Audio Engineering Society

Convention Paper

Presented at the Convention
2018 August 20 – 22, Redmond, WA, USA

Evaluation of Binaural Renderers: Sound Quality Assessment

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

[1]*New York University, 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)

ABSTRACT

Binaural renderers can be used to generate spatial audio content over headphones for use in a number of different media applications. These renderers take individual audio tracks with associated metadata and transform this representation into a binaural signal. A large multi-phase experiment evaluating six commercially available renderers was carried out. This paper presents the methodology, evaluation criteria, and main findings of the tests which assessed perceived sound quality of each of the renderers. In these tests, subjects appraised a number of specific sound quality attributes - *naturalness, spaciousness, timbral balance, clarity*, and *dialogue intelligibility* - as well as overall preference. Results indicated that binaural renderer performance is highly content-dependent, 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 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.

Binaural renderers use object-based means for binaural reproduction over headphones. An audio object is an audio waveform with associated metadata describing the location of the sound source in virtual space at any given time, its reverb character, and its directivity, amongst other properties. In interactive settings, the object's metadata is updated in real-time based on the user's location and head orientation in the virtual space. The goal is for objects to appear as naturally occuring within the virtual, or augmented, environment [1]. The flexibility of object-based audio contrasts with traditional channel-based content, such as surround-sound reproduction, in which the location of audio sources are baked into the transmitted audio. Binaural renderers can also be used in non-tracked displays as a mean for creating static binaural content.

1.1 Methodology overview

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. The overall methodology has been detailed by the authors in a previous work [2]. In the experiment, six different renderers were compared using a number of qualitative and quantitative metrics. 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 were presented in [3], while the results of the horizontal localization test are found in [4]. 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 by the participants and serves as a global assessment of perceived sound quality. This assessment is critical as it can be treated as an outcome variable in a multiple regression to determine which of the sound quality attributes best predicts preference for a binaural renderer. The results of both Phase II and III, along with the multiple regression analysis, are presented in this work.

1.2 Previous Work

It has been pointed out by Rumsey [5] that spatial audio quality is not necessarily a purely objective measure, related to things such as *localization accuracy* or *externalization*. It is also reliant on subjective psychological assessments of various factors that contribute to the listener's experience of immersion.

The search for appropriate descriptor labels to be used as sound quality attribute scales has been the subject of numerous studies. In [6], Le Bagousse reviews the range and variety of elicited quality descriptors used in subjective spatial sound quality assessment studies. Although the understanding of the definitions has been shown to vary, all reported studies generally agree on presenting attributes related to various complexities of spatial impression [5, 7], timbral qualities and coloration [8] and, occasionally, presence and naturalness [9].

Similar to this work, a previous study [9] explored the interactions of reproduction methods and stimulus on overall preference of spatial audio content. An important finding was that the choice of preferred method significantly interacted with the type of content presented. The paper concluded that while attributes like *presence* and *readability* were important to listeners, no universally optimal reproduction method could be determined.

Another related work [10] 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 [11], sensory judgements were correlated to hedonic preferences for surround-sound quality. Three clusters of attributes were selected - *timbre, space* and *defects* - the latter of the three being deemed the most influential in overall preference 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 content type and rating methodology.

2 METHODOLOGY

2.1 Rendering Procedure, Stimuli, and Presentation

Six different binaural renderers were tested in a 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 through head-related transfer functions (HRTFs). Though each renderer has head-tracking capabilities in its native application, the experiment content was presented under a static condition.

A total of six different stimuli were tested in phase II and III - three music and three movie stimuli. The "music" stimuli were three different short musical excerpts. These stimuli were recorded works cut to approximately 20 seconds in length. The stimuli were of varying style, one jazz and two distinct symphonic orchestral works. The jazz piece was mixed for 5.0 surround sound. 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 each no more than 30 seconds in length and each included dialogue, music, and sound effects. For each

stimuli, the individual channels were treated as independent virtual audio objects by each of the processing renderers. 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 [12]. These channels were rendered to a single piece of static binaural content without additional room information; all settings regarding room reverb and early reflections were turned off. All other renderer properties were set to their highest quality.

