Audio Engineering Society

Convention Paper

Presented at the 137th Convention
2014 October 9–12 Los Angeles, USA

The Correlation Between Distortion Audibility and Listener Preference in Headphones

Steve Temme[1], Sean E. Olive[2], Steve Tatarunis[3], Todd Welti[4], and Elisabeth McMullin[5]

[1,3] Listen, Inc., Boston, MA, 02118, USA
stemme@listeninc.com, statarunis@listeninc.com

[2,4,5] Harman International Industries, USA
sean.olive@harman.com, todd.welti@harman.com, Elisabeth.McMullin@harman.com

ABSTRACT

It is well-known that the frequency response of loudspeakers and headphones has a dramatic impact on sound quality and listener preference, but what role does distortion have on perceived sound quality? To answer this question, five popular headphones with varying degrees of distortion were selected and equalized to the same frequency response. Trained listeners compared them subjectively using music as the test signal, and the distortion of each headphone was measured objectively using a well-known commercial audio test system. The correlation between subjective listener preference and objective distortion measurement is discussed.

1. INTRODUCTION

There has been much research published on how a loudspeaker's *linear* performance, e.g. frequency, time and directional responses, affects perceived sound quality. However, there is little research published on how *non-linear* distortion affects perceived sound quality. In recent years, the increasing availability and affordability of high quality headphones and personal digital music players e.g. MP3 players, has brought high quality music playback to the masses. The transducer performance is critical to listener enjoyment and Dr. Olive and others have presented research on what they believe the target frequency response of the headphone should be for optimum sound quality [1]. The attempt of this research is to determine what level and what kind of distortion is audible and how it affects the perceived sound quality.

Five different pairs of good quality over-the-ear headphones with varying levels of distortion were objectively measured and subjectively rated for their perceived sound quality. First, each headphone was equalized to the same target frequency response. Several different kinds of distortion metrics including harmonic, intermodulation, and non-coherent distortion, were measured for each headphone. A listening test was then conducted where the five headphones were rated by eight trained listeners based on preference and distortion using four short musical excerpts. The program material was selected for wide dynamic and frequency ranges to excite mechanisms in the headphone transducers that would cause distortion.

The different headphones were presented virtually to listeners via binaural recordings of the headphones reproduced through a calibrated low-distortion reference headphone, Stax SR-009. This virtual headphone test method minimized headphone leakage effects, and removed the influence of non-auditory biases (brand, price, visual appearance, comfort, etc.) from listeners’ judgment of sound quality.

In this paper, correlations between subjective and objective ratings of distortion are examined (as was done previously [2]) in an attempt to develop an objective metric for measuring distortion audibility in headphones and other loudspeakers. This could possibly be extended to other types of audio devices such as amplifiers.

2. DISTORTION MEASUREMENTS

2.1. Selection of Headphones

Five popular over-the-ear headphones, retailing for $250 or more, that cover a wide range of technologies were selected for this study (see Table 1). More importantly, these headphones produce different amounts of measured nonlinear distortion making them good candidates for the purpose of this study.

Table 1 Details on the headphones selected for these experiments

Headphone	**Price on Amazon.com**	**Description**
AKG K701	$261	Dynamic, open back
Beats by Dre Studio Edition	$270	Dynamic, closed back with active noise cancellation
Bose QC15	$269	Dynamic, closed back with active noise cancellation
Sony MDR-V600	$350	Dynamic
Stax SR-009	$4055	Electrostatic

2.2. Measurement Setup

The headphones were placed on a Bruel & Kjaer Head and Torso simulator (HATS) Type 4128 which has two, left and right, ear simulators (Fig. 1). The calibrated pressure microphones inside the ear couplers are connected to a Listen SoundConnect 2 microphone power supply. The microphone power supply is connected to a RME Fireface UCX audio interface which converts the signal from analog into digital for the SoundCheck® measurement software to analyze the measured waveforms. The digital test signals, created and played by the SoundCheck software, are passed through the audio interface for conversion to analog signals, and then amplified by a Channel Island headphone amplifier. Four of the five headphones are plugged into the Channel Island headphone amplifier and the reference Stax SR-009 headphones have their own special amplifier, the SRM-007tII.

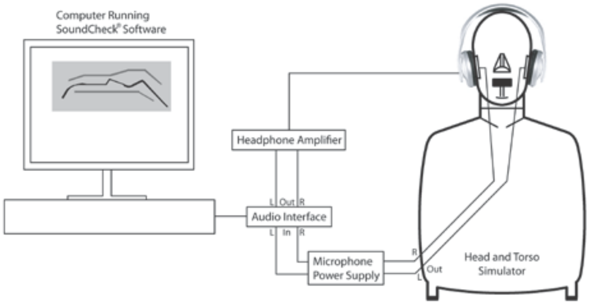


Figure 1 Experimental Hardware Setup

2.2.1. Headphone Sensitivity Calibration & Frequency Equalization

A custom SoundCheck calibration sequence was developed which:

1. Determines the combined sensitivity (in Pa/V) of the headphone under test and headphone amplifier.

2. Produces equalization and correction curves which, when applied to the stimulus (EQ) and analysis (correction), will result in a flat frequency response of the diffuse-field corrected HATS response.

During the sensitivity calibration process, SoundCheck applies a calibrated 1 kHz tone from the audio interface to the headphone amplifier/headphone under test. The tone is recorded by SoundCheck via the calibrated HATS ear signal path and the ensuing analysis derives the combined sensitivity (Pa/V) of the headphone amplifier and headphone under test. Knowing this sensitivity value, allows precise stimulus playback level control of the DUT at the HATS ear in dBSPL.

