Brief Papers

Results From Subjective Testing of the HD Codec at 16–96 kbps

Ellyn G. Sheffield, John Kean, Mike Starling, Jan Andrews, Kyle Evans, and Sunny Khemlani

***Abstract*—As continuing support of National Public Radio's Tomorrow Radio project, this paper presents results from a consumer study conducted with iBiquity's HD Radio codec. Forty participants judged the quality of the HD codec (HDC) at various bit rates, ranging from 16 kbps to 96 kbps. A range of musical and speech genres were tested. This study provides in-depth information intended to help broadcasters select optimal bit rates when HD Radio's 96 kbps data stream is shared between primary and supplemental channels.**

***Index Terms*—Codec, consumer testing, perceptual testing.**

I. Introduction

With the introduction of HD Radio, important questions have arisen concerning optimal allocation of the 96 kbps data stream. To understand how allocation schemes would potentially affect consumer satisfaction and listening behavior, rates from 16 to 96 kbps were incrementally tested in a subjective audio experiment. The study was designed to explore whether: (a) listeners could detect quality differences in HDC at specific bit rates; (b) listeners rated these differences as meaningful and significant; and (c) listeners would change their listening behavior based on differences in quality.

The study was conducted in two phases during the months of July and August. The first phase narrowed the field of testable bit-rates in order to limit the number of test conditions on which the general public would be tested. This phase was conducted with a small sample of NPR audio engineers and personnel. The second phase was designed to obtain mean opinion scores (MOS) for a wide range of HDC bit-rates and to test specific bit-rate comparisons that were found to be of interest from phase 1 testing. This phase was conducted with 40 listeners from the general public.

A. Testing Environment

Testing was conducted at National Public Radio, Washington, DC, in Studio 4A. The test area was approximately 16 m $\times$ 10 m, with a ceiling height of 4.6 m. The observed Preferred Noise Criteria for the studio was measured at PNC-19.

The studio was divided into six listening stations. Audio samples were presented to listeners binaurally over Sennheiser HD-600 open-backed headphones. Because the audio samples were delivered over open-back headphones, there was concern that leakage would create audio interference between participants. Therefore, large foam blocks, measuring 4 feet square by 2 feet thick, separated stations from each other. The blocks, fabricated of 2 lb. per cubic foot open cell urethane foam, were stacked 4 feet high, providing acoustic and visual insulation.

B. Source Material and Sample Presentation to Listeners

For both phases of this study, source material was taken from previous NRSC test material, NPR and Sun Sounds of Arizona programming material, and music CD's. Speech, voice-overs, and music (rock, jazz and classical) were included.

The playback of samples to listeners was controlled using a software package developed by iBiquity Digital Corporation. Sound samples were presented to listeners individually. The software collected and stored listener responses, requiring no experimenter control or interaction once the testing session began. Participants were free to take the test at their own pace, and were given instructions to play samples as many times as necessary to make their decision.

C. Preparation of HD Codec Audio Samples

Audio samples were prepared on the FM test bed shown in Fig. 1. The system produced a hybrid digital and analog FM-band signal with stereo subchannels in compliance with the FCC rules and applicable standards.[1]

The test bed passed audio samples from an audio CD through a transmission/receiving chain. The resulting HD-encoded and decoded audio was recorded on audio CD, for later transfer to playback equipment used by the listeners. The stereo generator and analog Host FM generator side chain did not contribute to the audio sample transfer. It was included only to provide compliance with hybrid DAB transmission standards.

Audio transferred through the IBOC DAB side chain remained digital at all times. Playback of samples from CD were connected by AES/EBU link, at 44.1 kHz sampling rate, to a digital audio processor. Broadcast audio processors were provided for this test by Omnia (6EX-HD), Optimod (8400HD) and Harris/Neural (Neustar). The stereo generator portion of the Omnia or Optimod provided an analog stereo signal for the FM Host generator. These processors were used in the production of HDC-coded samples for Phase 1 testing, which are not reported herein. For the main Phase 2 testing project the digital audio from the CD player was fed directly to the IBOC

Manuscript received October 4, 2004. This work was supported by a grant from the Corporation for Public Broadcasting.

