Audio Engineering Society
Convention Paper

Presented at the 140th Convention
2016 June 4–7 Paris, France

Design and Subjective Evaluation of a Perceptually-Optimized Headphone Virtualizer

Grant Davidson[1], Dan Darcy[1], Louis Fielder[1], Zhiwei Shuang[2], Richard Graff[1], Jeroen Breebaart[3], and Poppy Crum[1]

[1] Dolby Laboratories, Inc., 1275 Market Street, San Francisco, CA 94103
Address correspondence to Grant Davidson, gad@dolby.com

[2] Dolby Laboratories Intl. Services Co. Ltd., Room 907-916, Level 9, West Building, World Financial Center, No. 1, East 3rd Ring Middle Road, Chaoyang District, Beijing 100020 China

[3] Dolby Australia Pty Ltd., Level 3, 35 Mitchell Street, McMahons Point, NSW 2060 Australia

ABSTRACT

We describe a novel method for designing echoic headphone virtualizers based on a stochastic room model and a numerical optimization procedure. The method aims to maximize sound source externalization under a natural-timbre constraint. The stochastic room model generates a number of binaural room impulse response (BRIR) candidates for each virtual channel, each embodying essential perceptual cues. A perceptually-based distortion metric evaluates the timbre of each candidate, and the optimal candidate is selected for use in the virtualizer. We designed a 7.1.4 channel virtualizer and evaluated it relative to a LoRo stereo downmix using a single-interval A:B preference test. For a pool of 10 listeners, the test resulted in an overall virtualizer preference of 75%, with no stereo test item preferred over binaural.

1. INTRODUCTION

While the end-user benefit of delivering multichannel audio content to consumers in place of two channels for loudspeaker reproduction is well-established, compelling evidence that binaural processing for headphones is generally preferred by listeners has been challenging to find. Nearly all of the evidence published over the last two decades suggests that stereo and 5.1 channel spatial enhancement methods for headphone presentation (including binaural) offer at best a marginal improvement over stereo [1-5], with the exception of a balanced, near-field cross-talk simulation method described in [6]. Instead, most results suggest listeners actually prefer stereo over a variety of spatial enhancement algorithms [7]. One hypothesized

explanation for these findings is the change in perceived timbre [5, 7] introduced by head-related transfer function (HRTF) convolution and cross-talk induced comb filtering as opposed to using spectrally-flat, phase-coherent amplitude panning functions. Studies suggest that timbre and naturalness are dominant attributes influencing listener assessments of basic audio quality with loudspeaker playback [8, 9]. Taken together, these findings motivate the development of algorithms with a minimum of spectral coloration by so-called 'balanced' networks [5]. A second reason for low preference ratings associated with binaural processing that has been postulated is the reduction in the perceived spatial 'width' [6], which can be overcome by near-field crosstalk simulation instead of far-field HRTFs.

Besides the application of HRTFs to introduce sound source localization cues, binaural renderers often employ some sort of room simulation, by simulating early reflections, late reverberation or both; the combined effect of HRTFs and room simulation can be captured by binaural room impulse responses (BRIRs). It seems intuitively plausible that the addition of such an acoustic environment influences the perceived timbre of a binaural presentation; however, the authors are not aware of any work describing analyses or methods for timbre optimization, nor the effect on listeners' preference assessments in echoic binaural conditions.

In this paper, we describe a method and report on a subjective evaluation of an echoic headphone virtualizer called OptiBRIR that aims for higher listener preference over amplitude panning techniques. Our hypothesis was that if aspects of stereo that listeners value most were present in a binaural presentation along with a realistic sense of space, listeners would prefer the binaural system. The virtualizer design decisions were strongly guided by a minimalistic signal processing philosophy and the objective of preserving the sound mixer's intent. Sound source externalization was considered desirable, but only to the point where virtualizer side-effects such as timbral coloration and temporal distortion began to degrade the experience. An additional aspect that differentiates this study from previous ones is the use of highly immersive audio content in the form of Dolby Atmos printmasters [10]. Each printmaster contains multiple audio channels and audio objects. The channels are intended for loudspeakers at nominal positions, while the objects describe sound sources that can be authored to arbitrary locations in three-dimensional space. The objects can also change location and size over time. The printmasters were rendered to 7.1.4 channel files comprised of seven virtual loudspeakers in the horizontal plane and four overheads. The addition of virtual height channels offers the potential for listener preference gains over stereo and virtualized 5.1 channel systems.

