Audio Engineering Society

Conference Paper

Presented at the Conference on
Audio for Virtual and Augmented Reality
2018 August 20 – 22, Redmond, WA, USA

Evaluation of Binaural Renderers: Multidimensional Sound Quality Assessment

Gregory Reardon[1], Andrea Genovese[1], Gabriel Zalles[1], Patrick Flanagan[2], and Agnieszka Roginska[1]

[1]*New York University, Music and Audio Research Lab, 35 W. 4th St., New York, NY 10012*
[2]*THX, 1255 Battery St, Suite 100, San Francisco, CA 94111*

Correspondence should be addressed to Patrick Flanagan (patrick@thx.com) or Agnieszka Roginska (roginska@nyu.edu)

ABSTRACT

A multi-phase subjective experiment evaluating six commercially available binaural audio renderers was carried out. This paper presents the methodology, evaluation criteria, and main findings of the tests which assessed perceived sound quality of the renderers. Subjects appraised a number of specific sound quality attributes - *timbral balance, clarity, naturalness, spaciousness*, and *dialogue intelligibility* - and ranked, in terms of preference, the renderers for a set of music and movie stimuli presented over headphones. Results indicated that differences between the perceived quality and preference for a renderer are discernible. Binaural renderer performance was also found to be highly content-dependent, with significant interactions between renderers and individual stimuli being found, making it difficult to determine an "optimal" renderer for all settings.

1 INTRODUCTION

Recent interest and advances in augmented reality (AR) and virtual reality (VR) technologies have highlighted the need for coherent and high-fidelity spatial audio. Audio plays a significant role in orienting the user to their 360° environment, providing information about the location of virtual objects outside the user's field of view and directing the user's attention. A number of different binaural audio technologies, known in this work as *binaural renderers*, have recently become commercially available for use in AR and VR applications. These renderers can also be used to generate immersive audio for more traditional music, movie, and computer game settings to significantly enhance the experience.

One of the key concerns of psychoacoustics research is to understand the auditory system with respect to sound localization and "higher-level" perceptual judgements about the quality and characteristics of the perceived auditory image. Psychoacoustic evaluation of commercial binaural renderers serves a primary purpose of gauging the variability of performance across a number of metrics. Further, because each renderer is distinct in its rendering methodologies (where methodology refers to the entire signal chain, including the specific Head-Related Transfer Functions [HRTFs] used for spatial-

ization), it also presents an opportunity to research the perceived sound quality space and its partition. Finally, this initial study provides an opportunity to examines the proposed methodology for the comprehensive evaluation of binaural renderers.

This work presents part of the results obtained from a larger three-phase experiment that was conducted on the performance of commercially available binaural renderers. It is beyond the scope of this study to identify the specific renderers tested and evaluate their internal mechanisms. The methodology presents a framework for a blind user-level assessment of available renderers for the interest of gauging the high-level performance over specific applications without necessarily having access to the internal software.

In the experiment, six different renderers were compared using a number of qualitative and quantitative metrics [1]. Phase I of the experiment was concerned with analyzing the prevalence of 3D sound localization errors - externalization, front/back and up/down confusions, and horizontal localization accuracy - for each of the renderers under test. The results on externalization, front/back and up/down confusions, are presented in [2], while the results of the median-plane localization test are found in [3]. Phase II of the experiment was concerned with evaluating specific sound quality attributes believed to be important for appraising spatial audio scenes. Phase III consisted of a forced choice ranking of the renderers in terms of *preference* and is interpreted, in the context of this work, as a global assessment of perceived sound quality. The results of both Phase II and III, along with subsequent joint analysis that looked at the correlations between spatial sound quality attributes and listener preference, are presented.

1.1 Spatial Audio Evaluation

Perceived spatial sound quality has been found to be comprised of distinct perceptual dimensions related to more specific sound qualities [4, 5]. The identification of spatial sound quality attributes has been a large subject of research. Most approaches combine some free verbalization method, whereby subjects are permitted to construct their own vocabulary to represent auditory sensations, with subsequent perceptual evaluation and statistical analyses [6, 7, 8, 9]. A number of different attribute sets have been proposed by various authors [4, 7, 10, 11]. Generally, there are three main classes of sound quality attributes - *timbral, spatial,* and *technical*. Timbral attributes relate to tone color, spatial attributes to the three-dimensional aspects of the audio scene, and technical attributes to distortion and artifact-related sensations. Some attributes are more global, like *naturalness*, and are often a combination of both timbral and spatial features. Thus the distinction between attribute classes in the above formulation is not sharp, but serves as a generic model of the sound quality space [12].

The statistical analyses of multidimensional sound quality data has a rich history. Many authors have explored the use of principal component analysis (PCA), principal factor analysis (FA), cluster analysis, correlation analysis, and regression analysis, often in conjunction with one another, to understand the sound quality space [10, 13, 14, 15, 16]. These tests give different perspectives on the data. In some of the experimental designs, preference is treated as a sound quality and assessed with the other direct attributes [17]. Other authors have investigated evaluating preference in a separate assessment from the other sound quality attributes. The goal of such work is then to learn the mapping from sound quality attributes to preference. For instance, Susini et al. [18] used a multidimensional scaling technique, in which ratings of dissimilarity between pairs of stimuli for each sound quality were assessed, to predict the probability of one sound being preferred over another. Zacharov and Koivuniemi [7] attempted to learn the preference mapping using a partial least squares regression. The sound quality attributes of different multichannel recording techniques (reproduced over loudspeakers) were measured in pairwise (on 100-point scales), while a single preference judgment (on a 200-point scale) was gathered for that same pair. *Movement, depth, broadness,* and *tone color* (and their interactions) were strong predictors of preference, while characteristics like *richness and hardness* did not contribute significantly to prediction.

Similar to this work, Guastavino and Katz [11] explored the interactions between reproduction methods (1D, 2D, and 3D loudspeaker reproduction) and stimuli (3 distinct soundscapes, two of which included musical elements) on overall preference of spatial audio content. An important finding was that the choice of preferred reproduction method had a statistically significant interaction with the type of content presented and therefore no universally optimal reproduction method could be determined. The authors also concluded that attributes

like *presence* and *readability* were important to listeners. Marins et al. [19] used a MUSHRA-type test for relating subjective perceptual changes on degraded multichannel audio, concluding that *timbral balance* was the main factor for basic audio quality. In [20], sound quality attributes, measured using a MUSHRA-based test protocol with stimuli presented over loudspeakers, were correlated with surround-sound quality using a multiple linear regression. Three clusters of attributes were selected - *timbre, space* and *defects* - the latter of the three being deemed the strongest predictor of overall sound quality judgements. While this work specifically focuses on a methodology for evaluating static binaural audio content over headphones, previous studies about multichannel spatial audio can help to understand the context of choosing and relating quality attributes to overall renderer preference, and interpret the role of stimuli and rating methodology in evaluating binaural renderers.