Though all mixes had a sub-woofer channel, this channel was not rendered. Spatializing the sub channel often results in a muddled low-end. Given that in the ideal case this channel would be included in the final static binaural mix as stereo headlocked content and therefore identical for each renderer, the authors deemed it unnecessary to include this channel. Furthermore, the frequencies in this channel range are very difficult to localize.

While Phase I was identical for each participant, for Phase II and III each subject was randomly assigned to either the "music" or "movie" stimuli 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) in a soundproof booth (NYU Dolan Isolation Booth). Custom software was used 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, 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 [5, 8, 9]. Ultimately, the descriptors were defined as follows:

- **Naturalness**: This attribute describes whether the sound gives a realistic impression, as opposed to artificial.
- **Clarity**: This attribute describes whether the sound appear to be clear or muffled.
- **Spaciousness**: This attribute describes how much the sound appears to surround you.
- **Timbral Balance**: This attribute describes how balanced (or colored) the different tone ranges of the sound appear to be.
- **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 scale of 1-5, with 1 being the worst, and 5 the best. The subject would then be 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; subjects were not forced to select a 1 and/or a 5. After the subject was satisfied with their assessment and ratings, they submitted and moved to the next trial. No hidden reference was provided; judgements were purely comparative.

Phase II took approximately 45 minutes to 1 hour to complete. To account for order of presentation, and thereby listener fatigue, influencing the subject's responses, both sound quality assessed and stimulus presented were randomized. That is, a random member of quality-stimulus pairs was drawn without replacement from the 12-element set. After a quality and stimulus had been selected, the six renderers were randomized for presentation. This also ensured that subjects were blind to the renderer they were evaluating.

2.3 Phase III

Phase III was concerned with evaluating total sound quality for each renderer. Total sound quality was assessed by forcing subject's to rank the renderers from

least preferred to most preferred. No additional information about what such an assessment entailed was provided; subjects were left to use their own internal reference for "quality." Subject's assigned to the music or movie condition in Phase II were again provided the same music or movie stimuli, respectively. This test had three trials, one for each stimulus. The order of stimulus to be presented in these trials was randomized for each subject.

In each trial a subject was tasked with constructing a ranking of the renderer 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 renderer for any period and 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 number of statistical tests and regressions were carried out on the data. The data from Phase II and III were treated separately at first and analyzed using repeated-measures multivariate analysis of variance (MANOVA) and repeated-measures analysis of variance (ANOVA) tests. The data was then analyzed in conjunction to understand how the various sound quality attributes were correlated with renderer rank. A significance value of α<0.05 was used for all statistical tests.

3.1 Phase II

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 - *naturalness, clarity, balance* and *spaciousness*.

At the multivariate level, "content type" was not significant, but "renderer" (Hotelling's Trace=10.201, F(20,41)=20.913, *p<0.001**, Partial ETA Squared=0.911), and the interaction term renderer*type, (Hotelling's Trace=2.418, F(20,41)=4.956, *p<0.001**, Partial ETA Squared=0.707) were statistically significant. While there is no significance differences due solely to content type, the multivariate tests indicate that the content type interacts with the renderers, so individual renderer performance varies across the two conditions. At the univariate level, "content type" was found to be not significant for each of the dependent variables - naturalness, clarity, balance, and spaciousness. The results of the univariate tests for this statistical design, for all significant factors and for each of the dependent variables, are presented in *Table 1*.