The principal design objectives for the virtualizer were to make a promising step toward the objective of markedly outperforming stereo, and to deliver robust performance on critical audio test items. Accordingly, we established the following desired performance metrics:

- Exceed 70% overall listener preference over stereo, averaged across all test items and subjects.
- No test item should perform worse than stereo in a statistically-significant sense (95% confidence).

Our approach to achieving these goals is based on a numerical optimization procedure to design virtual channel BRIRs that maximize sound source externalization under a natural-timbre constraint. In an offline procedure, a stochastic room model generates a number of BRIR candidates for each virtual channel. A perceptually-based objective function evaluates the timbre performance of each candidate, and the optimal candidate is selected for use in the virtualizer. This process is performed once for each virtual channel.

As a means for evaluating virtualizer performance relative to stereo, we adopted the single-interval A:B preference test. Each stereo signal was generated as a LoRo downmix of the 7.1.4 file. An 11.1-channel LoRo downmix is a direct extension of the ITU-R standard 5.1 channel downmix [11] commonly adopted as a stereo reference in published virtualized surround test reports [2, 4]. This test simulates the application of delivering 7.1.4 channel content to an endpoint, whereby listeners can select from either a stereo downmix or a binaural presentation.

2. VIRTUALIZER DESIGN METHOD

OptiBRIR represents an idealized virtual room designed to provide a close timbre match to stereo while creating a natural sense of space. To accomplish this, we removed some of the acoustical constraints of actual rooms, to an extent that the virtual room is not physically realizable. Instead, the virtual room aims to capture only the most perceptually-relevant BRIR

features introduced by listener head and room acoustics. Here, our hypothesis was that relaxing the room constraints and applying numerical optimization would expand opportunities for creating a compelling listener experience relative to stereo. We generally avoided selecting model parameters to match any specific room, as our design goal was to deliver robust spatial performance independent of the listening environment.

A stochastic room model was developed for generating synthetic BRIRs that impart the desired perceptual cues. The model combines a direct response with early reflections and a late response that aim to impart the key perceptual cues. The reverberation tails are created from individual reflections having controlled directionality to enhance the listeners' sense of externalization. All room surfaces are modeled as perfect reflectors, and the reflectors for each virtual channel are independent of the others. Among other benefits, this approach yields negligible correlation between reverberation tails for different channels, a feature that minimizes combing artifacts. Furthermore, the use of individual reflections offers a flexible means to achieve both time- and frequency-dependent interaural coherence (IAC). Conventional Feedback Delay Networks (FDNs), although often more computationally-efficient than direct convolution of BRIRs, cannot realize a time-dependent IAC. Time- and frequency-dependent IAC has been found to provide slight performance gain over frequency-dependent IAC in the design of synthetic BRIRs that match the perceptual impression of a measured room response [12].

2.1. Perceptual Cues

A core objective was to identify various BRIR features (perceptual cues) that convey to listeners a sense of sound source externalization while retaining high spectral naturalness. This research was executed through a series of formal and informal subjective experiments. In a first group of experiments, we evaluated the naturalness/externalization tradeoff of various BRIR types and attributes. Virtualized single-channel sound sources rendered with both room measurements and room models were presented to listening subjects. The sound sources included a variety of speech, solo-instrument, and pink noise signals. For each sound source, subjects were asked to record the perceived source location and size in three-dimensional space, and a spectral naturalness rating relative to the unprocessed input signal. Listener assessments were captured with a system called ADA (Adaptive Audio) [13] that allows high-dimensional capture of subjective impressions of sound location and quality. One of the key outcomes from these experiments was that suitably-defined synthetic BRIRs achieved higher externalization with the same spectral naturalness as non-personalized BRIRs from physical room measurements. The subject assessments were made in a different listening environment than the BRIR measurement room. The performance of the measured BRIRs might improve if the experiment was repeated in a consistent listening environment.

In a second experiment, we investigated which time interval of the BRIR reverberation tail had the largest effect on sound source externalization. We began by creating personalized blocked ear canal measurements using loudspeakers at 0º and ±30º azimuth angles for 3-5 human subjects. The BRIRs were measured with the listener in the center of three rooms sized 48, 79, and 550 m^3 and at a distance of 2, 1.4, and 6 meters from the loudspeakers, respectively.