2 METHODOLOGY

2.1 Rendering Procedure, Stimuli, and Presentation

Six different binaural renderers were tested in the comparative study. These renderers are labelled 00 - 05. Three of the renderers (00, 01, and 05) use higher-order ambisonics (HOA) to spatialize content. Two of the renderers (03 and 04) use first-order ambisonics (FOA). The final renderer (02) uses direct virtualization (HRTF convolution/filtering). All of the ambisonics-based renderers use 3D ambisonics. Though each renderer has head-tracking capabilities in its native application, the experimental content was presented under a static condition and reproduced over headphones.

A total of six different surround-sound stimuli rendered for static binaural presentation were tested in Phases II and III[1] - three music and three movie stimuli. The "music" stimuli were short musical excerpts. These stimuli were recorded works cut to approximately 20 seconds in length. The stimuli were of varying style, one jazz, one wind quintet, and one symphonic orchestral work. The jazz piece was mixed for 5.0 surround-sound while the symphonic works were mixed for 9.0 surround-sound with height. The "movie" stimuli were excerpts taken from a 5.0 surround-sound mix of "*Star Wars: The Force Awakens*." These stimuli were no longer than 30 seconds and each included dialogue, music, and sound effects. For each stimuli, the individual channels were treated as independent virtual audio objects for processing by each renderer. These objects were placed at a distance of one meter from the listener in the auditory scene at azimuths and elevation corresponding to ITU-R guidelines for 5.0 and 9.0, respectively [21]. These channels were rendered to a single piece of static binaural content at 48 kHz sample rate and 24 bit depth without additional room information; all settings regarding room reverb and early reflections were turned off. All other renderer properties were set to their optimal settings (with matching sample rate and audio quality export settings).

Each subject was randomly assigned to either the "music" or "movie" condition; the condition for each subject was kept consistent throughout both phases in order to perform separate multivariate correlation analyses. The test was administered over circumaural headphones (Sennheiser HD-650) without additional equalization and in a soundproof booth (NYU Dolan Isolation Booth). Custom scripts were developed to run the experiment and collect data without experimenter intervention. A graphical user interface (GUI) was designed to allow subjects to play stimuli *ad libdum* (after a forced listening round), comment on specific trials, and indicate and submit their responses.

2.2 Phase II

Phase II was concerned with the evaluation of specific sound quality attributes. Subjects assigned to the music condition rated four sound quality attributes, while those assigned to the movie condition assessed five sound quality attributes. The descriptions of each of the attributes was inspired by previous literature [11, 17, 22]. Ultimately, the descriptors were defined as follows:

- **Timbral Balance**: This attribute describes how balanced (or colored) the different tone ranges of the sound appear to be.
- **Clarity**: This attribute describes whether the sound appear to be clear or muffled.
- **Naturalness**: This attribute describes whether the sound gives a realistic impression, as opposed to artificial.

[1] This section is concerned with Phase II and III of the larger methodology. For Phase I please refer to the previous documents about externalization, confusions and localization [2, 3].

- **Spaciousness**: This attribute describes how much the sound appears to surround you.
- **Dialogue Intelligibility** (movie stimuli only): This attribute describes the ease at which dialogue can be understood.

The description of each of these characteristics was provided to the subject before the experiment began. The subject completed twelve (music) or fifteen (movie) trials in this phase - one trial per characteristic per stimuli. In each trial a subject rated a single characteristic for each of the six renderers. The procedure was as follows: subjects played the first renderer, were forced to listen to the clip in its entirety, and then rate the characteristics on a 5-point scale, with 1 being the worst, and 5 the best. The subject was then free to move to the next renderer. After all six renderers had been preliminarily rated, the subject was free to replay any of the renderers, for any length of time, to refine their ratings. Subjects were free to use any range of the scale (i.e. were not forced to select a 1 and/or a 5). Once listeners were satisfied with their assessment and ratings, they submitted and moved to the next trial. No hidden reference was provided; judgements were purely comparative. All sound qualities and stimuli, along with the presentation of the renderers within a trial, were randomized.

2.3 Phase III

Phase III was concerned with determining a ranking of the six renderers in terms of user preference by forcing subjects to rank the renderers from least preferred to most preferred. No additional information about what such an assessment entailed was provided. The test presented three trials, one for each stimulus. The order of presentation of the stimuli was randomized for each subject. In each trial a subject was tasked with constructing a ranking of the renderers under the following procedure: the order of the six renderers was first randomized. The renderers were then automatically played for 7 seconds (in lieu of the full 20 or 30 seconds) in that order. After all renderers had been played, subjects were instructed to select their least preferred renderer from the set. They were free to replay any of the renderers for any period of time before making this selection. The renderer that was selected as the least preferred was removed and the remaining renderers were reshuffled and presented again with the same procedure. This process of elimination continued until a complete ranking of renderers from least preferred to most preferred was determined.

3 Results

A total of 80 paid subjects participated in Phase II and III. Some subjects did not complete Phase II. This totaled 45 "music" and 35 "movie" Phase III tests and 43 "music" subjects and 29 "movie" Phase II tests. A number of statistical methods were used to analyze the raw data. The data was not normalized as the absolute differences between renderers were of interest. The data from Phase II and III were treated separately at first. The analyses for Phase II included repeated-measures multivariate analysis of variance (RM-MANOVA), repeated-measures analysis of variance (RM-ANOVA), correlation analysis, and principle components analysis (PCA). Parametric statistical methods, of which ANOVA and Pearson correlations are categorized, have been shown to be robust to violations of traditional parametric assumptions and can be used to analyze 5-point interval data (and even ordinal Likert data) [23, 24]. The analysis for Phase III consisted of Friedman tests (because the data was ranked) [25] and follow-up Dunn-Bonferroni multiple comparison tests (to test for pairwise differences between nonparametric distributions) [26]. The data was then analyzed in conjunction to understand how well the various sound quality attributes were correlated with renderer rank. This was done using Spearman rank correlations analyses [27] and Friedman Tests (including follow-up Dunn-Bonferonni multiple comparison tests). Significance is reported at three levels, α <0.05, 0.01, 0.001 for all statistical tests, denoted with *, **, and ***, respectively. For the ANOVA test statistics, Greenhouse-Geisser corrections were used when sphericity assumptions were violated ($\alpha < 0.05$) (denoted with ^{a} in the appendix).