E. G. Sheffield is with the Department of Psychology, Salisbury University, Salisbury, MD USA (e-mail: egsheffield@salisbury.edu).

J. Kean, M. Starling, J. Andrews, K. Evans, and S. Khemlani are with the National Public Radio, Washington, DC USA (e-mail: jkean@npr.org; mstarling@npr.org).

Digital Object Identifier 10.1109/TBC.2006.875610

[1]Transmission standards for the analog Host stereo signal are prescribed by 47 CFR 73.322. As of this writing, rules for IBOC DAB are awaiting FCC adoption. However, transmission standards for the iBiquity system are detailed in Appendix B of the First Report & Order, MM Docket 99.325

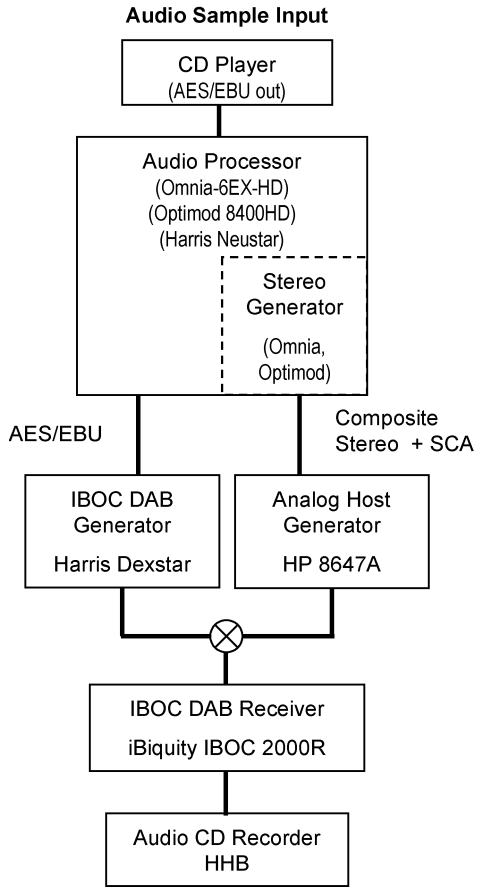

Fig. 1. Basic equipment configuration to prepare HD codec audio samples.

DAB Generator, completely bypassing the audio processors. All of the Phase 2 audio HD Codec samples were unprocessed, thereby providing comparability to the CD source references. However, care was taken to match loudness levels between all the samples and ensure that peak levels did not reach 0 dBFS.

II. Phase 1—Narrowing the Field of Bit-Rate Comparisons

Ten NPR employees participated in Phase 1. Listeners were presented with 98 sample pairs (i.e., two samples, back-to-back), and were asked the following questions about each pair:

- Which sample had **better** audio quality, "A" or "B"?
- How big was the difference, on a scale of 1–10, with 10 being "extremely different", 5 being "different", and 1 being "I really couldn't tell a difference but you made me pick"?
- Were you dissatisfied with either sample "A" or "B" (or both)?

Listeners were encouraged to play the samples as many times as they needed to make these determinations. Allowing unlimited access to sample-pairs afforded participants the greatest opportunity to discern small differences between the samples. Thus, we believe that their response data represents an extremely precise and stringent discrimination measure. Sample-pairs were randomized, such that each participant heard the pairs in a different order; pairs were counterbalanced, such that for half the pairs, the lower bit-rate was sample "A", and for the other half, the lower bit-rate was sample "B". Table I shows the samples pairs used.