The BRIRs were equalized to compensate for the headphone used in the experiment. Next, the BRIRs were time gated to various lengths between 4–200 ms from the first arrival sound. These BRIRs were used to binaurally render audio sources such as mono female speech and pink noise. The renders were then auditioned over headphones.

Informal experiments compared the perceived distance of the actual loudspeakers and virtualized loudspeakers. In the two smaller rooms, listeners found that the sound moved out of the head if more than 10 ms of the BRIRs were used and the externalization effect stopped increasing after 30 ms for a virtualized noise source and 50 ms for the speech source. In the largest room, the externalization stopped increasing at 70 ms for the speech source. As a result, the OptiBRIR reverberation duration was set to 80 ms to produce well externalized sound sources with minimally audible echos. Note that our results are similar to those independently reported for speech by Catic et al. [14] who found an 80 ms time limit.

A third experiment was an informal test in the 79 m^3 room again but with open canal BRIRs measured using microphones placed 20 mm down the ear canals. Loudspeaker sources were positioned 1.4 m from the listener at 0º, ±30º and ±110º azimuth angles. Additionally, room impulse response measurements were captured using two spaced omnidirectional microphones placed at the ear positions without the

head present. When these impulse response pairs were used to virtualize the audio sources, the sound externalized poorly compared to the measured BRIRs.

A hypothesis was tested that effective externalization is accompanied by significant short-term fluctuations in the early reflection azimuths over time. The apparent direction-of-arrival (DOA) of measured impulse responses was assessed by listening to 4 ms segments. The perceived direction was recorded as the time between the 4 ms segment and first arrival sound increased. We found that strong azimuthal fluctuations of impulse response segments occurring 15-50 ms after the first arrival sound were correlated with higher externalization.

Figure 1 presents the perceived azimuthal fluctuation as a function of time for the human subject BRIRs and spaced microphone impulse responses for a loudspeaker source at 110º to the right of center. The well-externalizing impulse responses had strong directional fluctuation, while the less effective responses did not. The correlation between externalization and strong directional fluctuation was observed for other loudspeaker positions and in another experiment employing a room reflection simulation. These results are independent from but in general agreement with Catic et al. [14] who reported that fluctuation of binaural cues in the BRIRs strongly affects externalization for frontal sound sources.

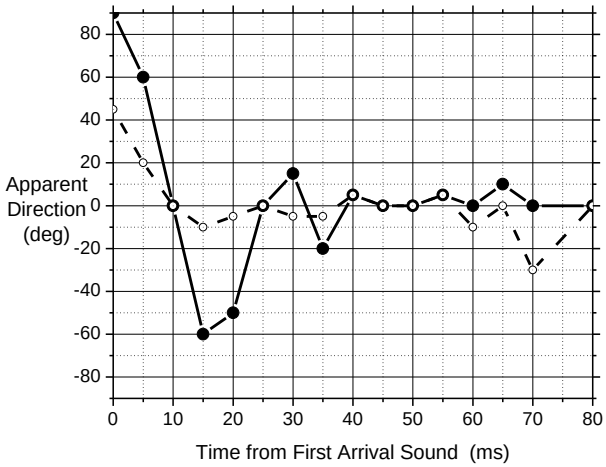

Figure 1: Perceived short-term direction-of-arrival for a human subject BRIR (solid line) and spaced microphone impulse responses (dashed line). Loudspeaker source was 110 degrees to the right of center. The response with higher directional fluctuation was found to externalize more effectively.

2.2. Stochastic Room Model

A stochastic room model was designed by combining the most effective of these deterministic BRIR features with a set of components driven by continuous and discrete random variables. Well-known perceptual cues from room acoustics such as energy decay, interaural coherence (IAC), and echo density are included in the model (cf. [12]). The random variables select individual reflection parameters such as DOA, time delay from the direct response, and energy. In accordance with six-sided physical rooms, the probability of a reflection occurrence increases quadratically with time for increasing echo density. The stochastic model generates multiple BRIR candidates for each virtual channel, all embodying the desired perceptual attributes while differing in the stochastic ones.

