Perceptual Tests of iBiquity's HD Coder At Multiple Bit Rates

Prepared for National Public Radio

October 14, 2004

by
Ellyn G. Sheffield, PhD
Sheffield Audio Consulting
Princeton, New Jersey 08540
(609) 737-9451
egsheffield@comcast.net

1. Introduction

With the introduction of HD Radio, important questions have arisen concerning optimal allocation of the 96 kbps data stream. This study was motivated by National Public Radio's (NPR) interest in exploring consumer acceptance of iBiquity's HD Radio coder (HDC) at multiple bit-rates in order to recommend to NPR Member Stations the best allocation schemes available for primary and secondary audio channels, given a total stream of 96kbps. Due to the variety of programming in today's marketplace and the flexibility of iBiquity's system, an exhaustive study of all bit-rate combinations was not possible. Therefore, in order to quantify consumer satisfaction and to establish patterns of potential consumer behavior, bit rates from 16 to 96kpbs were incrementally tested over a range of musical and speech genres typical to broadcast radio.

Specifically, this study was designed to explore whether:

(a) general public listeners could detect quality differences in the HD coder at particular bit-rates;
(b) listeners rated these differences as meaningful and significant;
(c) listeners would change their listening behavior based on the differences in quality.

The study was conducted in two phases during the months of July and August, 2004. 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 absolute category rating mean opinion scores (ACR-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. The details of both test phases are described in the remaining sections of this report.

2. Test Methodology

2.1 Test environment

Testing was conducted in a 1,700 square foot sound studio at National Public Radio, Washington, DC. The studio is approximately 53 x 32 feet, with a ceiling height of 15 ft. The ceiling has a spring-isolated acoustic lid at 18 ft., and the walls are built of concrete block. They sit on a 4-inch thick "poured in place" floating concrete floor slab. The observed 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 listening 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 isolation between the listening stations.

2.2 Audio Samples

For both phases of this study, sound samples were taken from NRSC test material, NPR and Sun Sounds of Arizona program material, and music CD's. Speech, voice-overs, and music (rock, jazz and classical) were included. Appendix 1 lists all of the samples.

Preparation of HD Coder Audio Samples

Audio samples were prepared on the FM test bed shown in Figure 2.2. The system produced a hybrid digital and analog FM-band signal with stereo subchannels in compliance with the FCC Part 73 rules and applicable industry standards.[1]

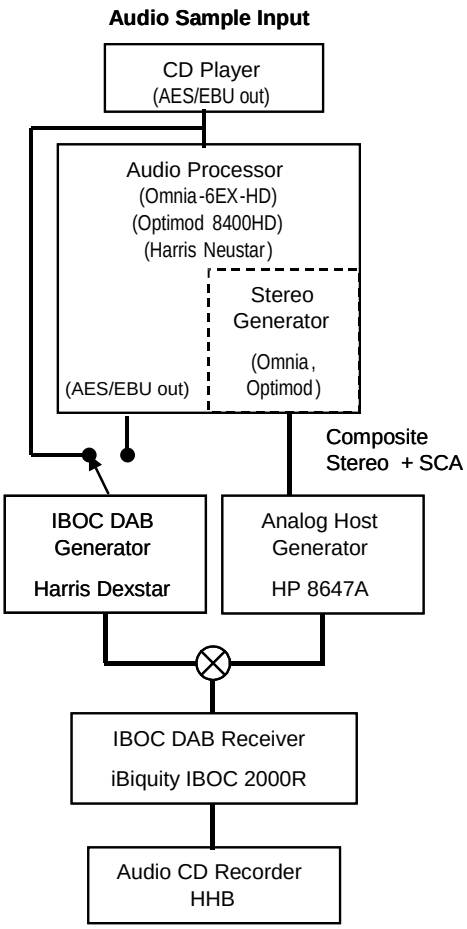

Figure 2.2: Basic equipment configuration to prepare HD Coder audio samples

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

[1] Transmission standards for the analog Host stereo signal are prescribed by 47 CFR 73.322. As of this writing, detailed service 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

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 DAB Generator, completely bypassing the audio processors. All Phase 2 audio HD Coder 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.