TABLE I
Sample Pairs Used in Phase 1 Testing

24 kbps	36 kbps	48 kbps	56 kbps	64 kbps	72 kbps
24 vs. 36	36 vs. 48	48 vs. 56	56 vs. 64	64 vs. 72	72 vs. 80
24 vs. 48	36 vs. 56	48 vs. 64	56 vs. 72	64 vs. 80	72 vs. 96
	36 vs. 64	48 vs. 72		64 vs. 96	

A. Results for Phase 1 Testing

Paired t-tests were conducted to see if the percentage of respondents claiming that the higher bit-rate sounded better than the lower bit-rate was statistically different from chance, or 50%. At lower bit-rates, listeners were able to accurately report that the higher bit-rate sounded better than the next adjacent bit-rate (24 vs. 36; and 36 vs. 48). However, with one exception (64 to 80 kbps), at mid-range bit-rates and above, listeners were unable to reliably tell the difference when the samples differed by 8 or 16 kbps.

Participants were additionally asked how big the difference was between the audio samples in an audio pair on a 1–10 scale, 1 being no difference at all; 5 being a difference; 10 being extremely different and noticeable. Participants claimed that they perceived larger differences between lower bit rate pairs (ranging from 3.11 to 5.06) and smaller differences at the higher bit rates (ranging from 2.20 to 3.02).

Finally, listeners were asked whether they would continue to listen to sample "A", sample "B", "neither" or "both" at various bit-rates. Over 40% of participants were dissatisfied with samples coded at 24 and 36 kbps, but this number dropped substantially to 15% at 48 kbps. This indicates that somewhere around 48 kbps a large majority of listeners begin to react favorably to HDC coded material.

Based on the NPR listeners' results, we selected a sub-sample of bit-rate comparisons for inclusion in Phase 2, consumer testing. We included two low bit-rate comparisons that were reasonably large (i.e., 24 $\times$ 36; 24 $\times$ 48), to see if the general public corroborated NPR listener views; and we included two additional comparisons that were potentially important to the allocation of 96 kbps (i.e., 48 vs. 64; 48 vs. 96). Finally, we included 64 vs. 96 to replicate test conditions in NRSC FM testing. Because Phase 1 indicated that NPR listeners could not reliably discern differences between 48 and 56 kbps, and 48 and 72 kbps, we did not include those comparisons in general consumer testing. However, we did include a 48 vs. 96 kpbs bit-rate pair to see if consumers could hear a difference between the more disparate bit-rates.

III. Phase 2—Consumer Testing

A. Participants

Fifty-nine total listeners (29 males and 30 females) initially participated, distributed between 18 and 65 years of age. Subjective data from 40 qualified listeners was collected, where qualification was based on performance on the initial screening test and a post-hoc screening test designed to eliminate outliers. Table II shows the demographic breakdown of general public

TABLE II
DEMOGRAPHIC BREAKDOWN OF PARTICIPANTS INCLUDED IN RESULTS

Age	Female	Male
18-29	6	6
30-39	5	4
40-49	5	4
50+	4	6

listeners. Listeners were recruited from several sources, including friends and family members of NPR staff, flyers posted in the downtown Washington area and outlying suburbs, and on-line postings.

B. Design and Procedures

Participants were tested individually over Sennheiser HD-600 headphones for approximately 2 1/4 hours. The test session included a screening test, a single stimulus, Absolute Category Rating Mean Opinion Score (ACR-MOS) test and a double stimulus, A/B comparison test on selected sample pairs. Because it has been argued that the ACR-MOS is not as sensitive to differences as directly comparing one audio sample to another the A/B comparison test methodology was used on pairs of extreme interest.

C. Screening Test

Screening was conducted to ensure that listeners could reliably distinguish between significantly different audio qualities. There were seven screening trials. For each trial, participants were asked to listen to 3 samples, 2 of which were the same and the 3rd different (for example, 2 female speech source samples and the same female speech sample processed through an FM receiver). The listener's task was to decide which of two "test" samples ("A" or "B") was different from the reference sample. In each trial, the first sample they heard was always the "reference" sample. They then listened to the "A" and "B" samples and judged which of the samples was different from the reference. Listeners were free to replay any or all of the three samples until they were ready to enter their response and proceed to the next trial. In order to "pass" the screening test, participants had to answer 6 of 7 screening triads correctly.