Each experimental condition was then analyzed individually to determine the effect that the different stimuli had on the dependent measures. In the music condition, a repeated-measures MANOVA was once again conducted, this time with two within-subject factors - *renderer* and *stimulus* -, no between-subject factors, and four dependent measures - *naturalness, clarity, balance* and *spaciousness*. 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*. The multivariate test of averages indicated a significant effect due to "renderer" (Hotelling's Trace=4.305, F(20,682)=36.702, *p<0.001**, Partial ETA Squared=0.518) and the "renderer*stimulus" interaction (Hotelling's Trace=0.252, F(40,1382)=2.178, *p<0.001**, Partial ETA Squared=0.059), but not due to stimulus. These results prompted further univariate tests for each of the significant factors. These results are reported in *Table 2*. The table also reports which univariate test statistic were used. A Greenhouse-Geisser correction was used when sphericity assumptions were not met. This correction is denoted in the table, and for all proceeding tables, in the F Statistic column as ^{a}.

The movie condition was also analyzed with a repeated-measures MANOVA with two within-subject factors - *renderer* and *stimulus* - but with five dependent measures - *naturalness, clarity, balance, spaciousness*, and *dialogue*. At the multivariate level, the statistics of the averaged variables are once again reported due to insufficient residual

Factor	Dependent Measure	F Statistic	Significance	Partial ETA Squared
Renderer				
	Naturalness	F(5,300) = 77.616	$p < 0.001$*	0.564
	Clarity	F(3.389,203.331) = 106.196[a]	$p < 0.001$*	0.639
	Spaciousness	F(3.019,181.124) = 44.438[a]	$p < 0.001$*	0.425
	Timbral Balance	F(3.990,239.386) = 87.218[a]	$p < 0.001$*	0.592
Renderer*Content Type				
	Naturalness	F(5,300) = 6.914	$p < 0.001$*	0.103
	Clarity	F(5,300) = 10.561	$p < 0.001$*	0.150
	Spaciousness	F(5,300) = 2.489	$p = 0.061$	0.040
	Timbral Balance	F(5,300) = 5.520	$p < 0.001$*	0.084

Table 1: Results of univariate repeated-measures ANOVAs with experimental condition as a between-subjects factor

Factor	Dependent Measure	F Statistic	Significance	Partial ETA Squared
Renderer				
	Naturalness	F(5,175) = 71.773	$p < 0.001$*	0.672
	Clarity	F(3.378,118.231) = 106.373[a]	$p < 0.001$*	0.752
	Spaciousness	F(2.444,85.552) = 21.698[a]	$p < 0.001$*	0.383
	Timbral Balance	F(3.104,108.646) = 65.537[a]	$p < 0.001$*	0.652
Renderer*Stimulus				
	Naturalness	F(10,350) = 2.615	$p = 0.004$*	0.070
	Clarity	F(6.855,238.908) = 1.412[a]	$p = 0.202$	0.039
	Spaciousness	F(10,350) = 3.414	$p < 0.001$*	0.089
	Timbral Balance	F(10,350) = 2.041[a]	$p = 0.063$	0.053

Table 2: Results of univariate repeated-measures ANOVAs for Music condition

Factor	Dependent Measure	F Statistic	Significance	Partial ETA Squared
Renderer				
	Naturalness	F(3.684,92.111) = 23.642[a]	$p < 0.001$*	0.486
	Clarity	F(2.792,69.794) = 29.051[a]	$p < 0.001$*	0.537
	Spaciousness	F(5,125) = 25.754	$p < 0.001$*	0.507
	Timbral Balance	F(5,125) = 32.550	$p < 0.001$*	0.566
	Dialogue Intelligibility	F(5,125) = 24.079	$p < 0.001$*	0.491
Renderer*Stimulus				
	Naturalness	F(10,250) = 5.045	$p < 0.001$*	0.168
	Clarity	F(10,250) = 4.897	$p < 0.001$*	0.164
	Spaciousness	F(10,250) = 7.031	$p = 0.001$*	0.220
	Timbral Balance	F(10,250) = 4.242	$p < 0.001$*	0.145
	Dialogue Intelligibility	F(10,250) = 5.344	$p < 0.001$*	0.176

Table 3: Results of univariate repeated-measures ANOVAs for Movie condition

degrees of freedom necessary to calculate the exact statistic for the renderer*stimulus interaction term. Similarly, the test indicated that renderer (Hotelling's Trace=2.217, F(25,597)=10.590, *p<0.001**, Partial ETA Squared=0.307) and renderer*stimulus (Hotelling's Trace=0.866, F(50,1222)=4.232, *p<0.001**, Partial ETA Squared=0.148) were significant. Stimulus was once again not significant at the multivariate level. Given the significant effects, five univariate ANOVAs were carried out. The results of which are reported in *Table 3*. Greenhouse-Geisser corrected statistics are indicated and reported for those cases in which the sphericity assumption is not met.