Each virtual channel BRIR contains a direct response derived from an HRTF pair associated with the source DOA and distance, followed by synthetic early reflections and late reverberation. The synthetic reverberation is comprised of approximately 3,000 individual reflections derived from an equalized HRTF dataset and then mixed into the BRIR. The reflections are directionally-controlled to enhance the sense of space. The predetermined directional pattern can be, for example, the combination of azimuthal fluctuation plus a diffuse (random) component within a predetermined azimuth/elevation range. The change in reflection direction imparts a time-varying IAC, providing a primary perceptual cue. A further advantage of this method is that each reflection in the BRIR, especially those occurring in the interval from 10-50 ms, contains an accurate interaural time delay (ITD) and frequency-dependent interaural level difference (ILD). To address the well-known virtualizer issue of upward elevation bias in front horizontal sound sources, a portion of the reflections have a direction of arrival from the ground. The BRIR direct, early and late response are all derived from an HRTF dataset obtained from measurements on an artificial head at a source distance of 1.8 m.

2.3. BRIR Optimization

The stochastic model forms the basis of a Monte-Carlo numerical optimization procedure, as outlined in Figure 2. Each one of a number of BRIR candidates are generated by the stochastic room model, converted to a perceptual-domain representation, and then assessed in performance by comparing with a perceptual-domain representation of a desired target response. A suitably-

defined perceptual-domain metric (objective function) provides the basis for deriving a figure of merit. A variety of objective functions and target responses are possible. For example, a target may be generated as a smoothed perceptually-banded representation of the direct response.

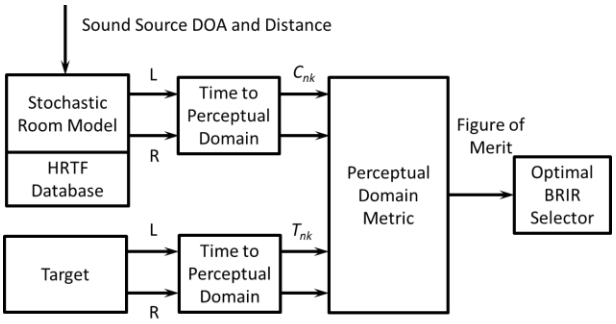

Figure 2: Numerical optimization based on a stochastic room model.

For the purpose of optimizing BRIRs for natural timbre, we have found that a critical band spectrum [15] provides a useful perceptual-domain representation. For an objective function, the weighted mean-squared log-spectral distortion measure is relevant. The critical band energies of the left and right ear BRIR responses C_{nk} are compared against the target critical band spectrum T_{nk} as shown in Eqns. (1-3):

$$D = \sum_{n=1}^{2}\sum_{k=1}^{B} w_{nk}[\log(C_{nk}) - \log(T_{nk}) + g_{\log}]^2 \qquad (1)$$

$$g_{\log} = \sum_{n=1}^{2}\sum_{k=1}^{B} w_{nk}[\log(T_{nk}) - \log(C_{nk})], \qquad (2)$$

where

$$\sum_{n=1}^{2}\sum_{k=1}^{B} w_{nk} = 1. \qquad (3)$$

Here, D is the total weighted mean-square log-spectral distortion for both ears (figure of merit), and B is the number of critical bands. To account for intra- and interaural level differences, the mean-square log-energy difference for each ear and band is scaled by a weighting factor w_{nk}. The weighting factors are proportional to an estimate of relative loudness derived from the critical band energies. In Eqn. (2), $g_{\log}$ represents a broadband gain offset between C_{nk} and T_{nk} computed to minimize D. The distortion metric expressed in Eqn. (1) is sensitive to the degree of spectral combing introduced by the early reflections and late reverberation. The candidate minimizing the objective function (and hence the timbral distortion) is selected as the survivor BRIR for that virtual channel. The final 7.1.4 virtualizer is defined by the 11 survivor BRIRs, each representing the highest degree of spectral naturalness amongst the candidates.

As a quantitative illustration of the results of numerical optimization, Figure 3 presents the spectral deviation from a target response for the worst-case and survivor BRIRs. The worst-case and survivor BRIRs are defined as the candidates having the highest and lowest log-spectral distortion D, respectively. The target and BRIR spectral magnitudes were computed using a DFT and then smoothed on a half-critical band scale prior to subtraction. The direct-to-late ratio was 13 dB, and the optimization considered 100 candidates. Spectral differences for the worst-case candidate show pronounced spectral combing at frequencies below 1 kHz, which impairs rendering quality primarily on speech and music signals that have pronounced harmonic structures. Spectral differences for the survivor BRIR are noticeably reduced.