D. Main Tests

In the ACR test, participants listened to 200 samples, one-by-one, and rated each sample individually. The ACR test yielded a Mean Opinion Score (MOS), a measure of overall audio quality. Listeners were required to judge the quality of an audio sample using a five category rating scale (Excellent=5, Good=4, Fair=3, Poor=2, and Bad=1). Listeners controlled playback of the audio samples but were not allowed to register their answer until the entire sample was played. Listeners were given the opportunity to adjust the playback volume during one practice trial, and this level was maintained throughout the remainder of the experiment.

In the double stimulus test, participants were given 30 sample-pairs and asked the same three questions that Phase 1 listeners were asked. Table III lists the sample pairs participants were asked to rate.

TABLE III
SAMPLES USED IN A/B TEST

	24 vs. 36	24 vs. 48	36 vs. 64	48 vs. 64	48 vs. 96	64 vs. 96
Speech (1M; 1 F)	2	2	2	2	2	2
Classical	1	1	1	1	1	1
Rock	1	1	1	1	1	1
Jazz	1	1	1	1	1	1

TABLE IV
MEAN OPINION SCORES

	Classical	Jazz	Rock	Speech	Voiceover
16	2.8*	3.3*	2.5*	2.0*	2.4*
24	3.2*	3.7*	3.1*	2.9*	3.0*
36	4.0	4.0	3.7*	3.4*	3.2
48	4.0	4.0	3.9	3.7	3.3
56	4.0	4.1	3.9	3.8	3.5
64	4.1	4.1	4.0	3.7	3.5
72	4.1	4.2	4.0	3.8	3.5
80	4.0	4.1	4.1	4.0	3.4
96	4.1	4.2	4.1	3.9	3.5
Source	4.1	4.2	4.2	4.1	3.4

*Significantly different from 96

E. Results for Phase 2 Testing

Preliminary analyses were conducted to examine whether participants rated audio quality of samples differently based on their age or gender. A 2 (gender) $\times$ 4(age) ANOVA yielded a main effect of age, but no main effect of gender. Newman-Keuls Multiple-Comparison tests indicated that, as with past audio testing, older participants rated samples less critically than younger participants. The range of mean scores, however, was rather small between the youngest and oldest groups: 18–29 year old participants' mean was 3.5; 50+-year-old participants' mean was 3.8. In this study, females and males rated samples similarly. Thus, because differences were minimal, participants' data was combined for all other analyses and total results are reported.

Table IV shows the results from ACR-MOS testing, listed by genres. A one-way analysis of variance (ANOVA) was conducted for each genre to see if the scores at various bit-rates were significantly different from each other. These analyses yielded significant differences, which are highlighted on the table by asterisks. Speech was divided into male and female speech in order to examine it more closely. Table V shows slight differences between participants' scores for female and male speech. For female speech, 48 kbps is significantly different from the reference (but not from 96, 80, 72, 64 or 56), whereas with male speech 48 kbps is significantly the same as all of the higher bit rates.

Taken together, these results suggest that there is a difference in people's perception of quality at lower bit rates than at higher bit rates, and that this difference emerges around 36 to 48 kbps, depending on the genre. With the exception of female speech, participants reported quality parity until 36 kbps. At 36 kbps,

TABLE V
MOS FOR FEMALE AND MALE SPEECH

	16	24	36	48	56	64	72	80	96	Source
Female	1.8*	2.8*	3.3*	3.5*	3.6	3.6	3.7	3.9	3.8	4.1
Male	2.1*	3.0*	3.6*	4.0	4.0	3.9	4.0	4.0	4.1	4.1

*Significantly different from reference

TABLE VI
RESULTS FROM A/B TESTING

Bit rates	% respondents claiming higher bit-rate sounded better	t-test	prob
24 vs. 36	77%	t = 9.2935	.0001
24 vs. 48	80%	t = 10.865	.0001
36 vs. 64	64%	t = 4.1145	.0001
48 vs. 64	54%	ns	
48 vs. 96	56%	t = 1.8491	.03
64 vs. 96	57%	t = 1.9932	.02