These results suggest sound quality assessments of renderers differ significantly for different content type and for each stimulus. Thus, the descriptive statistics displayed in *Figures 1 to 4* are broken down at both levels of analysis for four of the dependent measures. "Dialogue Intelligibility" was unique to the movie condition so this particular characteristic is not presented for content type (*Fig. 5*).

3.2 Phase III

Phase III was then analyzed at multiple levels. A repeated-measures ANOVA was conducted with a single between-subjects factor - *content type* - and a single within-subjects factor - *renderer*. The multivariate test indicated that "renderer" (Hotelling's Trace=9.318, F(5,64)=119.270, *p<0.001**, Partial ETA Squared=0.903) and "renderer*content type" (Hotelling's Trace=0.578, F(5,64)=7.403, *p<0.001**, Partial ETA Squared=0.366) were significant. "Content type" was not significant. Because the test indicated a significant interaction between renderers and the content type, each condition was also analyzed individually with a repeated-measures ANOVA with a within-subject design using two factors - *renderer* and *stimulus*. While the design of the experiment as a forced choice ranking of renderers for each stimulus prevents one from assessing the effect of stimulus on renderer rank, the interaction between renderer and stimulus can be interpreted. The music ANOVA indicated significant effects (Greenhouse-Geisser corrected) due to renderer (F(2.845,105.262)=67.711, *p<0.001**, Partial ETA Squared=0.647), but not due to "renderer*stimulus." On the other hand, under the movie condition, the ANOVA indicated significant effects (Greenhouse-Geisser corrected) due to "renderer" (F(3.651,113.171)=50.630, *p<0.001**, Partial ETA Squared=0.620) and "renderer*stimulus" (F(16.026,209.673)=6.181, *p<0.001**, Partial ETA Squared=0.166). The average ranking for each renderer under both levels of analysis are presented in *Fig. 6*.

3.3 Predicting Preference

The above results indicate two important points. First, that the experimental condition influences ratings of sound quality attributes and ranking. Second, that renderers also interact with the individual stimulus within a condition. Given that the movie condition had *dialogue intelligibility* as an additional dependent measure, type-specific regression lines for the music and movie conditions were predicted. In the procedure used, each of the dependent measures was treated as regressors for predicting rank. Each subject's measures were also aggregated across stimuli such that each subject had six observation, one for each renderer, of what is called in this paper a *renderer profile*. Such a renderer profile consists of measures of a renderer's naturalness, clarity, balance and spaciousness (and dialogue) with rank treated as an outcome variable. Each subject's six observations are not independent, thus an individual-specific fixed-effect design was used to capture an individual's group of observations. This amounts to assuming that each individual is estimating a single regression line with a variable intercept that predicts renderer rank for their six observations. The fixed-effects design implies that omitted variables, such as sound quality attributes not directly tested for, are implicitly captured by the individual's intercept. The lack of explicit modeling of the effect that stimulus has on this regression line is a weakness of this approach, but the decision was made for the sake of providing generality to the resulting preference prediction. This regression model also assumes that there is no interactions between the renderer profile regressors.

The results from the linear regression under the music condition are presented in *Table 4*. The results from the linear regression under the movie condition are presented in *Table 5*. In each table, the subject-specific effects (intercepts) are excluded. Standardized beta coefficients for each of our renderer profile regressors are reported and interpreted as the effect size of each of the sound quality attributes on preference. Variance inflation factor is a measure of multicollinearity between regressors.