3.1 Phase II

3.1.1 Analysis of Variance

The data in Phase II was captured on a 5-point scale. It was treated as interval data in an initial set of typical statistical models. This was used to gain a preliminary understanding of the variance structure of the data. The Pillai's Trace F-Statistic was used in the multivariate

tests as it displays the most robust behavior to deviations from assumptions of homoskedasticity [26]. In order to get an understanding of the differences between the two experimental conditions - music and movie - a MANOVA test was first carried out. Subjects' answers across each of the different stimuli for each experimental condition were averaged, resulting in a balanced design with a single between-subjects factor, *content type*, a single within-subject factor, *renderer*, and four dependent measures - *balance, clarity, naturalness* and *spaciousness*.

At the multivariate level, *content type* was not significant, but *renderer* (Pillai's Trace=0.951, F(20,51)=49.305, $p<0.001$***, Partial ETA^2=0.951), and the interaction term renderer*type, (Pillai's Trace=0.739, F(20,51)=7.228, $p<0.001$***, Partial ETA^2=0.739) were statistically significant. While there was no significance differences in rating due solely to *content type*, the multivariate tests indicated that the *content type* interacts with the renderers, meaning that the individual renderer performance varies across the two conditions. These results prompted further univariate tests for each of the significant factors. At the univariate level, *renderer* was significant ($p<0.001$***) and *renderer*content type* was significant ($p<0.001$*** for *balance, clarity, and naturalness*; $p=0.021$* for *spaciousness*) for all four dependent measures (see *Appendix A*).

The results of the multivariate and univariate tests indicated that each experimental condition needed to be analyzed separately. This also permitted analyzing the individual stimuli in each condition, in lieu of averaging. In the music condition, a repeated-measures MANOVA was conducted, this time with two within-subject factors - *renderer* and *stimulus* -, no between-subject factors, and four dependent measures - *balance, clarity, naturalness* and *spaciousness*. The multivariate tests indicated a significant effect due to *renderer* (Pillai's Trace=0.977, F(20,23)=49.475, $p<0.001$***, Partial ETA^2=0.977) and *renderer*stimulus* (Pillai's Trace=0.998, F(40,3)=33.599, $p<0.007$**, Partial ETA^2=0.998), but not due to *stimulus*. Univariate tests were once again conducted for each of the significant factors. At the univariate level, *renderer* was significant for each dependent measure ($p<0.001$***) and *renderer*stimulus* was significant for all dependent measures ($p<0.25$*) except *balance* ($p=0.062$) (see *Appendix B*). Given the significance of the interaction term, in the music condition, renderer performance for each of the sound quality attributes shows high content-dependency, with only some minor invariance between stimuli for ratings of timbral balance.

The movie condition was also analyzed with a repeated-measures MANOVA with two within-subject factors - *renderer* and *stimulus* - but with five dependent measures - *balance, clarity, naturalness, spaciousness*, and *dialogue*. The multivariate results reported are the F statistics of averaged variables as opposed to the exact statistic; insufficient residual degrees of freedom prevented the calculation of an exact test statistic for the interaction term *renderer*stimulus*. Similarly, the test indicated that *renderer* (Pillai's Trace=1.010, F(25,700)=7.084, $p<0.001$***, Partial ETA^2=0.202) and *renderer*stimulus* (Pillai's Trace=0.667, F(50,1400)=4.232, $p<0.001$***, Partial ETA^2=0.133) were significant. *Stimulus* was once again not significant at the multivariate level. Given the significant effects, follow-up univariate ANOVAs were carried out. *Renderer* and *renderer*stimulus* were highly significant ($p<0.001$***) for each sound quality (see *Appendix C*).

Follow-up Bonferroni-corrected multiple comparison tests were computed for all of the univariate ANOVA tests in each experimental conditions (music and movie). These are not reported, but generally, many of the groups were significant, indicating that the renderers can be distinguished from one another. Significant group differences for renderers 04 and 05 against each of the other renderers were typically found. Group differences between the other four were also found, but much less frequently. Given the significance of the

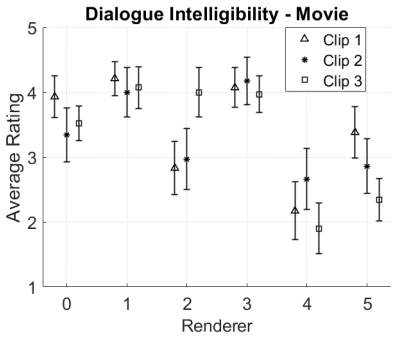

Fig. 1: Dialogue - Average ratings and 95% confidence intervals for movie condition.