TABLE VII
SIZE OF DIFFERENCE WHEN QUALITY WAS IDENTIFIED CORRECTLY

Bit-rate	Size of difference – Phase 2 listeners	Size of difference – Phase 1 listeners
24 vs. 36	5.23	4.17
24 vs. 48	5.27	5.06
36 vs. 64	2.75	3.53
48 vs. 64	2.13	2.51
48 vs. 96	2.55	Not given during phase 1
64 vs. 96	2.04	2.49

TABLE VIII
PARTICIPANTS CLAIMING DISSATISFACTION

Bit rate	Phase 2 dissatisfaction	Phase 1 dissatisfaction
24	34%	43%
36	21%	43%
48	20%	15%
64	15%	17%
96	16%	17%

participants' scores ranged from "fair" (3.0–3.5) in voiceover and speech, to "good" in classical and jazz (4.0). Notice that at the lowest bit-rates the quality ratings dropped dramatically: at 24 kbps, participants rated most genres as "fair", and at 16 kbps participants rated samples between "poor" (2.0) and "fair" (3.3).

F. A/B Test

Table VI shows results for which sample had better audio quality. Paired t-tests were conducted to see if the percentage of respondents claiming that the higher bit-rate sounded better than the lower bit-rate was statistically different from chance, or 50%. Listeners were able to correctly identify the higher bit rate of the bit-rate pair at very low bit-rates.

A large majority of participants heard differences between 24 and 36 kbps; 24 and 48 kbps; and 36 and 64 kpbs. There was no significant difference at 48 vs. 64 kbps, but a slight majority accurately reported hearing differences between 48 and 96 kbps and 64 and 96 kpbs. The t and p values indicate that while significantly different from chance, the percentage of people accurately reporting differences was minimal.

As in phase 1, phase 2 participants were also asked how big the difference was between the audio samples in an audio pair on a 1–10 scale, 1 being no difference at all; 5 being noticeable; 10 being extremely different and noticeable. Table VII shows these results for sample-pairs that were identified correctly. Notice that participants claimed larger differences at lower bit-rates and smaller differences at higher bit rates. Further, a comparison of results from both phases indicates that NPR listeners and general public listeners rated the size of the difference similarly.

Finally, listeners were asked whether they were satisfied with sample "A", sample "B", "neither" or "both" at various bit-rates. Table VIII shows the rate of "dissatisfaction" for each bit rate, with column three (phase 1 dissatisfaction) taken from phase 1 participant scores. Notice that dissatisfaction remained relatively constant from 48 kbps to 96 kbps (the highest reference), suggesting that a portion of the measure of listener dissatisfaction was influenced by participant's opinions of program content. Fewer general public listeners reacted negatively to samples coded at 24 and 36 kbps than did NPR participants, but at 48, 64 and 96 kbps, the numbers were statistically equivalent. Nonetheless, taken together, these results corroborate MOS and A/B findings that only around 36 kpbs participants' dissatisfaction begins to climb in a significantly noticeable way.

IV. CONCLUSION

Results from ACR-MOS and A/B testing support the notion that for most music and speech listeners either do not notice differences between HDC bit rates of 48 kbps or higher, or notice very small differences. However, participants do notice significant differences at lower bit-rates of 16, 24, and 36 kbps. As with previous testing, participants were more sensitive to differences when rating speech than when rating music and voiceovers. This may be due to masking effects (i.e., there is less masking of digital artifacts with speech) or because humans are particularly sensitive to voices and voice quality. In any event, results from this study indicate that for the HDC coder it is possible to separate 96 kbps into two 48 kbps streams with minimal disturbance to listeners. Interestingly, when making choices about bit allocation for HDC, it is apparent that music may require fewer bits than speech for transparency.

ACKNOWLEDGMENT

The authors gratefully acknowledge the efforts of the audio engineers and personnel at NPR who donated their time to help them with the phase 1 testing. Additional thanks to iBiquity Digital Corporation for providing some of the equipment used in the preparation of audio samples. This study would not have been possible without the generous grant from the Corporation for Public Broadcasting.