The Frequency Equalization process involves playing a 20 Hz – 20 kHz 1/12 octave resolution frequency stepped sweep at a user defined level (in our case 90 dB SPL) from SoundCheck and recording the output from the headphone at the HATS ear. The recorded time waveform is analyzed to yield a diffuse field corrected fundamental response and its reciprocal is then applied as equalization to the stimulus during a subsequent sweep. The recorded time waveform of the equalized sweep is then analyzed to yield a diffuse field corrected fundamental. At this point the test operator determines if the equalized curve is flat enough or if another pass at the equalization process is desired. In some cases, multiple iterations of the equalization process are required to achieve the target response.

The final output from this calibration sequence is an equalization curve which can be applied to the stimulus during playback and a correction curve (figure 2) which can be applied during analysis. The following example shows the two curves for one of the test headphones.

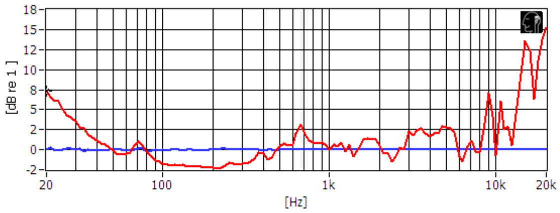


Figure 2 Equalization (red) & Correction (blue) Curves

As can be seen from the almost flat curve at 0dB in figure 2, the correction curve, after equalization, is typically within +/-0.5dB.

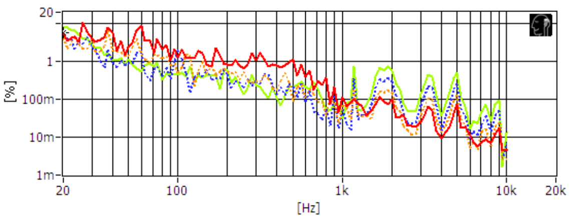
2.3. Test Signals for Measuring Distortion

Unlike linear measurements, non-linear distortion measurements depend heavily on the stimulus signal's spectral content and level [3]. For this reason, a variety of test signals were used: swept-stepped sine, two-tone, one fixed and one moving tone (IM), multitone, and even music as a test signal. After equalization, all five headphones were measured at four different calibrated levels; 82, 88, 94, & 100 dB SPL.

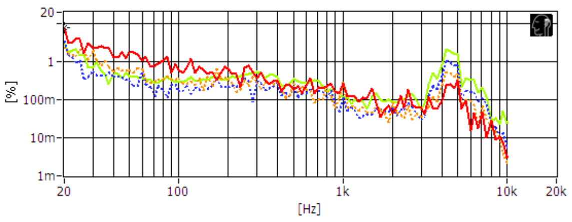
2.3.1. Harmonic Distortion

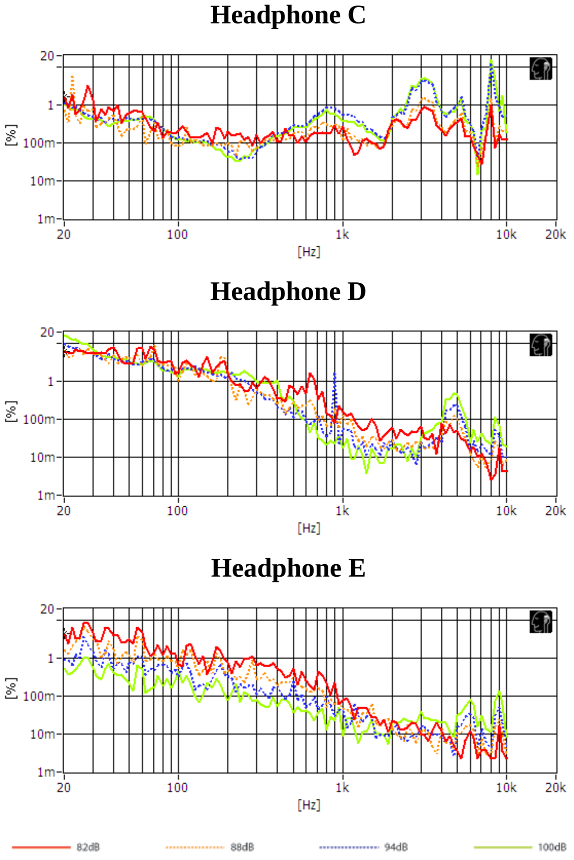
Traditional harmonic distortion measurements using a single test tone or sweeping tone are easy to calibrate and perform. They are also a good indicator of the type of non-linearity, e.g. asymmetric (even order harmonics), symmetric (odd order harmonics) and even rub & buzz (high order harmonics) [3]. Figure 3 shows the Total Harmonic Distortion (THD) measurements at the four different test levels.

Headphone A



Headphone B




Headphone C

Headphone D

Headphone E

Figure 3 Headphones A, B, C, D & E THD vs. Level

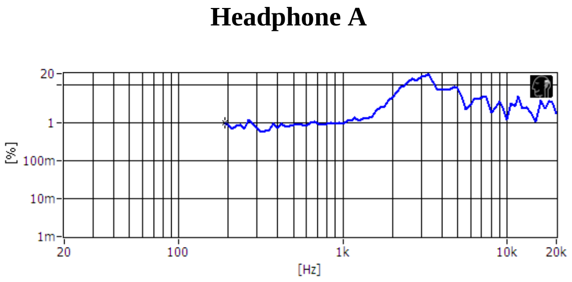
An interesting observation is that for most of the headphones, THD level decreases with increasing level at the low frequencies below 100Hz. This is not expected as most loudspeakers' distortion increases with test level as traditionally the voice coil starts to move out of the linear region of the magnetic gap. At first, it was suspected that this was an indication of poor signal to noise ratio at low test levels but the validity of the measurements was confirmed by repeating the test several times and confirming the measurement noise floor was well below the measured headphone distortion. It was also interesting that the distortion at higher frequencies increased with level as would typically be expected. This behavior has also been observed in microspeakers used for cell phones, tablet and laptop computers. Due to their thin profile, they typically do not have a spider to center the diaphragm or separate soft surround to attach the diaphragm to the basket. It is suspected that the driver's compliance decreases at low levels and this causes the added distortion. As the test level goes up, the diaphragm warms up and has to move more thus becoming more compliant.

