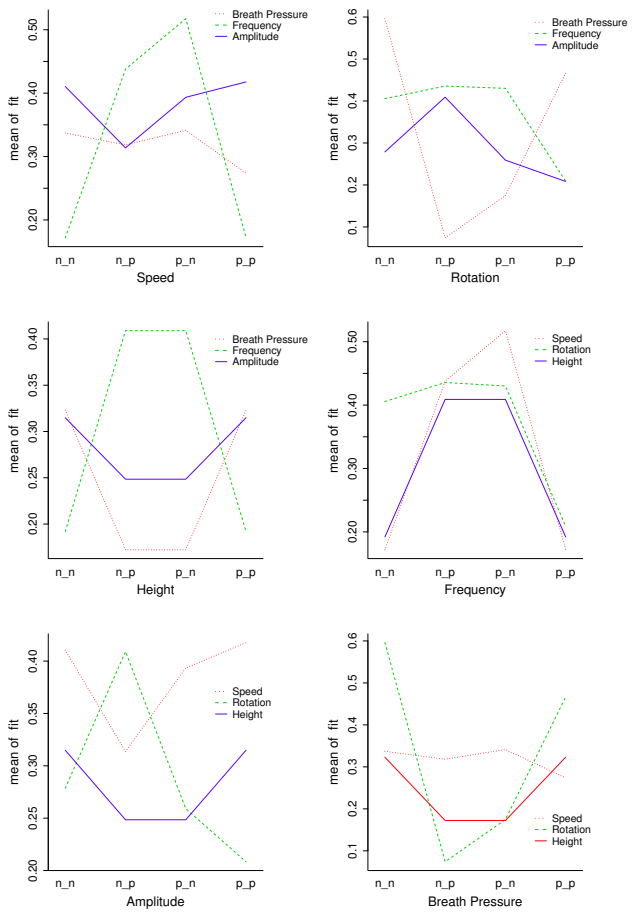

Figure 2: Interaction plots displaying mean of fit for motion parameters speed, rotation, height, and sound parameters frequency, amplitude and breath pressure, each plotted against the bimodal parameter directions positive (p) and negative (n).

7.3.3 Mean Fit for Height

The symmetrical nature of the one height example used in this study shows clearly symmetrical relationships for both correlated and uncorrelated mappings. Examples in which height was mapped to frequency exhibit strong mean preferences for uncorrelated mappings, with both **n_p** and **p_n** mappings scoring twice as high as **n_n** and **p_p** mappings. For both breath pressure and amplitude mappings, the inverse case can be seen where correlated mappings are higher than uncorrelated mappings.

7.3.4 Mean Fit for Frequency

Examples mapping attributes of motion to frequency show a preference for mappings in which mapping direction is inverted between motion and sound. These mappings are strongly symmetrical for both speed and height mappings. Frequency mappings with rotation exhibit the same decrease for positively mapped rotation to frequency increase, while showing little change in mean fit for inversely correlated mappings.

7.3.5 Mean Fit for Amplitude

Examples mapping attributes of motion to amplitude show the widest variance across all mappings, with few commonalities evident between mapping pairings with speed, rotation and height.

7.3.6 Mean Fit for Breath Pressure

Examples mapping attributes of motion to breath pressure were varied, with correlated mappings for rotation exhibiting the strongest overall fit. Both height and speed exhibited little variance regardless of whether mappings were correlated or not. Height and rotation exhibited strongly symmetrical results.

8 Discussion and Conclusions

The 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 homogenous environments. The speed and scale of the Mechanical Turk service also allowed for incredibly quick turnaround 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

The results answering the first primary goal of this exploratory study, namely "Which examples exhibited the strongest fit across all participants?" can be seen in Table 1 (the top five results). While these results themselves offer no explicit explanations for subjects' preferences, they do however suggest many interesting directions and new questions that can be addressed in subsequent more focused studies. Here let us consider some of the potential approaches that can be considered.