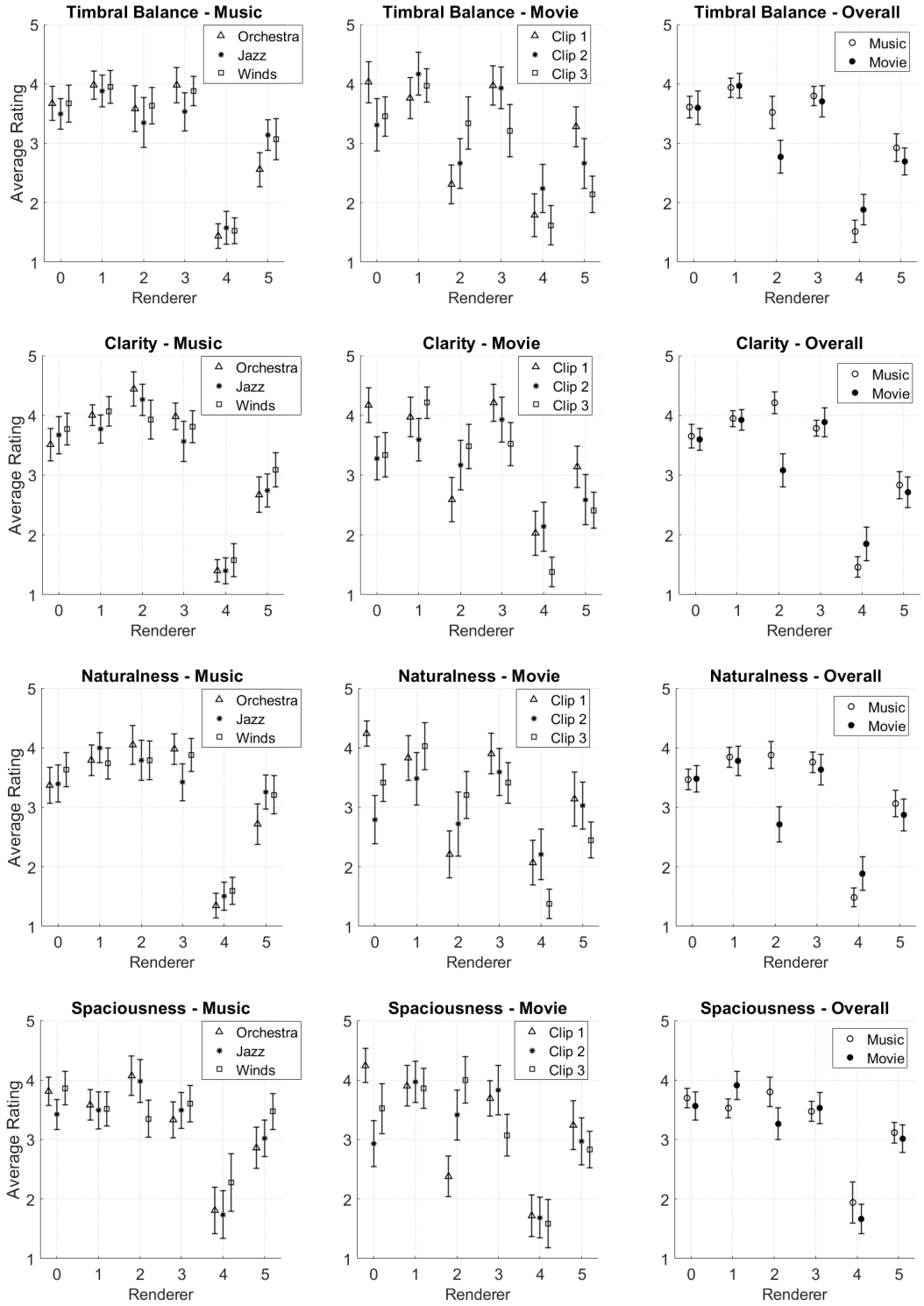

Fig. 2: Average ratings and 95% confidence intervals of each spatial sound quality attribute for the music condition *(left)*, movie condition *(center)* and across conditions *(right)*.

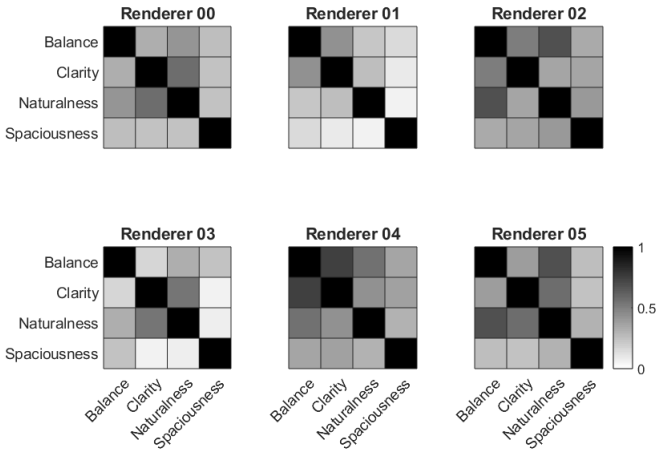

Fig. 3: Corrgram of Phase II Music data, averaged across stimuli.

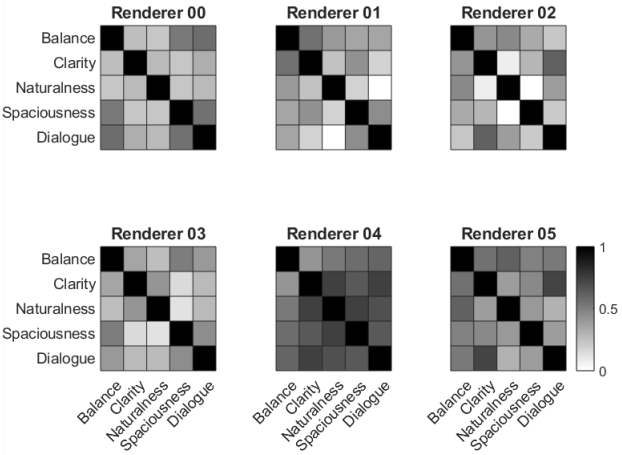

Fig. 4: Corrgram of Phase II Movie data, averaged across stimuli.

interaction between renderer and stimulus and the difference between experimental conditions, the average ratings and 95% confidence intervals for four of the spatial sound quality attributes - *balance, clarity, naturalness, spaciousness* - pictured in *Fig. 2* are broken down at both levels of analysis (across experimental condition and within experimental condition). *Dialogue Intelligibility* was unique to the movie condition, so this particular characteristic is plotted separately (*Fig. 1*). The averages and confidence intervals confirm the results of the ANOVA and post hoc testing. From the figures it is clear that renderer 04 was the weakest performer across all metrics and renderer 05 also performed quite poorly when compared to the remaining four renderers. Comparing between the two conditions (right column), the variability of renderer 02's performance is evident. While being one of the strongest performers in the music condition, it tended to cluster with renderers 04 and 05 in the music condition. Other minor differences between the other renderers due to content type can be observed, but they are not as drastic as with renderer 02. Further, there appears to be a stronger interaction between renderer and stimulus in the movie condition, with much greater variance of renderer performance across stimuli. The renderers performed more consistently in the music condition, although the interaction term was still significant for *clarity, naturalness, and spaciousness* and nearly significant (p=*0.063*) for *balance*.

3.1.2 Multicollinearity Analysis

After understanding the general behavior of the independent variables, the dependent variables were explored. This was done using two methods, Pearson correlations and PCA. The variance analysis indicated the experimental conditions deserved separate treatment. To compute the Pearson correlations without violating assumptions of independence, the correlations between each of the sound qualities for each renderer were averaged across the three stimuli and are reported as corrgrams in *Figs. 3-4*. Thus each subject has one observation of sound quality averages per renderer. The corrgram (the upper triangle of which is redundant) indicates that for each condition there were either near-zero correlations or positive correlations between the sound qualities; increases in ratings of a single quality predict increases in the other qualities. Consistent with *Figs. 1-2*, renderer 04 and 05 exhibit abnormal behavior, with very strong correlations between many of the sound quality attributes. The poorer quality of renderer 04 is apparent and the multicollinearity likely reflects the scale used in the experiment and the comparative nature of the study. Because all renderers were tested comparatively, with little room for nuance in judgements due to the 5-point scale, renderer 04 was likely relegated to the bottom of the scale for all sound quality judgements. The other four renderers have less drastic correlations, with minor to mild correlations being found. There does generally appear to be less collinearity between *spaciousness* and the other qualities, though this is less apparent in the movie condition.

The raw Phase II music and movie datasets, without averaging across stimuli, were then analyzed using PCA. For the music condition, the process returned 4 components that were a linear combination of the original four sound qualities. For the movie condition,

	Balance	Clarity	Naturalness	Spaciousness	% Explained
PCA-1	0.51	0.53	0.52	0.42	64.81
PCA-2	-0.29	-0.16	-0.27	0.90	15.75
PCA-3	0.79	-0.54	-0.29	0.07	10.30
PCA-4	-0.16	-0.63	0.75	0.07	9.14

	Balance	Clarity	Naturalness	Spaciousness	Dialogue	% Explained
PCA-1	0.47	0.44	0.43	0.44	0.45	60.62
PCA-2	-0.147	0.137	0.77	-0.59	-0.15	11.74
PCA-3	-0.03	0.39	-0.40	-0.58	0.59	10.49
PCA-4	0.76	-0.55	0.00	-0.33	0.08	9.28
PCA-5	-0.42	-0.57	0.24	0.13	0.65	7.85

Table 1: PCA Components for Phase II Music (top) and Movie (bottom).