2.3 Presentation software

The playback of samples to listeners was controlled using a software package developed by iBiquity Digital Corporation, which has been utilized in prior testing submitted to the National Radio Systems Committee (NRSC). Sound samples were stored on the hard-disk drives of PC's and presented to listeners individually at each station. The software collected and stored listener responses, requiring no experimenter control or interaction once the test session commenced. 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 good decisions.

3. Narrowing the field of bit-rate comparisons (Phase 1)

3.1 Participants

Ten NPR employees participated in Phase 1. Listeners included 4 audio engineers and 6 additional staff members employed in various departments at NPR. By virtue of working at NPR, this listening population may be described as "well educated" in terms of sound quality, but would not necessarily be characterized as "golden ears". However, all of the audio engineers who participated work extensively with sound, and thus are likely to be more sensitive to very small changes in sound quality than the general public.

3.2 Design and procedures

Listeners were presented with a total of 98 bit-rate pairs (i.e., 2 samples, back-to-back), and were asked the following questions about each pair:

(a) Which sample had **better** audio quality, "A" or "B"?

(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"?
(c) Would you turn either sample "A" or "B" off?

The test was divided into 2 sections, with listeners answering 49 trials before receiving a 5-minute break. Listeners were encouraged to play the samples as many times as they needed to make these determinations. See Appendix 2 for the Experimenter's script. 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". Figure 3.2 shows the PC response display used for the A/B discrimination task.

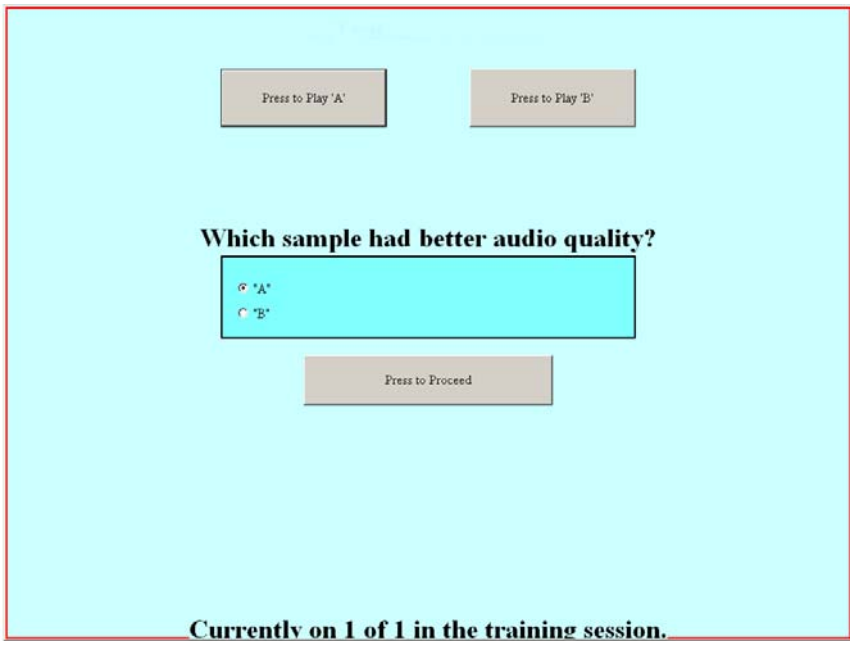

Figure 3.2: PC response display for A/B discrimination task

Table 3.2 shows the sample pairs used for this test. Notice that at each bit-rate sample-pairs that were quite close (8 and 16 kbps difference) were included. At points of special interest, pairs that were further apart (36 vs. 64, 48 vs. 72, and 64 vs. 96) were included.

At each bit-rate, samples included: (a) male speech; (b) female speech; (c) classical music; (d) jazz; (e) rock; (f) male voice-over; (g) female voice-over.

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	

Table 3.2: Sample pairs used in Phase 1 testing

3.3 Results for Phase 1 Testing

3.3.1 Accuracy

Table 3.3.1 shows total results for discrimination 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 (see 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. Notice that although a majority of NPR listeners were able to reliably tell the difference between 64 and 80 kbps, the percentage of correct responses was closer to chance, as evidenced by the lower percentage, the lower t- and p value.