In any case, since the human ear is less sensitive to distortion in this low frequency range, it is not expected to significantly affect the perceived distortion. Human hearing is more sensitive to distortion frequencies around the 1 – 5 kHz region [3]. Headphones B&C had noticeable distortion in this region.

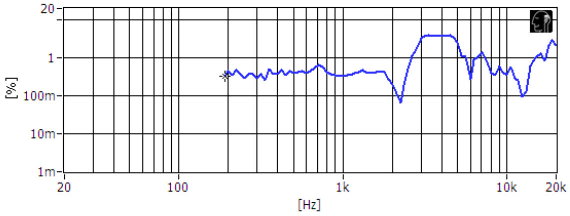
2.3.2. IM Distortion

An alternative to harmonic distortion (which does not reveal intermodulation products and is very different from music) is to use two test tones and measure intermodulation distortion. System theory shows that more than two tones are needed to fully characterize a non-linear system [4]. System nonlinearities cause intermodulation distortion (IMD) to occur due to amplitude and/or frequency modulation of the higher frequency components by the lower frequency components. The sum and difference components arising in two-tone interaction distortion have no harmonic musical relationships and can therefore be quite annoying. The difference components, in particular, are unlikely to be masked by the two test tones since they appear at lower frequencies, outside the effective masking curve region [5]. Since most headphones, including all the headphones in this sample group, use a single driver to cover the entire frequency range, it was expected to see some IM distortion at the higher frequencies where the high excursions of the driver at the low frequency tone is modulating the higher frequency tone.

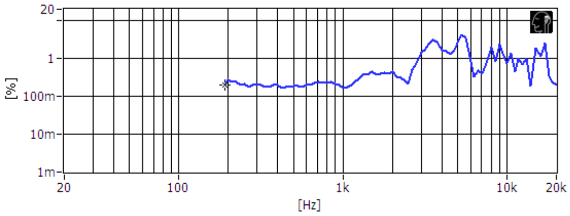
The test procedure fixed one tone at 43.1Hz @ 94dB SPL and swept the other tone from 190 to 20k Hz at 94dB SPL. In figure 5, Headphone A had the highest IM distortion of the five.


Headphone A

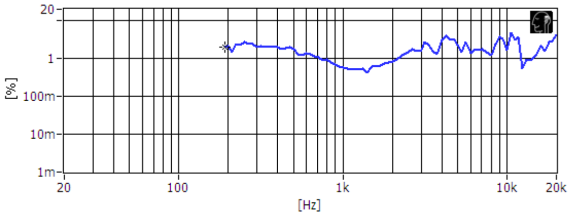
Headphone B



Headphone C



Headphone D



Headphone E

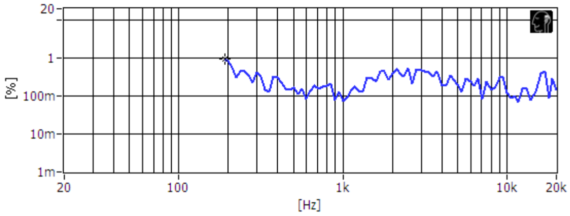


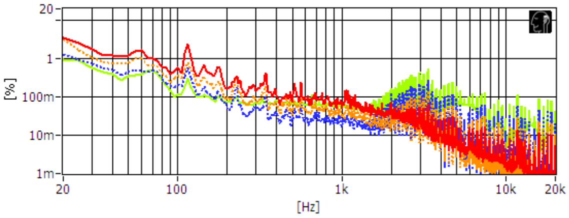
Figure 4 Headphones IMD at 94dB SPL

2.3.3. Multitone Distortion

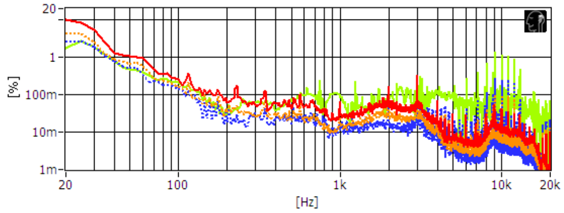
Multitone is a popular test signal for fast frequency response measurements. It is also a more rigorous test signal for assessing system non-linearities because it produces many frequency components simultaneously and excites both harmonic and intermodulation distortion products. Statistically, it is closer to music than a sine wave or two tones. [6,7, 8].

A 12th Octave multitone from 20 to 20kHz was created and measured at 82, 88, 94 & 100dB SPL. In figure 5, Headphones A, B, & C show more significant multitone distortion than the other two headphones, especially in the mid-frequencies.