5 components were returned. These components are reported in *Table 1*. In both conditions, the first component accounts for over 60% of the data variance in each experimental condition. And the contribution of each of the characteristics to the component is comparable. The transformed data was then projected onto various dimensions of the PCA space. Pictured in *Fig. 5* are the data projections onto the subspaces created by the first three PCA components. The projections of the sound quality attributes used to construct the space are also included (and multiplied by a factor of 1.5 to improve readability of the plots). The projected data is grouped by renderer (using the interquartile ranges as estimates of the spread [11]) and represented as ellipses. The renderer groupings in the PCA figures are consistent with *Figs. 1-2* with regards to the performance of renderers 04 and 05, and the variable performance of renderer 02 between experimental conditions. In the music condition, the PCA projections involving the first component show clustering of renderers 00, 01, 02, and 03. The projections onto the second and third component indicate that most of the data variance due to renderer differences have been factored out and thus all renderers cluster together. The movie condition is similar, with renderers 00, 01, and 03 clustering together in projections involving the first component and all renderers clustering together for the later components. The second and third component projections show much clearer separation of the sound qualities. Taking the tables and the projections together, its clear that the principle component factors out the variance due to renderer differences, of which all sound quality attributes have similar contributions, before parsing the differences between sound quality attributes. The remaining components, which compose less than half of the data variance, provide further evidence that there is strong collinearity between the factors tested, specifically clarity and naturalness in the music condition and clarity and dialogue in the movie condition. Some of the attributes (such as spaciousness and naturalness) do appear to function along distinct perceptual dimensions, which is discussed further in section 4.

3.2 Phase III

The Phase III tests resulted in 3 rankings of renderers per subject. Renderers were ranked 1-6, with rank 1 being the most preferred renderer and rank 6 being the least preferred renderer. The frequency bar char pictured in *Fig. 6* presents the number of observations of each rank for each renderer, with rank increasing from 1 to 6 as one moves to the right within each renderer grouping, and the two experimental conditions stacked above one another. The renderer rankings mirror the Phase II results, with renderer 04 and 05 being the weakest and second weakest performers, respectively. The other renderers are comparable and renderer 02's variability in experimental condition is also visible. Before running nonparametric tests, the renderer ranks were averaged across stimuli, resulting in a single average set of rankings for each subject. The average-rank data distributions for each renderer were then compared using a Friedman test for k mutually correlated samples [28, 25]. Kendall's coefficient of concordance (W) is also reported and interpreted as a measure of agreement between subjects' rankings. Given that that the Friedman test involves a rank-transformation, rank averages are valid inputs for the test. And while averaging stimuli loses important information about the role of the stimuli in the ranking, the interaction term is often difficult to interpret in rank data. The Friedman test indicated

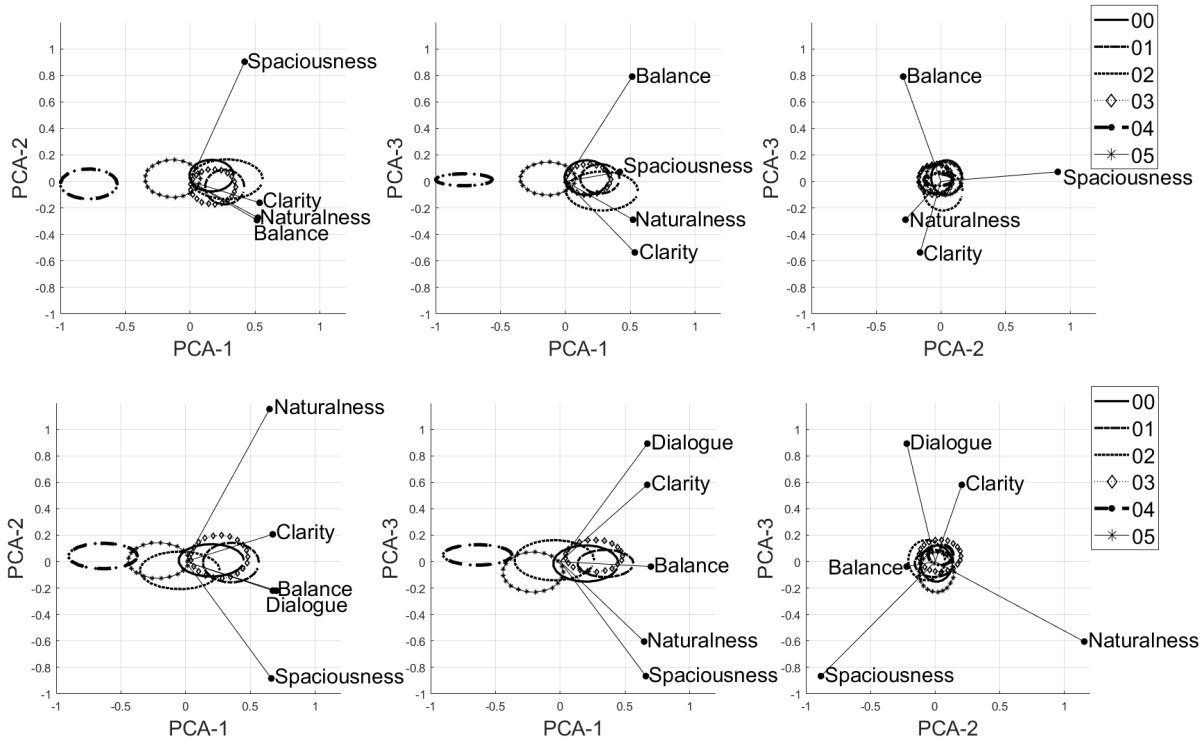

Fig. 5: PCA projections onto the subspaces composed by components 1-3 for Music (top) and Movie (bottom). Data is grouped by the interquartile ranges of each renderer (as denoted in the legend) in the transformed space.

significant differences between the renderers for the music condition (X^2(5,N=45)=137.641, *p<0.001****, Kendall's W=0.612) and for the movie condition (X^2(5,N=35)=102.647, *p<0.001****, Kendall's W=0.587). These are reported in *Table 2*, along with the other summary statistics discussed below. Kendall's *W* indicates that there was a strong consensus between subjects in the rankings of the renderers. Follow-up Dunn-Bonferroni comparisons were used to analyze the pairwise differences between average renderer rankings [29] but are not reported. In both conditions, many of the pairings were significant, and generally confirmed the trend that renderers 04 and 05 were poor performers. The movie condition post hoc tests also indicated that renderer 02 performed poorly when compared with renderers 00, 01, and 03. The pairwise comparisons did not reveal which renderer was outright the strongest performer, but suggested that renderer 01 and 03 when compared across both conditions are consistently ranked in the top group of renderers, a trend which is discernible, though to a lesser degree, in *Fig. 6*.