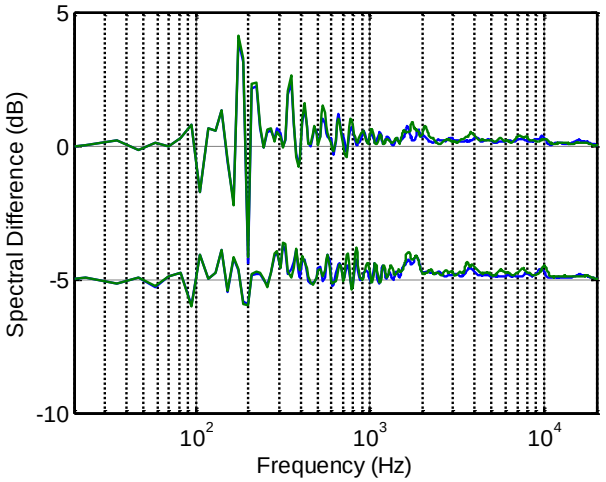

Figure 3. Spectral deviation from the target response for two BRIR candidates. The topmost blue/green lines represent the left- and right-ear differences for the worst-case (rejected) candidate, while the lines below represent the survivor BRIR. Spectral differences for the survivor BRIR have been offset 5 dB for clarity.

2.4. Algorithm Tuning

To support algorithm parameter selection, we conducted multiple small-scale (4-7 listeners) single-interval A:B preference tests of the 7.1.4 virtualizer against stereo. Algorithm changes were made after each test in response to the results. The objective was to drive the design not only by listener preference ratings, but also by written comments explaining why one version was preferred over another. Naturally not everyone agreed on areas of strength and weakness for each system; however, clusters of listener preferences emerged. The feedback was instrumental in increasing the overall listener preference ratings.

3. PREFERENCE TEST

The methodology selected for this test was a single-interval A:B preference test. This forced-choice testing methodology was identified for this evaluation to satisfy multiple testing requirements. Desired test constraints included: a) identification of listener preference in the absence of a reference target, b) controlled content playback and randomized presentation of sequences, c) double-blind representation of the two systems at test, and d) reduced bias and criterion effects associated with use of rating scales where multiple stimulus attributes are considered in the system assessment. Forced-choice tests are known for demonstrating increased sensitivity to system differences. In this way, use of a forced-choice testing paradigm is a good alternative for tests of multidimensional system experiences where the user response target is not identification of a specific impairment artifact and is, instead, making a judgment on a qualitative holistic experience. Inclusion of rating scales for these types of judgments can introduce more internal noise and response variance and/or bias in cross-user data collection.

Regarding item (a), many previous virtualizer tests have used fixed references as key qualitative comparisons in the evaluations. For example, a well-known study performed by the BBC [4] compared different virtualizers to stereo and mono and included a dialog reference for many of the test conditions. This inclusion established a specific qualitative dimension that users were to listen for rather than the overall quality of the experience of the content. Regardless of instructions given the listener, inclusion of the reference introduces an experiential comparison target. The tests used in this study were reference free with the acknowledgement and discussion of the different feature attributes contributing to overall quality included as part of the listener instructions and training. This method allowed users to evaluate the overall quality of the experience free of the targeted performance of a single feature dimension (e.g. dialog).

3.1. Audio Sequences

The 13 audio sequences selected for this test all originated from Dolby Atmos printmasters. Three printmasters were used in a listener familiarization phase, and the other ten were used in a grading phase. Ten are generally regarded as known "critical material," meaning that they frequently reveal differences among signal processing systems under test. Two were added as dialog+effects to test the important and challenging application of spatialized voice, and the last item was a helicopter flyover to evaluate the effect of elevated sound objects. The ten items used for the grading phase are described in Table I.

Item #	Description
1	Bird in flight with jungle ambience.
2	Female and male cinematic dialog with subtle outdoor ambience.
3	Bass-heavy electronic music with male voice narration at a low level.
4	Helicopter takeoff and flyover.
5	Fixed and panned clock chimes, mechanical sounds, gears, and bells with strong transients.
6	Panned creature dialog with strong cave reverberation. Subtle running water sounds.
7	Forest ambience with numerous wind sound effects.
8	Heavy rainfall with subtle thunderclap.
9	Electronic music with tempo-synchronized panned percussive elements. Cheering crowd and applause ambience.
10	Male voice-over narration with subtle, sparse ambience.