Bit rates	**Percentage of respondents claiming higher bit-rate sounded better**	**t-test, probability level**
24 vs. 36	77%	t = 5.3693; p = .0001
24 vs. 48	76%	t = 4.9812; p = .0001
36 vs. 48	71%	t = 4.2696; p = .0001
36 vs. 56	67%	t = 3.3230; p = .0001
36 vs. 64	68%	t = 3.4320; p = .0001
48 vs. 56	50%	not significant
48 vs. 64	46%	not significant
48 vs. 72	54%	not significant
56 vs. 64	46%	not significant
56 vs. 72	54%	not significant
64 vs. 80	61%	t = 2.2103; p = .01
64 vs. 96	54%	not significant
72 vs. 80	59%	not significant
72 vs. 96	54%	not significant

Table 3.3.1: Results of A/B discrimination testing

3.3.2 Size of difference

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. Table 3.3.2 shows these results for participants who had correctly

identified the higher bit-rate sample[2]. Notice that participants claimed small differences at the higher bit rates, indicating that although they heard a difference between the two samples, the perceived quality difference was minimal.

Bit rate	Size of difference
24 vs. 36	4.17
24 vs. 48	5.06
36 vs. 48	3.11
36 vs. 56	3.48
36 vs. 64	3.53
48 vs. 56	3.09
48 vs. 64	2.51
48 vs. 72	3.64
56 vs. 64	2.70
56 vs. 72	3.02
64 vs. 80	2.61
64 vs. 96	2.49
72 vs. 80	2.20
72 vs. 96	2.82

Table 3.3.2: Size of difference when quality was identified correctly

3.3.3 Listener Behavior

Finally, listeners were asked whether they would continue to listen to sample "A", sample "B", "neither" or "both" at various bit-rates. Table 3.3.3 shows the rate of discontinue listening for each bit-rate. Notice that over 40% of participants reacted negatively to samples coded at 24 and 36 kbps, but this number dropped substantially to 15% at 48kbps. This result indicated that somewhere around 48kbps the great majority of listeners began to react favorably to HDC.

Bit-rate	Discontinue
24	43%
36	43%
48	15%
56	17%
64	17%
72	21%
80	11%
96	17%

Table 3.3.3: Percentage of participants who would discontinue listening

Based on listening results of NPR personnel, 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

[2] Correct identification means participants judged "higher" bit-rates as having "better" quality. While it may be argued that this is not necessarily an appropriate assumption when different coders are being judged against each other, we assume that because only the HD coder was tested as the bit rate increased the quality improved.

reasonably large (i.e., 24 x 36; 24 x 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 available in the Main Audio Program stream (MAP) (i.e., 48 vs. 64; 48 vs. 96). Finally, we included 64 vs. 96 to replicate test conditions in previous NRSC FM testing. Because Phase 1 indicated that NPR listeners could not reliably discern differences between 48 and 56kbps, and 48 and 72kbps, we did not include those comparisons in general consumer testing. However, we did include a 48 vs.96kpbs bit-rate pair to see whether consumers could hear a difference between the two more disparate bit-rates.

4. Phase 2 Consumer testing

4.1 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. Four males and 5 females were excluded from final results because they failed the screening test. Seven participants were excluded because they did not complete the test. Three more female participants were excluded in order to make even the number of responses from each gender. Table 3.1 shows the demographic breakdown of general public 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.

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

Table 4.1: Demographic breakdown of participants included in results

4.2 Design and Procedures

General consumer testing was conducted between July 19th and August 3rd. Participants were tested individually over Sennheiser HD-600 headphones for approximately 2 ¼ hours. The test session was divided as follows:

1. Experimenter welcomed participants and described the equipment and test procedures
2. Participants were given a screening test, followed by a short break
3. Participants were given an ACR-MOS test, followed by another short break
4. Participants were given an A/B pair-wise comparison test
5. Participants were de-briefed, paid and escorted out

Notice that in this study participants rated the same samples in two ways: (a) they completed an ACR-MOS test and (b) they completed an A/B comparison test on selected sample-pairs. Why use

both methodologies? The ACR opinion scores derived from a single stimulus presentation test tend to be highly predictive of real-world consumer satisfaction. Listeners are rating samples one at a time, using their internal reference to guide their decisions. This is how most consumers judge audio on an everyday basis. However, it has been argued that the ACR-MOS is not as sensitive to differences as other kinds of testing, such as directly comparing one audio sample to another in an A/B presentation. Therefore, in order to test stringently and thoroughly, both the ACR-MOS and A/B comparison test methodologies were included in this study. See Appendix 2 for the Experimenter Script.