3.3 Joint Analysis

The joint analyses involved combining the data from Phase II and III to better understand the mapping of spatial sound quality to listener preference. Given the ordinal nature of the rank data, nonparametric methods were used for the joint analysis [30]. Each experimental condition was handled separately. Correlation analysis was carried out to determine which sound quality was most correlated with renderer ranking. To perform this, each subject's answers were averaged across stimuli, resulting in 6 observations per subject, one for each renderer, of average measures of a renderer's naturalness, clarity, balance, and spaciousness (dialogue), and an average rank. For each subject, the rank correlations (Spearman correlations) between each sound quality and rank was computed [27]. Thus each subject had one observation of a *balance-rank, clarity-rank, naturalness-rank,* and *spaciousness-rank* (*dialogue-rank*) correlation. A boxplot displaying the distribution of quality-rank correlations for each sound quality is pictured in *Fig. 7.* A positive correlation is in the di-

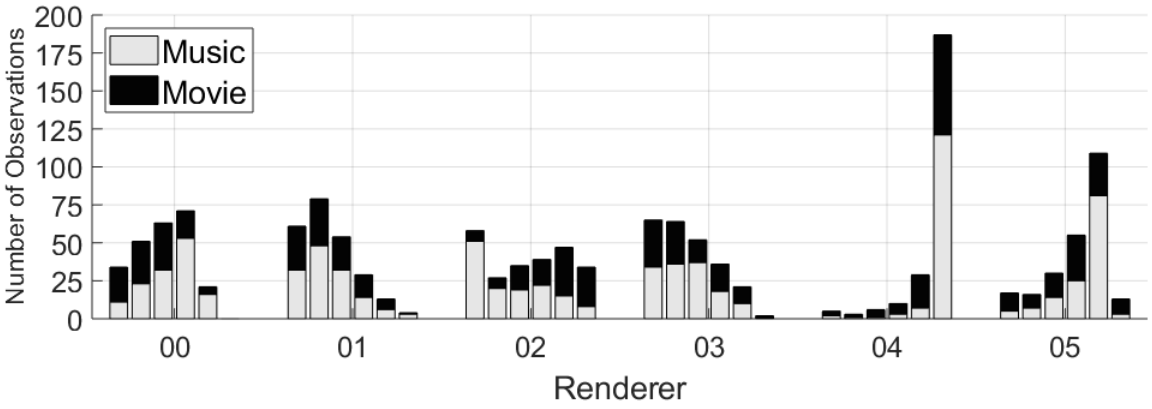

Fig. 6: Rank counts for each renderer. The leftmost bar in each grouping is the number of observations of rank 1 (best rank) and the rightmost bar in each grouping is the number of observations of rank 6 (worst rank).

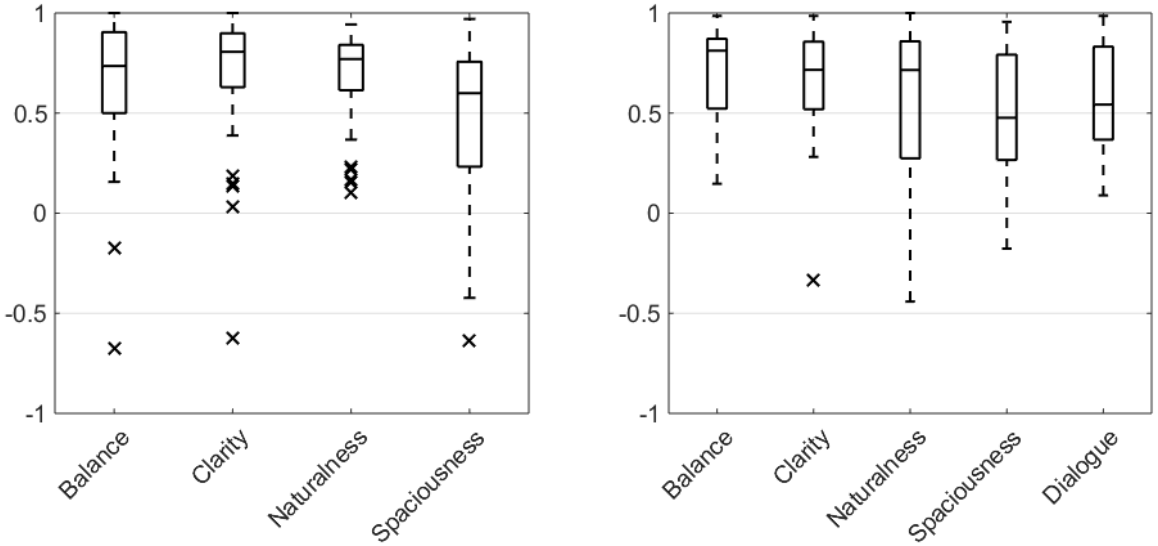

Fig. 7: Boxplot of Attribute-Rank Spearman Correlations for Music (left) and Movie (right).

rection of improved rank. The boxplot indicates that most of subject correlations were positive, though there exist some outliers and heteroskedastic behavior. In the music case, spaciousness appears least correlated with rank. In the movie case, there are fewer outliers and naturalness appears least strongly correlated with preference. Across the two conditions, clarity appears to have the strongest relationship with rank, though all groups are generally quite similar.

Given the distributions, a Friedman test was again used, with the null hypothesis being that there are no differences between the distributions of attribute-rank correlation coefficients. In the music case, significant differences were found (X^2(3,N=43)=16.385, $p<0.001$***, Kendall's W=0.127). Follow-up post-hoc Dunn-Bonferroni tests indicated significant differences between naturalness and spaciousness ($p=0.021$*) and clarity and spaciousness ($p=0.001$**), and near-significance between balance and spaciousness ($p=0.058$). No statistically significant differences were found in the movie condition (X^2(3,N=29)=6.000, $p<0.199$, Kendall's W=0.052).