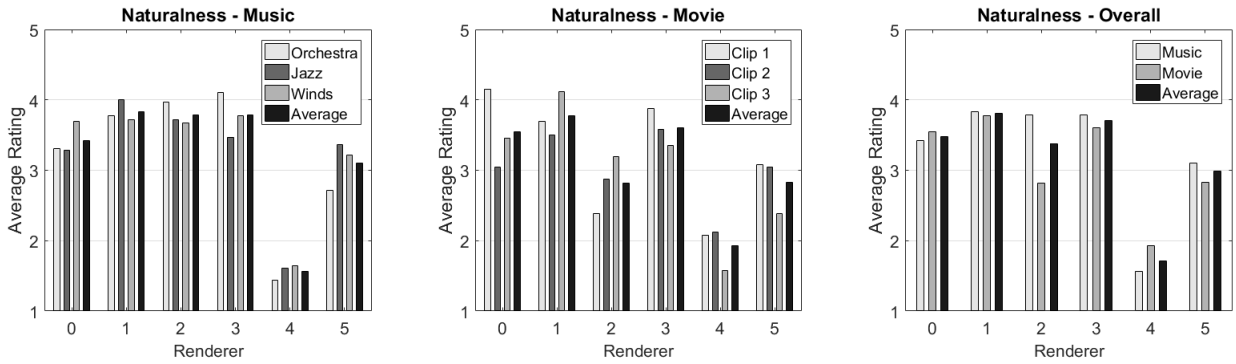

Fig. 1: Naturalness - Average Rating for music condition *(left)*, movie condition *(center)*, and across conditions *(right)*.

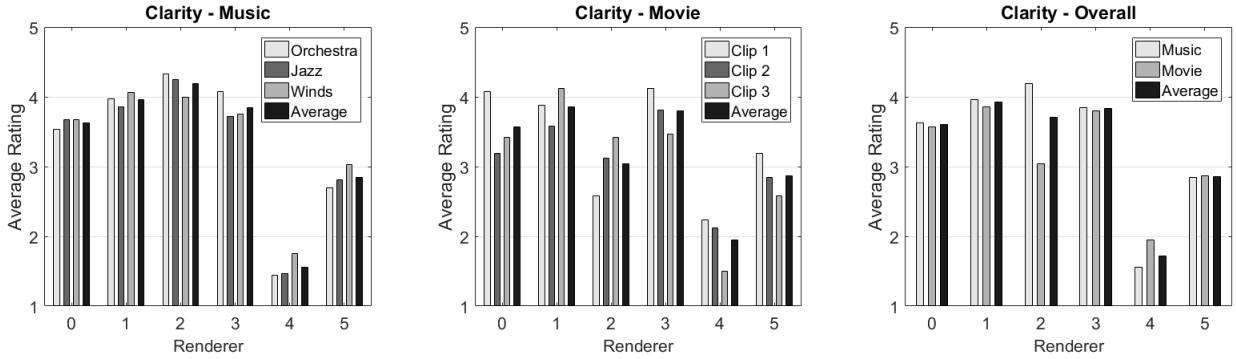

Fig. 2: Clarity - Average Rating for music condition *(left)*, movie condition *(center)* and across conditions *(right)*