Table 1. Description of 10 audio sequences used in the listener grading phase.

The printmasters were rendered to 7.1.4 channel files and then in turn rendered to LoRo and binaural. Since there are no standardized downmix coefficients for 7.1.4 to stereo, the LoRo downmix coefficient for each input channel was selected to provide approximately the same

perceived loudness as the associated virtualizer channel. This downmix strategy minimized loudness differences for each audio object in the stereo and binaural presentations. The low-frequency effects channel was attenuated by 3 dB and mixed equally into left and right channel binaural and LoRo content.

3.2. Leveling and Equalization

The LoRo references were leveled with respect to each other. All binaural audio sequences were then time-aligned and leveled to their associated LoRo reference using an excitation-based loudness estimation algorithm. Both LoRo and binaural sequences were then equalized for playback on Stax electrostatic headphones. The equalization filter was designed using measurements from 7 heads, 3 headphone placements each. A group of expert listeners verified that the loudness was consistent by listening prior to running the test.

3.3. Test Interface and Signal Path

The test was run using custom software called LisTest running on a Windows PC. The PC was connected via USB to a Grace m903 DAC, Stax SRM-252S headphone amplifier, and Stax SR-207 headphones. Volume was set to a fixed output level and listeners were not allowed to adjust it.

3.4. Test Subjects

Ten expert listeners (6 women, 4 men, age range: 21-47, average age: 30) having no affiliation with Dolby completed the test. These individuals had never heard the test content, and had no familiarity of the OptiBRIR design, the method of stereo downmixing, or the room in which the test was conducted. All of the listeners who completed the test have previously participated in several headphone preference tests and all are audio professionals (audio engineers, musicians, etc.).

3.5. Test Procedure

The test was divided into two phases – the familiarization phase and the grading phase. Each phase was comprised of a set of trials. In each trial, listeners were presented a stereo-binaural audio file pair described as A and B, and then asked to indicate a preference based on overall quality of the experience. To reduce experimental bias effects, the trial presentation order and stereo-binaural presentation order were randomized across listeners, and each stereo-binaural pair was presented twice with opposite orders.

The 6 trial familiarization phase was conducted first and implemented for three reasons:

- Familiarize listeners with the test interface UI and methodology
- Familiarize listeners with the type of cinematic and musical content that would be presented in the test
- Provide a period of time allowing listeners to adapt to the set of HRTFs embodied in OptiBRIR.

Additionally, this phase exposed the listener to the two conditions under test. The results for this phase were for training and familiarization only and were not used in the final analysis.

For the 20 trial grading phase, the listeners were instructed to listen to all conditions in their entirety at least one time through before indicating a preference. Looping of the audio clips was encouraged, as well as real-time switching between the two clips during playback to help them make a more thorough comparison of A and B. Clips varied in length, but none of the clips in the grading phase was longer than 19 seconds. Participants were not allowed to advance to the next trial without selecting a preference. The total duration of the two phases was approximately 60 minutes.

4. RESULTS

Results from the single-interval A:B preference test are shown in Figures 4-6. Total listening time per system and content was recorded and assessed following testing to ensure adequate time was spent to evaluate each system. Across all content, listeners spent 23.7 +/- 5.5 seconds on OptiBRIR clips on average, and 18.5 +/- 5.1 seconds for stereo.

Content rendered through OptiBRIR was preferred over stereo for 60%-95% of listeners, with average overall listener preference of 75% (Figures 4 and 6, n=10 listeners). Only one content item, Item 3, was equally split (50%/50%) among listeners' virtualizer preference. All other items showed a majority preference for OptiBRIR, with Items 1 and 4-9 preferred significantly higher than stereo ($p<0.05$, Student's t-test, n=10 listeners). OptiBRIR was overall preferred across

content for all but one listener, with 8/10 listeners significantly preferring it over stereo ($p<0.05$, Student's t-test, n=10 items).