4.2.1 Screening Test

Screening was conducted to ensure that listeners were reliably able to distinguish between significantly different audio qualities. There were seven screening trials. For each trial, participants were asked to listen to three samples, two of which were the same and the third different (for example, two female speech source samples and the same female speech sample processed through an AM receiver; two rock source samples from a CD and the third sample coded at HDC 24 kbps). 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 **six** of **seven** screening triads correctly. Listeners were provided no feedback on the "correctness" of their responses during the screening test nor were they informed of their specific performance after they were finished. Playback of samples was under listeners' control, but the screening software required them to listen to all three samples, from beginning to end, before the response options became available. Figure 4.2.1 shows the PC response display that was used for the screening task.

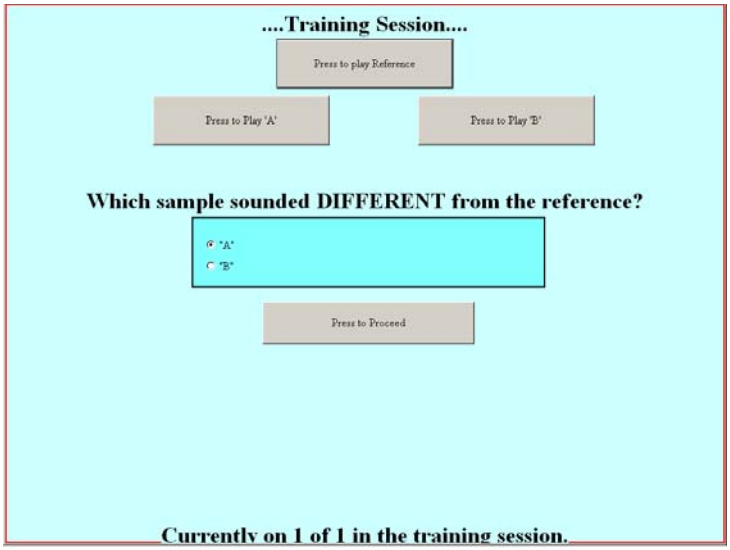

Figure 4.2.1: PC response display for screening test

4.2.2 Single stimulus, absolute category rating (ACR) test

In the ACR test, participants listened to 200 samples, one-by-one, and rated each sample individually. The test was divided into several sub-sections, with participants answering 67 trials and receiving five-minute breaks, until all trials were finished. 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. Figure 4.2.2 lists single stimulus samples used in the ACR test.

	16	**24**	**36**	**48**	**56**	**64**	**72**	**80**	**96**	**CD Source**	**Total**
Speech (2 male; 2 female)	4	4	4	4	4	4	4	4	4	4	40
Classical	4	4	4	4	4	4	4	4	4	4	40
Jazz	4	4	4	4	4	4	4	4	4	4	40
Rock	4	4	4	4	4	4	4	4	4	4	40
VoiceOver	4	4	4	4	4	4	4	4	4	4	40
Total	**20**	**20**	**20**	**20**	**20**	**20**	**20**	**20**	**20**	**20**	**200**

Figure 4.2.2 Samples used in ACR test

4.2.3 Double stimulus, A/B test

In the double stimulus test, participants were given 30 sample-pairs and asked the same three questions that Phase 1 listeners were asked:

(a) Which sample had **better** audio quality, "A" or "B"?
(b) How big was the difference, on a scale of 1-10, with 10 being "extremely different", and 1 being "I really couldn't tell a difference but you made me pick"?
(c) Would you discontinue listening to sample "A" or "B", neither or both?

Table 4.2.3 lists the sample pairs participants were asked to rate.

	24 vs. 36	**24 vs. 48**	**36 vs. 64**	**48 vs. 64**	**48 vs. 96**	**64 vs. 96**	**Total**
Speech (1 male; 1 female)	2	2	2	2	2	2	12
Classical	1	1	1	1	1	1	6
Rock	1	1	1	1	1	1	6
Jazz	1	1	1	1	1	1	6
Total	**5**	**5**	**5**	**5**	**5**	**5**	**30**

Table 4.2.3: Samples used in A/B test

5. Consumer Test Results

5.1 Preliminary analyses

Preliminary analyses were conducted to examine whether participants rated audio quality of samples differently based on their age or gender. A 2 (gender) x 4(age) ANOVA yielded a main effect of age, but no main effect of gender. Newman-Keuls Multiple-Comparison tests (p=.05) 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 quite minimal, participants' data was combined for all other analyses and total results are reported.