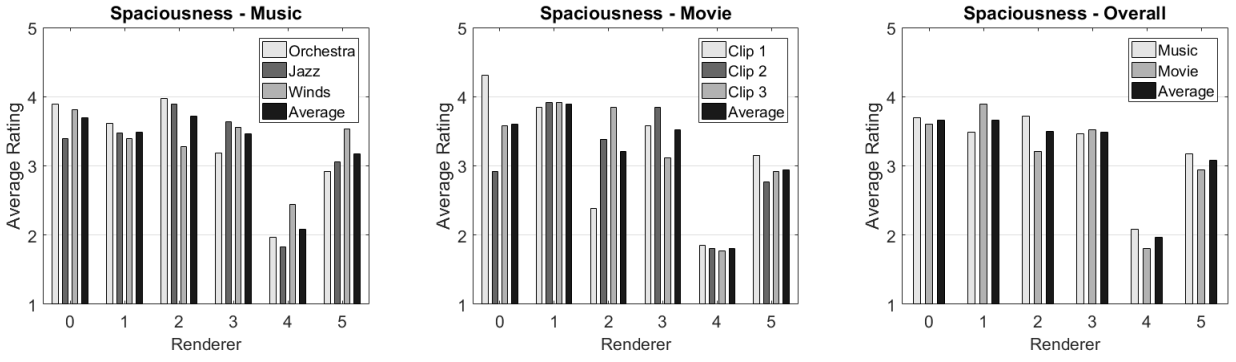

Fig. 3: Spaciousness - Average Rating for music condition *(left)*, movie condition *(center)* and across conditions *(right)*

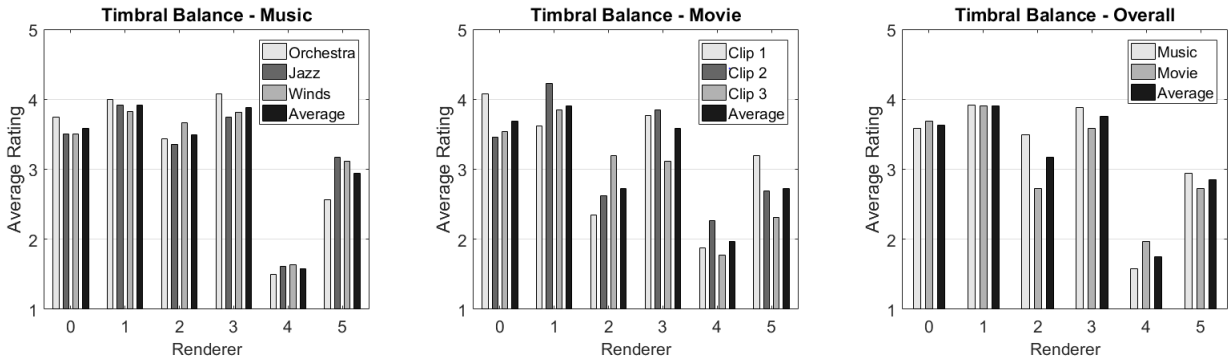

Fig. 4: Timbral Balance - Average Rating for music condition *(left)*, movie condition *(center)* and across conditions *(right)*.

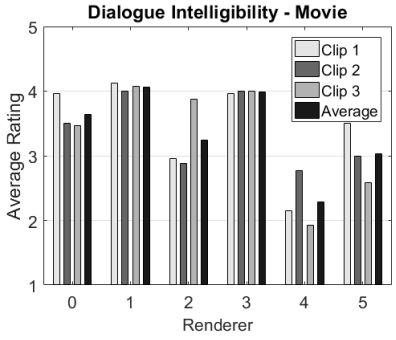

Fig. 5: Dialogue - Average Rating for movie condition.

4 DISCUSSION

The univariate tests are broken up into four sections. The first are the tests for across experimental condition (Table 1). This table breaks down the effect of our significant factors on each of the dependent measures. Taking each measure independently begins to give an understanding of which dependent measure is responsible for the significant interaction term renderer*type in the F test. Measures of spaciousness tend to be more robust across conditions, while for all other measures the interaction term still accounts for a significant amount of the variance of the dataset. In this first univariate table, renderer is still highly significant and explains much more of the variance than that of the interaction term.

Tables 2 and 3 display the univariate tests when each condition is taken individually. Table 2 indicates that for only two of the measures have a significant interaction term renderer*stimuli - clarity and timbral balance. This confounds with Table 1 which implies that spaciousness was most robust to changes in type of content, indicating that averaging stimuli within a condition might severely alter the error distributions of the dependent measures. But, it should be noted that the effect size is relatively small compared to renderer. Table 3 on the other hand, indicates that for the movie condition, interactions between stimuli and renderer amount for large portion of the data variance (Partial ETA Squared from 0.145 to 0.220). This is mirrored in the Phase III univariate tests which are reported in section 3.2. These tests indicate that for the music condition, the renderer*stimulus interaction term is not significant. While for the movie condition, this term is significant and has a strong effect (Partial ETA Squared=0.166). It is not surprising that the stimuli in the movie condition have a stronger interaction with each renderer than those in the music condition. The movie stimuli each include dialogue, music and sound effects; the variability of the content is mirrored in the results.