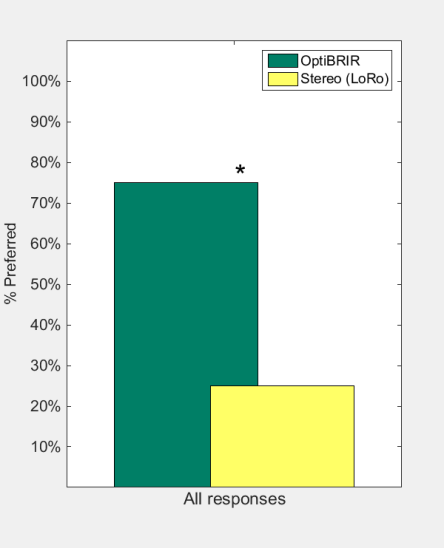

Figure 4: Overall preference test results. Asterisk indicates ratings that are significantly different using Student's t-test ($p<0.05$).

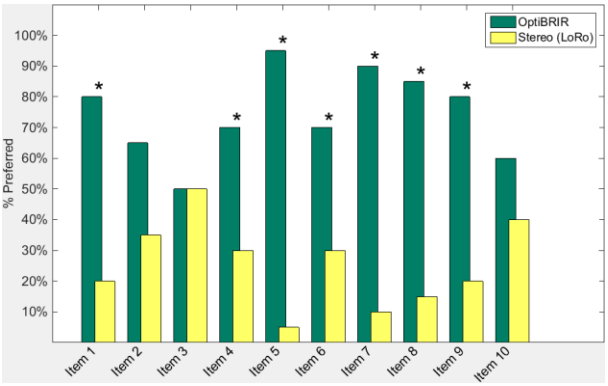

Figure 5: Preference test results by audio test item. Asterisks indicate ratings that are significantly different using Student's t-test ($p<0.05$).

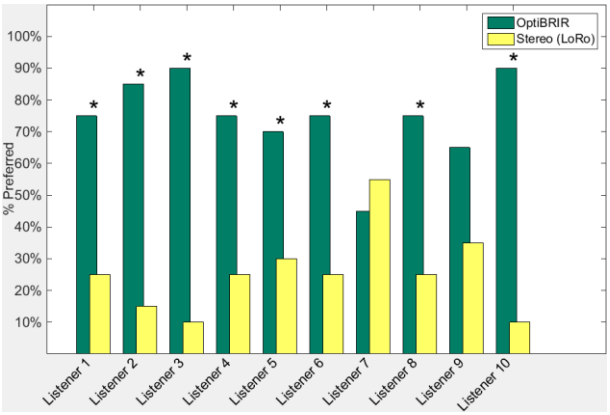

Figure 6: Preference test results by listener. Asterisks indicate ratings that are significantly different using Student's t-test ($p<0.05$).

5. SUMMARY AND DISCUSSION

System preference was dependent on the specific content that was presented: all listeners favored Items 1 and 4-9 rendered through OptiBRIR over stereo by a statistically significant margin ($p<0.05$, Student's t-test, n=10). These items represented a variety of content, from cinematic scenes with detailed spatialization (Items 1 and 7) to music (Item 9). The synthetic Items 4, 5, and 8 were included as known challenging items that would stress the system's ability to localize sound sources and preserve a sense of scene, elevation, and staging.

The remaining content, Items 2, 3, and 10, were not statistically different, although the trend favored OptiBRIR in all cases. Interestingly, the non-significant items all feature prominent dialog. Dialog is especially sensitive to perceived degradation with timbre changes and coloration, so the OptiBRIR design approach was supported by the observation that it performed as well as stereo for dialog. Furthermore, for Item 6 containing dialog that is heavily affected by cave reverberation and sweeping overhead pans, OptiBRIR was strongly preferred over stereo (70%).

We considered how downmixes created directly from the Dolby Atmos audio objects might mitigate some of the unwanted effects of conventional channel-based LoRo downmixing, such as left-right ear pressure differentials caused by the lack of crosstalk. However, we noticed that the two items with the most noticeable pressure differentials, Items 4 and 9, received about the same relative preference grades as content without

audible pressure differentials. Furthermore, in our earlier preference tests, some listeners reported preferring the hard pans introduced by LoRo stereo. For instance, Item 4 rendered in stereo resulted in a prolonged full left pan and associated extreme localization; however, the full Dolby Atmos rendering positions the helicopter elevated and partially to the left, which was more accurately conveyed by the virtualizer. Therefore, we were unable to find any clear evidence that the LoRo pressure differentials significantly affected the overall preference ratings. A second test would be needed to definitively answer this question.