4 DISCUSSION

The analyses of the multidimensional sound quality assessment revealed a number of trends and relationships about the behavior of the renderers, the stimuli, and the sound quality attributes. The initial Phase II MANOVAs and ANOVAs tested whether the indepen-

Scenario	Test Type	Test Statistic	Significance
Phase III - Music			
	Friedman	$X^2(5, N = 45) = 137.641$	$p < 0.001$***
	Kendall's Coefficient	$W = 0.612$	$p < 0.001$***
Phase III - Movie			
	Friedman	$X^2(5, N = 35) = 102.647$	$p < 0.001$***
	Kendall's Coefficient	$W = 0.587$	$p < 0.001$***
Joint Analysis - Music			
	Friedman Test	$X^2(3, N = 43) = 16.385$	$p = 0.001$**
	Kendall's Coefficient	$W = 0.127$	$p = 0.001$**
Joint Analysis - Movie			
	Friedman	$X^2(3, N = 29) = 6.000$	$p = 0.199$
	Kendall's Coefficient	$W = 0.052$	$p = 0.199$

Table 2: Summary Statistics of Nonparametric Tests.

dent variables in the experiment had statistically significant effects on ratings of sound quality attributes. These results indicated that renderer performance was highly content-specific, interacting with the renderers at both the experimental condition level - music versus movie - and at the level of the individual stimuli. The content-dependence of ratings of sound quality attributes are consistent with previous literature [11]. Further, there tended to be two groups of renderers, with renderer 00, 01, and 03 having similarly strong performance and renderers 04 and 05 having similarly weak performance. Renderer 02 was interesting because it performed well in the music condition, but quite poorly in the movie condition, once again highlighting that the binaural rendering process can have complex interactions with the content to be rendered and so the selection of a renderer for use in creating content should be consistent with the end-goals of the application (ie: computer games, music, or movie).

The Phase II correlation analysis looked at the correlations between sound quality ratings for each renderer (*Figs. 3-4*) to understand the multicollinearity of the sound quality attributes tested. It also provided a means to understand the variability of renderer multicollinearity between sound quality attributes. Renderers that performed strongest also demonstrated the weakest correlations. Even so, generally, there was strong collinearity between most of the characteristics, which was further substantiated by the PCA analysis. The multicollinearity analysis makes clear a few things though. First, there is a distinction between spaciousness and the other quality attributes tested. In the music condition, the second PCA component (*Table 1*) indicates that the remaining data variance is partitioned based on differences between the spaciousness and the other three sound quality attributes. Thus for music content, spaciousness appears to be perceptually distinct from the other quality attributes. Second, naturalness, which does not appear wholly distinct from the balance and clarity in the music condition, is clearly separated out in the movie condition. This suggests that for music content, where naturalness does not necessarily have an ecological meaning, it is perceptually similarly to clarity and balance. But for movie content, where the auditory reference point is an entire spatial scene (including sound effects, music, and dialogue), naturalness is more clearly differentiated from the other characteristics [31].

The Phase III analysis indicated that renderer 04 was the least preferred renderer by a large margin. The ranking data generally agreed with the Phase II results, with renderers 01 and 03 being the strongest performers across both the music and movie conditions, renderer 02 being strong in the music condition and weak in the movie condition, and renderer 00 being middle-of-the-table. The final joint analysis attempted to understand the mapping of sound quality attributes to rank by correlation analysis. This was not unilaterally successful, as statistically significant differences between the distributions of quality-rank correlations were found in the music case, but not the movie case. Thus in the movie case, no judgements on the relative importance of the characteristics on rank can be made. The boxplots taken together show a trend towards clarity being a strong predictor of rank and spaciousness a weak predictor of rank (especially for music content).

Following from this initial study and analyses, a number of possible improvements can be identified with

respect to the proposed methodological approach for Phase II and III. Though a 5-point rating scale in Phase II was fine enough to discriminate between renderers given a large number a subjects, the corrgrams (*Fig. 3-4*) suggest that the scale might not permit enough nuance to meaningfully differentiate between all of the different sound quality attributes. This issue can be exacerbated when one of the tested renderers is relatively weak and many renderers are being tested comparatively. An 11-point rating scale or perhaps even a much larger scale (100-points) might prove useful. Further, electing to determine user preference in a separate assessment with a different methodological approach than the other sound quality attributes presents some difficulties with respect to statistical analyses and requires appropriately handling the different data types (interval vs. ordinal). This should be well-understood before beginning binaural renderer evaluation.

5 CONCLUSIONS

The results of the sound quality assessment strongly indicate that renderer performance is highly content dependent, making it unlikely that a given renderer will be optimal for multiple different applications. There were significant interactions between the renderers and experimental conditions - music and movie - and between the renderers and the individual stimuli within the conditions. Thus the context in which the renderer will be realized must be considered before choosing a renderer. Further, a diverse set of stimuli, consistent with said context, must be used for renderer evaluation. The methodological approach used and analyzed in this work performs well for the set of six commerical binaural renderers selected, with the renderers able to be discriminated between. But some improvements have been suggested, the most significant of which is using a finer rating scale. Future work will include regression analysis to determine a more precise and significant mapping of sound quality attributes to preference in this setting.

6 ACKNOWLEDGEMENTS

The authors would like to thank THX Ltd for their support on this research. Special thanks to Dr. Johanna Devaney for her statistics guidance.

References

[1] Reardon, G., Calle, J. S., Genovese, A., Zalles, G., Olko, M., Jerez, C., Flanagan, P., and Roginska, A., "Evaluation of Binaural Renderers: A Methodology," in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[2] Reardon, G., Zalles, G., Genovese, A., Flanagan, P., and Roginska, A., "Evaluation of Binaural Renderers: Externalization, Front/Back and Up/Down Confusions," in *Audio Engineering Society Convention 144*, Audio Engineering Society, 2018.

[3] Reardon, G., Genovese, A., Zalles, G., Flanagan, P., and Roginska, A., "Evaluation of Binaural Renderers: Localization," in *Audio Engineering Society Convention 144*, Audio Engineering Society, 2018.

[4] Gabrielsson, A., "Dimension analyses of perceived sound quality of sound-reproducing systems," *Scandinavian Journal of Psychology*, 20(1), pp. 159–169, 1979.

[5] Letowski, T., "Sound quality assessment: concepts and criteria," in *Audio Engineering Society Convention 87*, Audio Engineering Society, 1989.

[6] Berg, J. and Rumsey, F., "In search of the spatial dimensions of reproduced sound: Verbal protocol analysis and cluster analysis of scaled verbal descriptors," in *Audio Engineering Society Convention 108*, Audio Engineering Society, 2000.

[7] Zacharov, N. and Koivuniemi, K., "Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training," in *Audio Engineering Society Convention 111*, Audio Engineering Society, 2001.

[8] Choisel, S. and Wickelmaier, F., "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound," *Journal of the Audio Engineering Society*, 54(9), pp. 815–826, 2006.