Given the results of the multivariate and univariate tests, it is necessary to breakdown the performance of renderers by stimulus and content as displayed in the figures. It is clear from the figures that the music ratings across each stimuli (left column) are less variable those in the movie condition (center column), consistent with the results of the statistical tests. These figures also provide a summary of renderer performance. The consistent poor performance of renderer 4 is evident. In almost all instances this renderer performs worst. Another interesting trend is the variable performance of

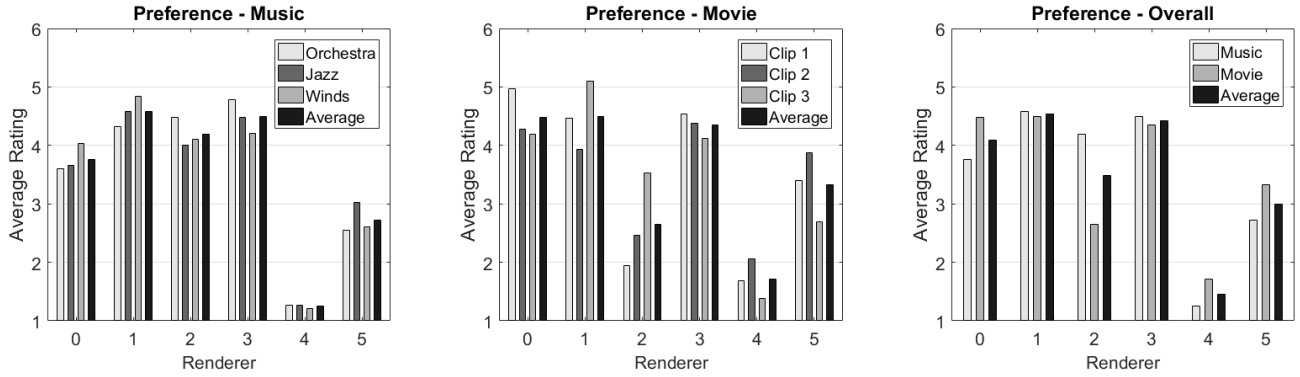

Fig. 6: Preference - Average Ranking for music condition *(left)*, movie condition *(center)* and across conditions *(right)*.

Regressor	Standardized Beta	Significance	Variance Inflation Factor	Adjusted R^2
Naturalness	0.239	*p = 0.003**	4.064	
Clarity	0.400	*p < 0.001**	3.798	
Spaciousness	0.136	*p = 0.009**	1.794	
Timbral Balance	0.240	*p = 0.002**	4.008	
				0.672

Table 4: Predicting Rank (Music) - Results of Linear Regression for music condition

Regressor	Standardized Beta	Significance	Variance Inflation Factor	Adjusted R^2
Naturalness	0.197	*p = 0.055*	2.625	
Clarity	0.270	*p = 0.833*	4.200	
Spaciousness	0.067	*p = 0.520*	2.720	
Timbral Balance	0.392	*p = 0.001**	3.560	
Dialogue Intelligibility	0.176	*p = 0.131*	3.385	
				0.389