5.2 Absolute Quality Rating

Table 5.2.1 shows the ACR-MOS for bit rates from 16 kbps to 96 kbps, as well as the CD source reference samples. The results are listed by genres (for a complete listing of MOS by cut, see Appendix 3). 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. In classical and jazz, 16 and 24 kbps were rated significantly lower than all other bit rates and the reference. In Rock, 16, 24 and 36kbps were rated significantly lower than all other bit-rates. In Voiceover, 16, 24 and 36kbps were rated significantly lower than all higher bit-rates. In Speech, 16, 24, 36, 48 and 64kbps were all rated statistically lower than the reference. However, while 16, 24 and 36 were rated significantly lower than 96kpbs, 48kpbs and 96kpbs were rated equivalently. In order to examine the speech genre more closely, it was divided into male and female speech. Table 5.2.2 shows slight differences between participants' scores for female and male speech. For female speech, 48, 56 and 64kbps were rated significantly different from the reference (but not from 96, 80, 72), whereas with male speech 48kbps was rated significantly the same as all of the higher bit rates.

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

Table 5.2.1: Mean opinion scores for genres

	16	24	36	48	56	64	72	80	96	CD Source Reference
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

Table 5.2.2: Mean opinion scores for female and male speech

Taken together, these results suggest that in general there is a difference in people's perception of quality at lower bit rates than at higher bit rates, and that this difference emerges between 36 and 48kbps. With the exception of female speech, participants reported quality parity until 36kbps. At 36kbps, 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 24kbps, participants rated most genres as "fair", and at 16kbps participants rated samples between "poor" (2.0) and "fair" (3.3).

5.3 A/B Test Comparisons

5.3.1 Accuracy

Table 5.3.1.1 shows results for which sample had better audio quality. As with Phase 1 participants, 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%. Again, in keeping with Phase 1 participants, general public listeners were able to correctly identify the higher bit rate of the bit-rate pair at very low bit-rates. The majority of participants heard differences between 24 and 36kbps; 24 and 48kbps; and 36 and 64kpbs. The majority did not hear differences at 48 vs. 64kbps, but a slight majority accurately reported hearing differences between 48 and 96kbps and 64 and 96kpbs. The t and p values indicate, however, that while significantly different from chance, the percentage of people accurately reporting differences was minimal.

Bit rates	**Percentage of respondents claiming higher bit-rate sounded better**	**t-test, probability level**
24 vs. 36	77%	t = 9.2935; p = .0001
24 vs. 48	80%	t = 10.8652; p = .0001
36 vs. 64	64%	t = 4.1145; p = .0001
48 vs. 64	54%	not significant
48 vs. 96	56%	t = 1.8491; p = .03
64 vs. 96	57%	t = 1.9932; p = .02

Table 5.3.1.1: Results from A/B discrimination testing

In order to explore specifically where participants were hearing differences, t-tests were run for each genre at 48 vs. 96 and 64 vs. 96kbps. Table 5.3.1.2 shows these results. T-tests showed significance in "speech" at 64 vs. 96kpbs, but not at 48 vs. 96kpbs. The inability to find significant differences by genre at 48 kbps is most likely an artifact of statistical testing: the smaller the number of responses, the larger the difference must be for statistical significance. Because the number of responses in the genre analyses was substantially smaller than the number included in analyses conducted for total responses, statistical differences did not show up. However, if results from 48 vs. 96 and 64 vs. 96 are taken together, there is a strong indication that more participants heard differences in "rock" and "speech" than they did in "jazz" and "classical".

Bit rate 48 vs. 96	Percentage of respondents claiming higher bit-rate sounded better	t-test, probability level
Jazz (n = 40)	55%	not significant
Rock (n = 40)	60%	not significant
Speech (n = 80)	58%	not significant
Classical (n = 40)	53%	not significant
Bit rate 64 vs. 96		
Jazz (n = 40)	55%	not significant
Rock (n = 40)	55%	not significant
Speech (n = 80)	64%	t = 2.5423; p = .0001
Classical (n = 40)	48%	not significant

Table 5.3.1.2: Results from discrimination testing at 48 and 64 kbps by genre

5.3.2 Size of difference

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; 10 being extremely different and noticeable. Table 5.3.2 shows these results for participants who correctly identified the higher bit-rate sample. Note 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.