These test results are a promising step towards the goal of creating a headphone virtualizer that markedly outperforms a channel-based LoRo downmix. In our view, this can be accomplished by supporting highly immersive content, preserving stereo attributes that listeners value most, and rendering to impart a natural sense of realism (soundstage depth, height, immersion, etc.). Further gains in spatial performance are possible by direct binaural rendering of individual audio objects from the Dolby Atmos printmaster, rather than from an intermediate channel-based format as in this study. Direct object processing allows individual sound sources to be rendered more accurately at arbitrary locations, rather than relying on phantom imaging from multiple fixed-position virtual channels.

The specific components leading to high preference for the virtualizer were not fully explored. As an area for future research, a larger test (20 listeners) could be conducted to examine preference alongside specific audio attributes such as brightness, loudness, sensation of height, spectral naturalness, and spaciousness. This approach will determine if parameters that we attempted to align between stereo and binaural, like brightness and loudness, are in fact not a significant basis for differentiation. Other factors, like spaciousness and height, may underlie listener's preference for the virtualized listening experience.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions of Dr. Kuan-Chieh Yen to the formative concepts of this research. The authors also express their sincere appreciation to the 40 Dolby volunteers who participated in the small-scale single-interval A:B preference tests. The preference ratings and the written comments received were very insightful and in fact crucial to understand how the virtualizer design should evolve in order to increase the group preference relative to stereo.

7. REFERENCES

[1] G. Lorho, D. Isherwood, N. Zacharov, J. Huopaniemi, "Round Robin Subjective Evaluation of Stereo Enhancement Systems for Headphones," *AES 22nd International Conference: Virtual, Synthetic and Entertainment Audio*, June 2002.

[2] G. Lorho and N. Zacharov, "Subjective Evaluation of Virtual Home Theatre Sound Systems for Loudspeakers and Headphones," *116th AES Convention*, May 2004.

[3] G. Lorho, "Evaluation of Spatial Enhancement Systems for Stereo Headphone Reproduction by Preference and Attribute Rating," *118th AES Convention*, Barcelona, Spain, May 2005.

[4] C. Pike and F. Melchoir, "An Assessment of Virtual Surround Sound Systems for Headphone Listening of 5.1 Multichannel Audio," BBC Research and Development White Paper, October 2013.

[5] O. Kirkeby, "A Balanced Stereo Widening Network for Headphones," *AES 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, June 2002, conference paper 000249.

[6] E. Manor, W. Martens, A. Marui, D. Cabrera, "Nearfield Crosstalk Increases Listener Preferences for Headphone-Reproduced Stereophonic Imagery," *Journal of the Audio Engineering Society*, Vol. 63, May 2015.

[7] S. Olive, "Evaluation of Five Commercial Stereo Enhancement 3D Audio Software Plug-ins," *110th AES Convention*, May 2001, convention paper 5386.

[8] R. Mason and F. Rumsey, "An Assessment of the Spatial Performance of Virtual Home Theatre Algorithms by Subjective and Objective Methods," *108th AES Convention*, Paris, Feb. 2000.

[9] F. Rumsey, S. K. Zieliński, R. Kassier, and S. Bech, "On the Relative Importance of Spatial and Timbral Fidelities in Judgments of Degraded

Multichannel Audio Quality," *J. Acoust. Soc. Am.*, Vol. 118, August 2005.

[10] C. Robinson, S. Mehta, and N. Tsingos, "Scalable Format and Tools to Extend the Possibilities of Cinema Audio," *SMPTE Motion Imaging Journal*, 121(85), pp. 63–69, 2012.

[11] Recommendation ITU-R BS.775-3, "Multichannel Stereophonic Sound Systems with and without Accompanying Picture," *International Telecommunications Union, Radiocommunication Assembly*, 2012.

[12] F. Menzer, "Binaural Signal Processing Using Interaural Coherence Matching," Thèse No. 4643 (2010), Ecole Polytechnique Federale de Lausanne, April 2010.

[13] D. Darcy, K. Terry, G. Davidson, R. Graff, A. Brandmeyer, P. Crum, "Methodologies for High-Dimensional Objective Assessment of Spatial Audio Quality," accepted for publication at the *140th AES Convention*, Paris, June 2016.

[14] J. Catic, S. Santurette, T. Dau, "Role of Reverberation-Related Binaural Cues in the Externalization of Speech," *J. Acoust. Soc. Am.*, Vol. 138, August 2015.

[15] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed., Emerald Group Publishing Limited, United Kingdom, 2012.