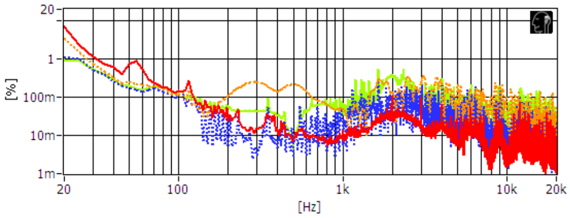
Headphone A



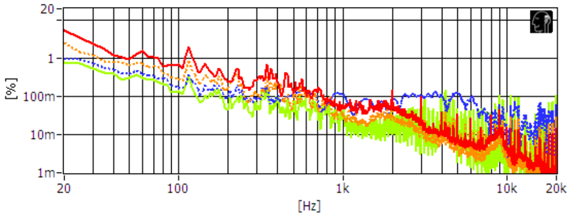
Headphone B



Headphone C



Headphone D



Headphone E

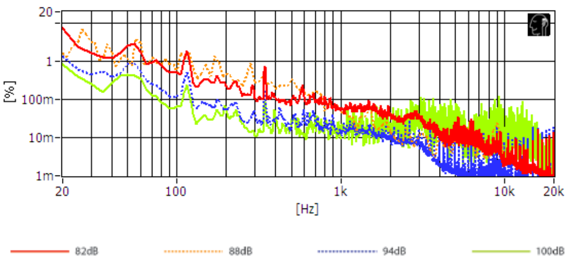


Figure 5 Headphones Multitone Distortion vs. Level

2.4. Non-Coherent Distortion using Music as the Test Signal

With dual channel analysis, non-coherent distortion can be measured using any stimulus signal including music [8]. It would be interesting to look at the different headphones' non-coherent distortion with the different music selections (Fig. 6). In order to get a good signal to noise ratio, the musical piece has to have a broad spectrum. Otherwise, we will be measuring noise instead of distortion. Most of the pieces had plenty of low to mid frequencies. Many did not have a lot of high frequency content above 10 kHz and in addition, the headphone target response rolls-off above 10 kHz. So it is best to ignore data above 10 kHz.

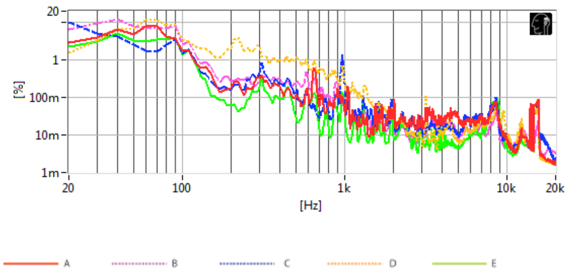


Figure 6 Headphones Non-Coherent Distortion measured using Music @ 80dB SPL average (B-weighted)

Headphone A is the red curve, B is pink, C is blue, D is orange and E is green. Headphone D clearly has the highest overall non-coherent distortion. Headphone C has a curious spike at 1 kHz.

3. LISTENING TESTS

This section describes the listening test method used for subjective evaluation of non-linear distortion in the different headphones

3.1. Binaural Recording and Playback of Headphones

The different headphones were subjectively evaluated by making binaural recordings of the headphones reproducing four music programs at different playback levels, and then reproducing the recordings through a calibrated reference headphone. In a previous paper [12], Olive et al. also used a virtual headphone listening test method where the replicator headphone was equalized to match the measured frequency responses of different headphones. The advantage of using binaural recordings is they capture both the linear and nonlinear distortions produced by the headphones. Since this paper is focused on only the *nonlinear* distortions, the frequency response of each headphone was equalized to a common target. In this way, we attempted to remove linear distortions so that any residual audible effects could be attributed to nonlinear distortions.

Another benefit of a virtual headphone test method is that it minimizes errors related to headphone leakage effects, and removes non-auditory biases (headphone brand, price, visual appearance, comfort, etc.) from listeners' sound quality assessments.

The procedure of making the binaural recordings and playing them back involved the following six steps:

1. Each headphone was measured on a G.R.A.S 45CA test fixture. This test fixture was equipped with a slightly modified ITU-T P57 Type 3.3 pinna, optimized to emulate leakage on human subjects. It also had an IEC 711 type coupler. A total of nine re-seats were measured and an average response taken.
2. The measured response for each headphone was used to create a correction filter, such that when applied to the headphone signal, would result in a headphone response matching a common target. This method is described in [1], and an example is shown in figure 7.
3. For each headphone, the source tracks were pre-equalized using the correction filter for that headphone. Additionally, overall gain differences between the headphones were measured and

accounted for. Playback of each headphone and its correction filter thus resulted in nearly identical acoustical output levels and spectral balance for the recordings.

4. The headphones were placed on the 45CA, in a quiet anechoic chamber, and all music tracks were recorded using a RME Fireface UC sound card, which has a low noise, low distortion headphone output, and a low noise microphone preamp for the 45CA.
5. All recorded tracks were then equalized to a flat response on the Stax SR-009 headphones. This was necessary because the headphone/pinna/ear canal response had already been included in the recording, and simply using the Stax headphone for playback would result in the inclusion of a second headphone/pinna/ear canal transfer function.
6. The binaural recordings were played back through the Stax SR-009 headphones and amplifier using a Windows PC connected to a Benchmark DAC2 audio interface.

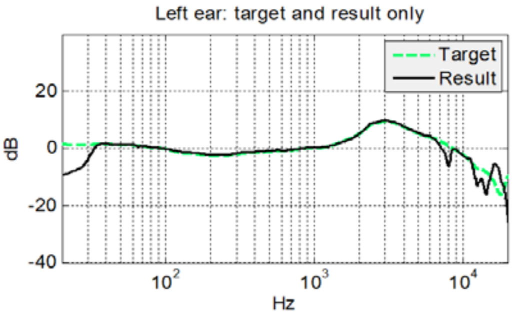


Figure 7 The frequency responses of the headphone target curve (dotted green), and an example of a headphone after equalization to the target (solid dark green).

3.2. Selection of Programs

A large number of commercial stereo programs were initially copied from Compact Disc and edited into short loops. The power spectrum and dynamic characteristics of the tracks were analyzed with custom MatLab scripts in order to narrow the selection to four music programs that would likely excite distortion mechanisms in the headphones. The programs generally had a large dynamic range and generous low frequency energy to excite nonlinear distortions in the headphones. At the same time, they had instruments, which produced spectra that would not overly mask the distortion components produced by the headphones. Details on the four programs used in the listening tests are summarized in Table 2.

Table 2 Details on the programs used in the listening tests

Program	*Artist / Album / Track / Year / Label / ASIN*
RP	Rebecca Pidgeon / The Raven / Spanish Harlem / 1994 / Chesky Records /B000003GGE.
JS	Joe Sample /Best of Joe Sample / Black and White / 1998/Warner Bros. Wea / B000009D1R.
MK	Mark Knopfler /Shangri-La / 5:15AM / 2004/ WEA-Reprise / B0002VKZL6.
SB	Shriekback / Sacred City / Signs / 1998 / World Domination / B000005LA6.