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 5.3.2: Size of difference when quality was identified correctly

5.3.3 Listening Behavior

Finally, listeners were asked whether they would continue to listen to sample "A", sample "B", "neither" or "both" at various bit-rates. Figure 5.3.3 shows the **difference** between the turn off rate for 96kbps and other bit rates[3]. In this figure, 96kbps was set to "0". The difference then is the

[3] 17% of Phase 1 and 16% of Phase 2 participants claimed that they would discontinue listening to samples coded at 96kbps. However, the mean opinion scores for 96kbps were between 3.5 and 4.2, and thus we believe that this inflated "discontinue" rate reflects participants' feelings about the source material, not the quality of the sound through the coder. Further, during this task we did not give participants explicit instructions to confine their judgment to audio quality. Because of these factors, we use 96kbps as our benchmark, set it to "0" and report on the difference between participants' rating of 96kbps and other bit rates.

additional rate of discontinuation participants claimed at various bit rates. Notice that fewer general public listeners reacted negatively to samples coded at 24 and 36 kbps than did NPR participants, but at 48, 64 and 96kbps, the numbers are virtually the same. These results again indicate that between 36 and 48kbps participants' behavior changes, with a large majority contending that they would maintain listening at 48kbps.

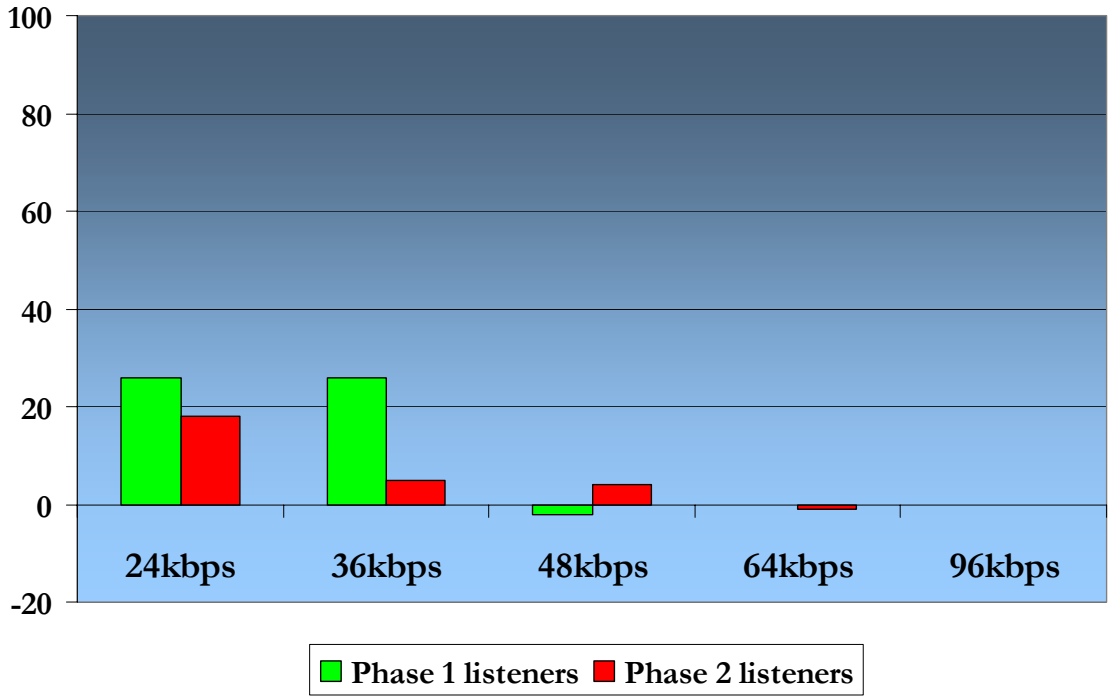

Figure 5.3.3: % difference between 96kbps and lower bit rates

6.0 Conclusions

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 48kbps or higher, or notice very small differences. As noted, these differences were heightened by allowing each participant to audition the choices as many times as needed to make a comparative decision, an opportunity obviously unavailable to radio consumers. However, participants do notice significant differences at lower bit-rates of 16, 24, and 36kbps. As with previous testing, participants were more sensitive to differences when rating speech than when rating music and voiceovers. This is presumably due to reduced psychoacoustic masking opportunities (i.e., there is less masking of digital artifacts associated with speech's overall lower acoustic density and frequent wave front pauses) or because humans are particularly sensitive to voices and voice quality -- or some combination of these factors.