[9] Olko, M., Dembeck, D., Wu, Y.-H., Genovese, A., and Roginska, A., "Identification of Perceived Sound Quality Attributes of 360° Audiovisual Recordings in VR Using a Free Verbalization Method," in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[10] Rumsey, F. and Berg, J., "Verification and correlation of attributes used for describing the spatial quality of reproduced sound," in *Audio Engineering Society Conference: 19th International Conference: Surround Sound-Techniques, Technology, and Perception*, Audio Engineering Society, 2001.

[11] Guastavino, C. and Katz, B. F., "Perceptual evaluation of multi-dimensional spatial audio reproduction," *The Journal of the Acoustical Society of America*, 116(2), pp. 1105–1115, 2004.

[12] Berg, J. and Rumsey, F., "Systematic evaluation of perceived spatial quality," in *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, Audio Engineering Society, 2003.

[13] Eisler, H., "Measurement of Perceived Acoustic Quality of Sound-Reproducing Systems by Means of Factor Analysis," *The Journal of the Acoustical Society of America*, 39(3), pp. 484–492, 1966.

[14] Lorho, G., "Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating," in *Audio Engineering Society Convention 118*, Audio Engineering Society, 2005.

[15] Andreopoulou, A. and Katz, B. F., "Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees," *Journal on Multimodal User Interfaces*, 10(3), pp. 259–271, 2016.

[16] Simon, L. S., Zacharov, N., and Katz, B. F., "Perceptual attributes for the comparison of head-related transfer functions," *The Journal of the Acoustical Society of America*, 140(5), pp. 3623–3632, 2016.

[17] Rumsey, F., "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *Journal of the Audio Engineering Society*, 50(9), pp. 651–666, 2002.

[18] Susini, P., McAdams, S., and Winsberg, S., "A multidimensional technique for sound quality assessment," *Acta acustica united with Acustica*, 85(5), pp. 650–656, 1999.

[19] Marins, P., Rumsey, F., and Zielinski, S., "Unravelling the relationship between basic audio quality and fidelity attributes in low bit-rate multichannel audio codecs," in *Audio Engineering Society Convention 124*, Audio Engineering Society, 2008.

[20] Le Bagousse, S., Paquier, M., and Colomes, C., "Assessment of spatial audio quality based on sound attributes," in *Acoustics 2012*, 2012.

[21] ITU-T, "Multichannel sound technology in home and broadcasting applications," Recommendation BS.2159-6, International Telecommunication Union, Geneva, 2013.

[22] Toole, F. E., "Subjective measurements of loudspeaker sound quality and listener performance," *Journal of the Audio Engineering Society*, 33(1/2), pp. 2–32, 1985.

[23] Norman, G., "Likert scales, levels of measurement and the "laws" of statistics," *Advances in health sciences education*, 15(5), pp. 625–632, 2010.

[24] Sullivan, G. M. and Artino Jr, A. R., "Analyzing and interpreting data from Likert-type scales," *Journal of graduate medical education*, 5(4), pp. 541–542, 2013.

[25] Sheldon, M. R., Fillyaw, M. J., and Thompson, W. D., "The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs," *Physiotherapy Research International*, 1(4), pp. 221–228, 1996.

[26] Olson, C. L., "Comparative robustness of six tests in multivariate analysis of variance," *Journal of the American Statistical Association*, 69(348), pp. 894–908, 1974.

[27] Hauke, J. and Kossowski, T., "Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data," *Quaestiones geographicae*, 30(2), pp. 87–93, 2011.

[28] Zimmerman, D. W. and Zumbo, B. D., "Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks," *The Journal of Experimental Education*, 62(1), pp. 75–86, 1993.

[29] Dunn, O. J., "Multiple comparisons using rank sums," *Technometrics*, 6(3), pp. 241–252, 1964.

[30] Conover, W. J. and Iman, R. L., "Rank transformations as a bridge between parametric and nonparametric statistics," *The American Statistician*, 35(3), pp. 124–129, 1981.

[31] Rumsey, F., "Spatial audio and sensory evaluation techniques-context, history and aims," 2006.

APPENDIX A - Phase II Univariate ANOVA Results

Factor	Dependent Measure	F Statistic	Significance	Partial ETA^2
Renderer				
	Balance	F(3.951,276.603) = 121.992[a]	$p < 0.001$***	0.635
	Clarity	F(3.725,260.750) = 166.169[a]	$p < 0.001$***	0.704
	Naturalness	F(5,350) = 102.293	$p < 0.001$***	0.594
	Spaciousness	F(3.310,231.691) = 69.420[a]	$p < 0.001$***	0.498
Renderer*Content Type				
	Balance	F(3.951,276.603) = 6.039[a]	$p < 0.001$***	0.079
	Clarity	F(3.725,260.750) = 14.377[a]	$p < 0.001$***	0.170
	Naturalness	F(5,350) = 11.403	$p < 0.001$***	0.140
	Spaciousness	F(3.310,231.691) = 3.168[a]	$p = 0.021$*	0.043

APPENDIX B - Phase II Music Only Univariate ANOVA Results

Factor	Dependent Measure	F Statistic	Significance	Partial ETA^2
Renderer				
	Balance	F(3.416,143.470) = 90.792[a]	$p < 0.001$***	0.684
	Clarity	F(3.635,152.677) = 153.816[a]	$p < 0.001$***	0.786
	Naturalness	F(5,210) = 100.104	$p < 0.001$***	0.704
	Spaciousness	F(2.575,108.153) = 35.799[a]	$p < 0.001$***	0.460
Renderer*Stimulus				
	Balance	F(7.380,309.955) = 1.918[a]	$p = 0.062$	0.044
	Clarity	F(6.871,288.590) = 2.356[a]	$p = 0.024$*	0.053
	Naturalness	F(10,420) = 2.474	$p = 0.013$*	0.056
	Spaciousness	F(10,420) = 3.355	$p = 0.002$**	0.074

APPENDIX C - Phase II Movie Only Univariate ANOVA Results

Factor	Dependent Measure	F Statistic	Significance	Partial ETA^2
Renderer				
	Balance	F(5,140) = 45.900	$p < 0.001$***	0.621
	Clarity	F(3.328,93.185) = 50.226[a]	$p < 0.001$***	0.642
	Naturalness	F(5,140) = 30.414	$p < 0.001$***	0.521
	Spaciousness	F(5,140) = 39.434	$p < 0.001$***	0.585
	Dialogue	F(5,140) = 33.644	$p < 0.001$***	0.546
Renderer*Stimulus				
	Balance	F(10,280) = 6.882	$p < 0.001$***	0.197
	Clarity	F(10,280) = 6.766	$p < 0.001$***	0.195
	Naturalness	F(10,280) = 7.616	$p < 0.001$***	0.214
	Spaciousness	F(10,280) = 8.805	$p < 0.001$***	0.239
	Dialogue	F(10,280) = 7.135	$p < 0.001$***	0.203