For example, 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. Looking to the directional attributes described in Table 2, the decrease in frequency exhibited by this example does correspond with the relatively strong predictive ability of the *frequency_decrease* attribute.

8.2 The Role of Perceptual Invariance

As a starting point for this body of research, the computer environments and interaction schemata used have retained clear perceptual and conceptual ties to the physics-based world in which we live. The avatar used in the crossmodal dataset displayed familiar human form and its motion was based in the visual and kinematic modalities intimately understood as belonging to our own physics-based reality. Subtly reinforcing the connection between these virtual spaces and the "real-world" has been an implicit reliance on certain perceptual invariants such as gravity, force and even the behaviors of light and shadow. The impact of these perceptual invariants has likely had the role of contextualizing the behaviors and interactions experienced in such virtual environments in terms that are at the same time familiar and potentially restrictive.

J. J. Gibson attributed our ability to perceive and understand objects during states of motion and change to the principle of *invariance*, labeling perceptual invariants as properties of "non-change that persists during change" [3]. In the context of multimodal audio-visual environments, the existence (or inexistence) of such invariants could affect the perceived causation and therefore the perceived coherence between crossmodal attributes. For future work in this direction, the use of less-familiar avatar forms, multiple camera viewpoints and more varied ecologies of multimodal interaction could provide insight into the role of invariants in our multimodal perception.

One hypothesis based in human experience and the perception of invariant physical phenomena such as Doppler shift and amplitude attenuation could suggest that as the 90 degree rotation directed the avatar away from the camera's point of view into the virtual "distance", subjects perceived the sound source as moving "away" and therefore a diminishing mapping schema was perceived as strongly coherent. If that were the case, it is interesting to note that the clearest example of such a diminishing mapping schema (i.e. "turn_rotation_amplitude_inverse", exhibiting a decrease in amplitude as virtual "distance" increases) was ranked as the twenty-third best fitting example. And the very similar "curve_rotation_frequency_inverse" example (a continuous curve inversely mapping rotation to frequency) which likewise exhibited a decrease in frequency as the avatar rotated and moved away from the camera location itself ranked quite poorly at rank 33.

References

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

[2] Z. Eitan and R. Y. Granot. How Music Moves: Musical Parameters and Listeners' Images of Motion. *Music Perception: An Interdisciplinary Journal*, 23(3):221–248, February 2006.

[3] J. J. Gibson. *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston, 1966.

[4] R. Hamilton. Udkosc: An immersive musical environment. In *Proceedings of the International Computer Music Conference*, pages 717–720, Huddersfield, UK, August 2011.

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

[6] Harmonix. Guitar hero, 2005.

[7] Harmonix. Rock band, 2007.

[8] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[9] P. J. Loewen, D. Rubenson, and A. Spirling. Testing the power of arguments in referendums: A Bradley-Terry approach. *Electoral Studies*, 31(1):212–221, March 2012.

[10] W. Mason and S. Suri. Conducting behavioral research on Amazon?s Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, 2011.

[11] J. Oh and G. Wang. Evaluating crowdsourcing through amazon mechanical turk as a technique for conducting music perception experiments. In E. Cambouropoulos, C. Tsougras, P. Mavromatis, and K. Pastiadis, editors, *... Conference on Music Perception ...*, pages 738–743, Thessaloniki, Greece, 2012.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[13] B. R.A. and T. M.E. Rank analysis of incomplete block designs i: The method of paired comparisons. *Biometrika*, 39:324–45, 1952.

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

[15] Thatgamecompany. Journey. [Digital Download], 2012.

[16] H. Turner and D. Firth. Bradley-Terry Models in R: The BradleyTerry2 Package. *Journal of Statistical Software*, 48(9):1–21, 2012.

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

[18] G. Wang, R. Fiebrink, and P. R. Cook. Combining Analysis and Synthesis in the ChucK Programming Language. In *Proceedings of the International Computer Music Conference*, Copenhagen, 2007.