Results from this study clearly indicate that for the HDC coder it is possible to separate 96kbps into two 48kbps streams with minimal, if any disturbance to listeners. Interestingly, when making choices about bit allocation for HDC, it is apparent that music may require fewer bits than speech to maintain transparency.

Appendix 1: Samples Used in Testing

Artist	Album	Song	ASIN/ISBN
Jacques Ibert	Summertime Music for Oboe	Entre'acte	B000000A9T
Georges Bizet	Carmen		B0000007DT
The Cars	The Cars	Just What I Needed	B00000IL66
Eric Clapton	The Best of Eric Clapton	Change the World	B00001U03Q
Tuatara	Cinemathique	Falling Pianos	B00005UWMB
Strunz & Farah	Primal Magic	Bola	B00001X53X
Male speech Philip Pullman	The Spoken Word (Children's Writers)	I was a Rat	ISBN 0712305181
Male speech John Glenn	A Memoir		NRSC Cut
Female speech Jacqueline Wilson	The Spoken Word (Children's Writers)	Wilson's Double Act	ISBN 0712305181
Katie Burton	The Vendetta Defense		NRSC Cut
Female Voiceover	provided by Sun Sounds of Arizona		
Male Voiceover	provided by Sun Sounds of Arizona		

Appendix 2: Experimenter Script

Welcome to our session! Today, you will be participating in a listening experiment, which should last about 2 hours. You will be listening to music and speech samples over headphones. There are three parts to this study. The first part is training, where you will listen to the music you will be encountering in your tests. The second part is a discrimination test, which we will explain in a few minutes, and the third part is an opinion test, again explained later.

In the training session, you will hear all of the sound samples we are going to be using today. We want to familiarize you with the material, so you will know what to expect during the test. You will be hearing the same material several times during the test, so don't be surprised when you come to a duplicate.

Screening Test

Now we are ready to examine your ability to hear different impairments. Let me first explain the task you will be doing, and then together we will try a practice trial. *(Experimenter: Show them Attachment 1, the REF-A-B task*). In this task, your only job is to decide which sample (either A or B) is DIFFERENT from the reference. Once you have decided, you will be prompted by the program to register your response. You are welcomed to play samples as many times as necessary to make your decisions. The differences between samples will sometimes be quite small, so you have to listen very closely. One of the samples *will always be identical to the reference*. The other *will always be different from the reference*.

In this part, there are 8 trials. The software will pretty much guide you through the trials. One important thing to note, you *must* play each of the sound samples at least once through (from beginning to end) before you can register your answers. The software will not let you continue until you have heard all three samples. Once you have played the samples, you can then register your responses. Once you have registered a response, you cannot go back and change it. Therefore, be careful to put in the response you really intend.

I can change the listening level during practice trials, but you may not change your listening level during the test. So, make sure the level you are listening at is comfortable before you start! I will help you determine this level.

When you are all done, you should open the door to find your experimenter.

Any questions? OK, now let's practice.

Experimenter: Start the program named "Pre-Screening" Make sure of the following:

- *the subject knows how to use the software*
- *the subject is wearing the headphones correctly*
- *the listening level is comfortable for the subject – volume may only be changed during the training sessions.*
- *the subject understands the entire procedure.*

The Main Test – Part 1

In this part of the study, you are going to hear over samples. You are asked to give your opinion of the "overall quality" of the sound samples you hear. You will be presented with one sound sample at a time. Listen carefully to the sample, and then rate it on an Excellent to Bad scale. The categories you can choose from are: Excellent, Good, Fair, Bad and Poor. After 67 samples, you will be asked to take a break. It is important that you rest for at least 5 minutes between 67-sample groupings. Also, if you feel that you are "burning out" during the 67, stop and relax at your terminal. There is absolutely no rush – you don't win a prize for getting done first!