3.3. Selection of Listeners

The listeners were all employees of Harman International who were trained listeners with normal audiometric hearing. All listeners had successfully achieved level 7 or higher in the Harman How to Listen Training software [11].

3.4. Absolute and Relative Playback Levels

The relative playback levels of the headphones were adjusted to the same level based on ITU-R 1770.2 level. All of the listening tests were conducted at an average playback level of approximately 80 dB (B-weighted, slow) for both listening test sessions.

An important feature of this test is that the playback level is lower than the level which was used in the recordings. The 90 dB recording level would have been uncomfortable and possibly unsafe for the listening subjects. Absolute ecological validity is thus not achievable in this case. If anything, this should result in an increase in the sensitivity of the subjects to audible distortion.

3.5. Listening Test Procedure

Each listener completed a total of 8 trials (4 programs x 2 observations). In each trial, the five headphones were evaluated with the same program and rated on an 11-point interval preference scale. The scale had semantic descriptors at every second interval to encourage consistent interpretation and use of the scale: 1 (Strongly dislike), 3 (Dislike), 5 (OK - Neither Like Nor Dislike), 7 (Like), and 9 (Strongly Like). To encourage a consistent use of the scale, listeners were given the following guidelines in choosing how they separated their ratings between two headphones: a ≤ 0.5 point separation implied a slight preference, 1 point separation a moderate preference, and ≥ 2 points separation implied a strong preference. Listeners could optionally provide comments to describe the underlying reasons for their preferences.

Listeners were instructed to rate their preference, with emphasis on the distortion aspect of the stimulus. This was intended to prevent subjects from focusing on the (unavoidable) slight timbre differences between the stimuli.

The entire listening test was administered via a custom software application to ensure the test was repeatable and double blind. The software provided a graphical user interface for the listener to switch among the headphones and store their responses. The software also randomized the presentation order of the programs and treatments to eliminate any order-related biases.

Table 3 The ANOVA table for Preference Rating

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	7	812.993	116.142				
Headphone	4	237.880	59.470	9.698	<.0001	38.790	.999
Headphone * Subject	28	171.708	6.132				
Program	3	4.870	1.623	1.345	.2866	4.036	.298
Program * Subject	21	25.342	1.207				
Observation	1	.639	.639	.981	.3550	.981	.134
Observation * Subject	7	4.561	.652				
Headphone * Program	12	100.372	8.364	2.064	.0281	24.770	.905
Headphone * Program * Subject	84	340.384	4.052				
Headphone * Observation	4	6.911	1.728	.816	.5256	3.265	.223
Headphone * Observation * Subject	28	59.270	2.117				
Program * Observation	3	5.929	1.976	3.051	.0510	9.154	.626
Program * Observation * Subject	21	13.602	.648				
Headphone * Program * Observation	12	28.987	2.416	1.092	.3774	13.107	.581
Headphone * Program * Observation * Su...	84	185.776	2.212				

4. RESULTS

4.1. Statistical Analysis

The listener responses were statistically analyzed using a repeated measures analysis of variance (ANOVA) where the within-subjects fixed factors were Headphones (5 levels), Programs (4 levels) and Observations (2 levels). The dependent variable was preference rating. All statistical tests were performed at a significance level of 5%.

The ANOVA table (see Table 3) indicated the main effect was Headphone; F (4, 28) = 97, p <0.001. There was also a significant interaction effect between Program and Headphone. These effects are discussed in the following sections.

4.2. Headphone Effect on Preference

A Scheffé post-hoc test was performed to determine which paired comparison between the five headphones were statistically significant (see Table 4). The only significant differences in preference were between Headphone D (the least preferred headphone) and Headphones A, B and E.

Table 4 The Scheffé post-hoc test results for the effect of Headphone on preference rating

Scheffe for Preference Rating
Effect: Headphone
Significance Level: 5 %

	Mean Diff.	Crit. Diff.	P-Value	
A, B	.647	1.295	.6639	
A, C	1.145	1.295	.1143	
A, D	2.430	1.295	<.0001	S
A, E	.234	1.295	.9888	
B, C	.498	1.295	.8402	
B, D	1.783	1.295	.0014	S
B, E	-.412	1.295	.9135	
C, D	1.284	1.295	.0533	
C, E	-.911	1.295	.3163	
D, E	-2.195	1.295	<.0001	S

Fig. 8 shows the mean preference ratings and 95% confidence intervals for the five headphones averaged across all listeners and programs. The mean preference ratings from most preferred to least preference were: Headphone A (6.5), Headphone E (6.3,) Headphone B (5.8), Headphone C (5.4) and Headphone D (4.1).

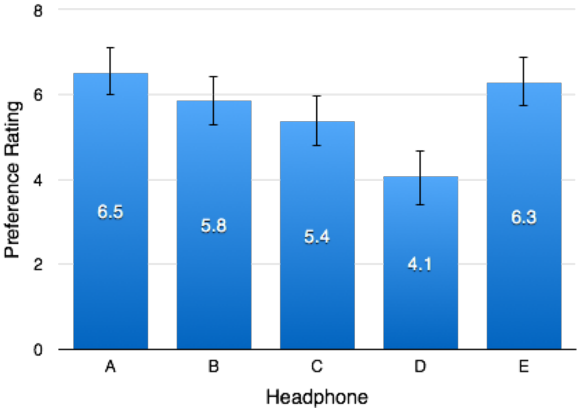


Figure 8 The mean preference ratings and 95% confidence interval for the five headphones averaged across all listeners and programs

4.3. Interaction Effect between Program and Headphone

There was a significant interaction effect between Program and Headphone (see section 4.1). This is graphically illustrated in Fig. 9, which reveals that the preference ratings for certain headphones varied depending on the music program. Note that the errors bars in the graph have been omitted in order to better see the trends. Headphone D tended to receive lower ratings for Programs SB and RP. Headphones B and C received much higher preference ratings on Program RP. The explanation for these program-headphone interactions are likely related to the mechanisms in the headphones that produce the distortion and the extent to which the physical characteristics of the program material can excite the distortions, yet not mask their audibility.