Table 5: Predicting Rank (Movie) - Results of Linear Regression for movie condition

renderer 2 under the different conditions. Renderer 2 is one of the strongest performers in terms of quality attributes and overall preference in the music condition but has a drop in performance in the movie condition. This is extremely surprising as the performance of the other renderers is relatively consistent across condition when averaged (right column). Table 1 indicated that spaciousness is the only attribute in which we do not have a significant interaction term between renderer and content type. Looking at *Fig. 3*, the difference between renderer 2's performance in the two condition is least pronounced. Thus it is likely that the performance of renderer 2 is responsible for the significant term renderer*type. Another interesting trend that can be gathered from the figures is that there does appear to be consistency in a renderer's performance on a particular stimuli. For instance, renderer 0 in the movie condition performs quite strongly on stimulus 1 across all dependent measures. This could imply one of two things. First, it might imply that the interaction between content and renderer is consistent. That is, the binaural rendering procedure interacts with the content presented in a predictable manner and results in a number of improvements in dependent measures. Or, it could speak to the multicollinearity of the dependent measures; the dependent measures might be acting as

surrogates for preference. In the later scenario, what is being observed in the figures is preference for the rendering procedure on a particular piece of content that inflates all sound quality attributes. Distilling the differences between the two is not simple, but it does have implications for how a company might go about improving a given renderer.

One of the main tasks of this study was to understand how the various sound quality attributes can be used to predict preference for a renderer. This can provide specific information about which sound quality attributes are most influential for improving a given renderer's performance and can therefore guide improvements in the renderer process. However, much of the above discussion highlights the content-specific nature of renderer performance. This motivates the decision to predict two different content-specific regression lines. While a significant interaction between renderer and stimulus has been reported, the authors sought to generalize, to some degree, by averaging responses within each of the conditions. This also means the authors do not have to explicitly model the effect that stimulus has on performance. The first thing to check in Tables 4 and 5 are the Variance Inflation Factors (VIF). The VIF is a measure of multicollinearity of the regressors. The VIFs in both tables indicate some multicollinearity between variables, but within the accepted tolerance (find reference). Table 4 indicates that each of the regressors are significant. The standardized beta coefficients are an estimate of effect size. These values should not be interpreted literally, only relatively. Thus, clarity is the strongest predictor of preference in the music condition and its effect on renderer preference is almost two times as strong as each of the other three predictors. Table 5 on the other hand indicates that most of the regressors are not significant in the movie condition. Only timbral balance is a significant predictor for preference in this condition. In the absence of significance of the other factors, it is difficult to interpret the standardized beta coefficient (0.392). The lack of significance of all other regressors might be attributed to averaged responses across stimuli. As was evident from the above discussion, the performance of different stimuli in the movie condition was extremely variable. Lack of explicit modeling of this effect likely made it difficult to return consistent estimates for the regressors.

5 CONCLUSIONS

The results of the sound quality assessment strongly indicate that renderer performance is extremely content dependent. There are significant interactions between renderer and movie/music condition and between renderer and stimulus. However, most of the observed variance is explained by differences between renderers. In terms of renderer performance, renderer 04 is the weakest performer in all metrics reported in this work. This finding is consistent with the previous experimental stages [3, 4]. On the other hand the performance of renderers 00, 01, 02, and 03 tend to cluster as the strongest performers for all sound quality attributes tested. Particular attention should be given to the performance of renderer 02 as presenting unusual differences between the music and movie conditions. In the overall preference assessment, renderers 01 and 03 are most preferred by subjects.

In order to understand the most salient attribute for binaural renderer preference under a specific context, music or movie content, the authors attempted to predict preference for a renderer. The regression results for the music condition indicate that clarity is the strongest predictor of preference for binaural rendering of music content. On the other hand, the variability of stimulus likely led to inconsistent results for the movie condition, resulting in timbral balance being the only significant predictor of preference for movie content. Further, as a result, the effect of improving the timbral balance of a renderer on its preference is difficult to quantify.

The presented results are part of a larger comprehensive evaluation of binarual renderers which were presented in previous works. Much insight on the subjective appraisal of immersive audio content can be gained through comprehensive evaluations of commercially available binaural renderers.

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] Begault, D. R. and Trejo, L. J., *3-D sound for virtual reality and multimedia*, NASA, 2000.

[2] 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.

[3] Reardon, G., Zalles, G., 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.

[4] 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.

[5] 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.

[6] Le Bagousse, S., Colomes, C., and Paquier, M., "State of the art on subjective assessment of spatial sound quality," in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, Audio Engineering Society, 2010.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

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