This test is different from the first test you took. There is no stated reference against which to compare the samples you are hearing. You simply hear a recording then rate it. You will have to use an internal reference to judge the "goodness" of the sample. By that I mean, when you are listening to a particular sample, think about how a radio show would sound in your car and over your home radio. Judge the sample in relation to your memory of those two references. Also, you will start to judge the sound samples against others you have heard during the test. This is ok, because your natural inclination will be to try to rate samples consistently. You may probably feel a little unsure of yourself on the first 3 or 4 trials. Don't worry! Just think about your internal reference, and you'll know how to rate those samples. After the first 3 or 4 samples, you'll feel like a pro!

Don't be afraid to use the entire scale to rate the samples. If you believe the sample sounded excellent, say so! If it sounded bad, again, say so.

Many things go into a quality rating. You'll be listening for impairments as well as the overall aesthetic value. By aesthetic we mean beauty, musicality, character, sound quality, etc. Try to judge each sample in an overall sense. This is especially hard to do if a big impairment happens to occur at the end of the sample. So, before you rate each sample, take a few seconds to think about the entire sample you just heard. In that way, it won't be just your last impression that carries the most weight.

Any questions? Great. Now, lets try a few trials to get used to the procedure and adjust the sound volume once again.

Experimenter: Start ACR-MOS test. During the training trials, allow the participant to set the volume to a comfortable listening level. Again, make sure of the following:

- *the subject knows how to use the software*
- *the subject is wearing the headphones correctly*
- *the listening level is comfortable for the subject*
- *the subject understands the entire procedure.*

Main Test – Part 2 (paired comparisons)

This part of the test will be pretty short. You will be played 2 samples, back-to-back and will be asked to simply report which one you liked better, and by how much. Often people are concerned because they think the samples are equal in quality. This is ok. We know that sometimes the differences may be very small, and that you may feel like you are "guessing". Other times you will be sure of yourself. Just remember, this is opinion test, and there are NO right or wrong answers! Again, the software will lead you through this procedure.

Appendix 3: ACR-MOS of short individual samples

Cut	16	24	36	48	56	64	72	80	96	Reference
BizetC1	2.6	3.1	3.8	3.8	3.6	4.0	3.9	3.8	3.9	3.8
BizetC2	2.4	3.0	4.0	4.0	4.1	4.0	4.2	4.1	4.2	4.2
CarsC1	3.0	3.3	3.7	4.0	3.9	4.0	4.1	4.1	4.2	4.1
CarsC2	2.0	2.7	3.0	3.4	3.7	3.6	3.6	3.7	3.7	3.9
ClaptonC1	2.5	3.0	4.2	4.1	4.0	4.2	4.3	4.3	4.3	4.5
ClaptonC2	2.5	3.3	3.9	4.0	4.0	4.1	4.2	4.2	4.2	4.2
FemaleA1	1.9	3.1	3.1	3.4	3.3	3.4	3.5	3.8	3.8	4.0
FemaleB1	1.7	2.5	3.4	3.6	3.9	3.8	3.9	4.0	3.9	4.1
FemaleC1	2.7	3.2	3.4	3.5	3.7	3.6	3.7	3.6	3.6	3.4
FemaleC2	2.5	3.1	2.9	3.3	3.4	3.5	3.3	3.1	3.3	3.0
IbertC1	3.1	3.5	4.2	4.2	4.2	4.1	4.3	4.1	4.4	4.2
IbertC2	3.1	3.2	4.2	4.2	4.2	4.1	4.1	4.2	4.1	4.4
MaleA1	1.7	2.5	3.4	3.7	3.9	3.7	3.8	3.9	3.9	4.0
MaleB1	2.6	3.4	3.9	4.2	4.2	4.1	4.2	4.2	4.3	4.3
MaleC1	2.3	3.0	3.2	3.1	3.5	3.5	3.6	3.5	3.6	3.5
MaleC2	2.2	2.6	3.1	3.4	3.5	3.3	3.4	3.2	3.5	3.5
StrunzC1	3.4	3.8	4.3	4.2	4.3	4.3	4.3	4.3	4.4	4.5
StrunzC2	3.3	3.4	3.9	3.8	4.0	4.0	4.2	4.4	4.1	4.2
TuataraC1	3.4	3.8	4.0	3.8	4.0	4.1	4.2	4.0	4.1	4.1
TuataraC2	3.2	3.7	4.0	4.0	4.3	4.1	4.1	3.9	4.2	4.2