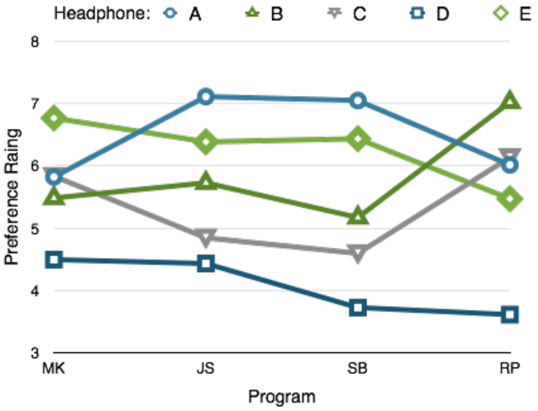


Figure 9 The mean preference ratings for each Headphone as a function of the program selection

4.4. Individual Listener Preferences

Fig. 10 shows the individual listener mean preference ratings and 95% confidence intervals for the different headphones. The first observation is the majority of listeners rated Headphone D as their least favorite headphone based on audible distortion. The second observation is that beyond Headphone D, most listeners had difficulty reliably discriminating among the headphones indicated by the small differences in headphone preference ratings, and the rather large overlapping error bars. Some listeners (53 and 363) apparently found it challenging to formulate preferences based on audible distortion. This is further confirmed in the following section where we examine the listener comments.

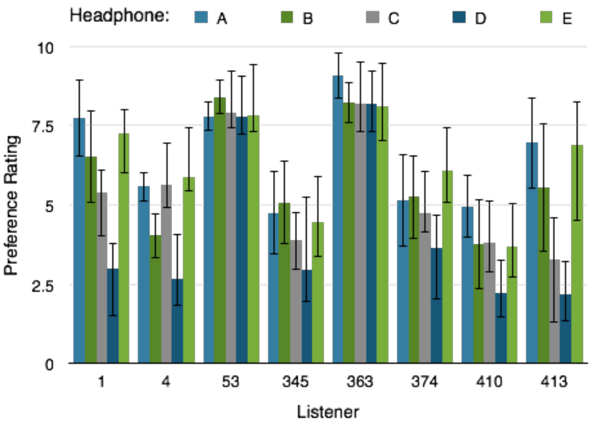


Figure 10 The individual listener mean preference ratings and 95% confidence intervals given to the headphones

4.5. Listener Comments

Listeners were encouraged to give comments to explain their underlying preferences. The headphone comments are summarized for each program in section 8. Each cell contains comments from both observations. Several observations can be drawn from reviewing the listener comments:

1. Other than Headphone D or to a lesser extent C, listeners generally had difficulty hearing audible nonlinear distortions. For Headphone D the effects were described as a noise or buzzing sound, muddy bass with less pitch, and some harshness at around 1-4 kHz. Many listeners contradicted themselves across observations saying that a particular headphone sounded distorted in one trial, but not distorted in another.

2. Most of the distortions were reported with the program RP (Rebecca Pidgeon's *Spanish Harlem*), and to a lesser extent SB (Shriekback's *Signs*).

3. Several listeners reported some small timbral or loudness differences among the headphone that could have been due to nonlinear distortions and/or small errors related to leakage and equalization when the headphones were recorded.

5. DISCUSSION

This section focuses on a discussion on the correlations between the objective and subjective measurements of nonlinear distortion in the headphones.

5.1. CORRELATIONS BETWEEN OBJECTIVE AND SUBJECTIVE MEASUREMENTS

The results from the listening tests presented in section 4 indicated that listeners had difficulty reliably discriminating among the different headphones even though the distortion measurements (section 2) indicated there were quantitative differences among them in terms of measured THD, IMD, multitone and non-coherent distortion measured with music.

The one exception was Headphone D, which listeners reported had audible distortion and was the least preferred. Headphone C received the second lowest preference ratings with some comments focused on its harsh, vocal sound reproduction. For this reason our discussion about correlation between subjective and objective measurements will focus mainly on Headphones D and C.

5.1.1. THD Measurements

The THD measurements (see Fig. 3) indicated that Headphone D had among the highest measured THD below 100 Hz regardless of the playback level. Listeners reported audible distortion on music programs with the highest amount of bass (Programs RP and DB). Listeners noted that one particularly loud bass note on RP produced audible distortion on Headphone D. Some listeners also reported that Headphone C sometimes produced overly bright, harsh, distorted vocal sounds. Headphone C had the highest amounts THD above 1 kHz.

In summary, there appears to be some moderate positive correlations between the amount of THD measured in the headphones and their sound quality rating.

5.1.2. IM Distortion Measurements

Looking at the IM distortion measurements of the headphones (Fig. 4), Headphone A had the highest amount above 1 kHz, yet listeners reported no audible distortion, and it received the highest preference ratings. The lowest rated Headphone D had the highest amount of IM distortion between 200 to 500 Hz but had less distortion than Headphone A above that frequency. Evidently, IM distortion was not a good indicator of its effect on sound quality in these particular tests.

5.1.3. Multitone Measurements

The multitone distortion measurements (see Fig. 4) indicated Headphones A, B and C had the highest amounts of distortion. Yet, all three headphones received higher preference ratings and less reported incidences of distortion than Headphone D. Unfortunately, in these tests multitone distortion was not a good predictor of audible distortion in the headphones.

5.1.4. Non-Coherent Distortion Based on Music

Figure 5 shows the non-coherent distortion measured for each headphone using the program RP, the program that produced the lowest preference ratings for the least preferred Headphone D (orange curve). Looking at figure 5, Headphone D has the highest percentage of measured distortion between 80 Hz and 1 kHz. The next lowest rated Headphone C (blue) has a large distortion spike at 1 kHz. The highest rated Headphone E (green) has the lowest amount of non-coherent distortion among the given headphones.

In summary, among the distortion metrics we chose in this study, non-coherent distortion based on music appears to be to be more correlated with listeners' preference ratings than the THD, IM and Multitone. It should not be surprising that this distortion metric produced the highest correlations since it used the same test signal (music) as used in the listening tests. We suspect that even better predictions of audible distortion in music could be achieved if a psychoacoustic model was applied to take into account the masking properties of the music on audibility of distortion. This will be the topic of a future study.

6. CONCLUSIONS

This paper investigated the relationship between the perception and measurement of nonlinear distortion in five popular over-the-ear headphones. The headphones were equalized to produce the same frequency response in order to focus listeners' attention on differences in nonlinear distortion. Four music programs were then binaurally reproduced at 90 dB SPL through each headphone and reproduced through a common low-distortion reference headphone at a lower, comfortable playback level. Eight trained listeners gave comparative preference ratings and comments for each headphone based on audible distortion.

Four different types of headphone distortion measurements (THD, IMD, Multitone and Non-coherent distortion based on music test signals) were made to explore which ones best correlate with the subjective results. The following conclusions can be made:

1. Headphone had a significant effect on preference ratings; this effect was largely isolated to one (Headphone D), which was less preferred to the other four headphones. There were no significant preferences among the other four headphones.

2. There were significant interaction effects between Programs and Headphones that influenced preference rating. Listener comments confirm that certain programs (SB and RP) produced more audible distortion on certain headphones than others.

3. Listener comments indicated Headphone D had audible distortion. For the other headphones, listeners' comments were often inconsistent across repeated observations, and expressed how difficult it was to hear audible distortion in the headphones.

4. While none of the headphone distortion measurements could reliably predict listener preference ratings based on audible distortion, the non-coherent distortion test based on the listening test music samples produced the best results. The least preferred headphone (Headphone D) had the greatest amount of THD below 100 Hz and greatest amount of non-coherent distortion between 100 Hz and 1 kHz. IMD and multitone distortion metrics were less successful.

5. Most of these headphones are considered to be high quality at least based on price and therefore

should have low distortion. A wider range of poorer quality headphones e.g. less expensive and more distorted should be evaluated to see if better correlation between subjective listening and objective measurements can be obtained.

In conclusion, the perception and measurement of nonlinear distortion is a fascinating but challenging area of research. The linear distortions in headphones are orders of magnitude higher and more audible than the nonlinear ones. The virtual headphone method used in this study solves one of the major research challenges: isolating the nonlinear distortions from linear distortions. Nonlinear distortion in headphones of this high caliber (Headphone D excepted) seems to not be a significant factor in how good it sounds. Finally, this study provides further experimental evidence that traditional nonlinear distortion measurements are not particularly useful at predicting how good or bad high caliber headphone sounds. A more perceptual-based approach for measuring the audibility of nonlinear distortion is warranted.

7. REFERENCES

[1] Olive Sean E., Welti Todd, and McMullin Elisabeth, "Listener Preferences for In-Room Loudspeaker and Headphone Target Responses," presented at the 135th Convention, Audio Eng. Soc., preprint no. 8867, (2013 Sept.).

[2] S. Temme, P. Brunet, and Parastoo Qarabaqi, "Measurement of Harmonic Distortion Audibility Using a Simplified Psychoacoustic Model," presented at the 133rd Audio Eng. Soc. Convention. Preprint 8704 (2012 Oct.).

[3] S. Temme, "Why and How to Measure Distortion in Electroacoustic Transducers", presented at the AES 11th Conference on Audio Test and Measurement, Portland, Oregon (29-31 May, 1992).

[4] M. Schetzen, "The Volterra & Wiener Theories of Nonlinear Systems", John Wiley, New York, 1980.

[5] S. Temme, P. Brunet, and D. B. (Don) Keele, "Practical Measurement of Loudspeaker Distortion Using a Simplified Auditory Perceptual Model," presented at the AES 127th Convention, preprint 7905, (2009 October).

[6] E. Czerwinsky, A. Voishvillo & Al., "Multitone Testing of Sound System Components- Some Results and Conclusions- Part 1: History and Theory", JAES, Vol. 49, No, 11, (2001 November).

[7] E. Czerwinsky, A. Voishvillo & Al., "Multitone Testing of Sound System Components- Some Results and Conclusions- Part 2: Modeling and Application", JAES, Vol. 49, No, 12, (2001 Dec.).

[8] S. Temme and P. Brunet, "A New Method for Measuring Distortion using a Multitone Stimulus and Non-Coherence," presented at the AES 121st Convention, preprint 6877, (2006 Oct.).

[9] S. Temme, "How to Graph Distortion Measurements" presented at the 94th Audio Eng., Soc. Convention, preprint 3602 (1993 March).

[10] A. Voishvillo, A. Terekhov, & Al., "Graphing, Interpretation, and Comparison of Results of Loudspeaker Nonlinear Distortion Measurements", JAES, Vol. 52, No. 4, (2004 April).

[11] Harman How to Listen: A Listener Training Software Program For Mac and Windows, www.harmanhowtolisten.blogspot.com (July 2014).

[12] Olive S.E., Welti T., and McMullin E., "A Virtual Headphone Listening Test Methodology," presented at the 51st Audio Eng. Soc. International Conference, Helsinki, Finland, (August 2013).

8. APPENDIX

The table below summarizes the listeners’ comments for each headphone and program. Each cell contains comments based on two observations. The cells highlighted in red indicate comments related to audible distortion. The cells highlighted in yellow indicate comments about timbre that could be indicative of distortion in the headphone.

Listener	Program	A	B	C	D	E
1	JS	No distortion; some HF emphasis; smooth; brighter than other top two	Bass bit unclear; no audible clipping or distortion	No Distortion; some 3-4 kHz emphasis	Brighter and thinner; timbre not as good; bass is fuzzy; less clear; some 2-3k emphasis; sharpness	No Distortion; bass a bit boomy
1	MK	No distortion; more proximity on voice; more bass than others	Less bass than others; big timbre difference; some 3-4k emphasis	No Distortion	Voice is colored; could be recording; distortion maybe heard as 3-k boost; nasal; bass is unclear	No Distortion
1	RP	I hear no distortion	Thin bass; no distortion	Thin bass; I hear mostly timbre differences	I hear some distortion on one bass note; third note of track	Good balance; no distortion; bass has pitch; a bit louder than others?
1	SB	Bass has pitch; no distortion; brighter than others	More bass; some HF emphasis; bit muddy' maybe LF distortion; bass lacks pitch; bass over boomy	Thinner bass; HF emphasis; harsh 3-4 kHz; bass not clear; distortion?	Some HF emphasis; has less bass; not timbrally neutral; bass not clear and lacks pitch	Low Distortion
4	RP				Can only really hear distortion n B	
4	SB		A lot of low bass in the right ear on this one			
4	SB				Obvious distortion	
53	JS	This is impossible to hear distortion differences	This is impossible		This is impossible	This is impossible
53	MK	Can’t hear any difference in distortion	Can’t hear any difference in distortion	I can't tell a difference	Can’t hear any difference in distortion	No Distortion
53	RP		Cleanest low end but also seems lacking in low end compared to the others, so I don’t know	Can’t hear any difference in distortion; cleanest sounding	Maybe some distortion on that highest bass notes? Tough; high bass note sounds funky	
53	SB		Low end sounds the best on the once, but the lyrics are still awful	Sounds fine		

345	MK	Most audible distortion; sounds slightly overdriven	Clear; no audible distortion; a little sharp	Mild vocal harshness; some low buzzing	Some audible distortion; vocal sounds unnatural; sharp	Some mild vocal harshness
345	RP	Sharp; clear	Clear	Some vocal distortion	A lot of bass distortion	Vocal harshness; some distortion on the bass
345	SB	Clear	Bass heavy; harsh vocals	Harsh; vocal distortion	Sharp; bass sounds fuzzy	Very neutral
363	JS	Good	Piano keys blend...not as distinct; Rough	Jingling less realistic	Sounds generally rougher; realistic/accurate	Background bass tones; good
363	MK	Good	Rough; PL	Good	Good	Good
363	RP	Rough	Sibilants; OK	Can hear in sibilants; less rough	Can hear distortion in low end; more rough	Distortion a little more obvious; rough
363	SB	OK; Good	A little rough; Good	Sounds compressed; Detail	Good; more HF, but don't hear more distortion necessarily	OK; Compression
374	JS	Bass seems a bit distorted, kind of smeared; Piano sounds best, most realistic on this one. I don't hear any obvious distortion.	Sounds clearest to me, hi hat sounds good; Bass sounds distorted.	Instruments not as easy to pick out, can't hear details as well as D and E; Piano sounds smeared, kind of distorted in a muddy way.	Seems to be some background hiss, bass sounds distorted; Piano sounds smeared, kind of distorted in a muddy way.	Similar to D, don't really notice any distortion; Similar to E, but piano not quite as clean sounding.
374	MK	I don't notice any distortion; Sounds distorted in the voice and bass.	Voice sounds compressed and grainy; Voice sounds kind of compressed, and fattened with distortion.	I don't notice any distortion; Voice sounds rough, distorted.	Voice sounds a little raspy; Voice sounds raspy, hard to tell if it's distortion or just brighter than the others.	Distorted guitars make it hard to pick out unwanted distortion; Sounds pretty clean
374	RP	Bass sounds distorted, and echo/delay on voice sounds distorted.	Seems like the clearest vocal, B and D sound more smeared on the lower mid; No obvious distortion	Voice sounds distorted, fattened up with distortion; voice sounds more sibilant	Bass sounds distorted on this one. Also, bass sounds compressed on all of them, as if the note gets louder after the string is plucked.	Bass sounds cleaner than A, voice sounds less raspy.
374	SB	Seems to be the cleanest; Bass sounds distorted, but not as badly as D.	Bass sounds boomy, but the fundamental seems to be there so maybe not as distorted as E.	Hard to tell with these sounds, but this one seems almost as clean as C; Percussion seems to have some distortion on the high frequencies.	Bass sounds very distorted; bass sounds very distorted	Bass sounds distorted, but not as bad as Possibly the least distorted.
410	JS	Feels like a little too much reverb	Quite distorted; bass Quite aggressive	Higher registers bass a little bit synth-like (also in C; a lot of reverb	Bass sounds more synth-like, more distorted	Bass not very natural; voice more clear because of less bass?

410	MK	Voice a little affected by a little intermodulation distortion, bass relatively clear	Guitar and bass a bit distorted	Low frequencies very distorted	The pronunciation of the mid also emphasizes the distortion; higher bass notes are not natural	Bass sounds a little rough, voice infected by some distortion
410	RP	It feels like the reverb creates some very low frequency amplitude modulation; voice a little affected by a little intermodulation distortion, bass relatively clear	Most natural; guitar and bass a bit distorted	Bass a little dull; low frequencies very distorted	Synth bass; the pronunciation of the mid also emphasizes the distortion	Bass a bit distorted; the pronunciation of the mid also emphasizes the distortion
410	SB	Very reverberant, but you can still hear the distortion	Very strong bass, voice a little distorted	A little bit more clean than C; most natural sound	More distorted bass; some bass notes sound a little bit like a synth	Bass sounds more clean (don't know if this is intended); he reverb is quite irritating, I think it adds quite some dist.
413	JS	Best; least coloration	Bright	Strange sounding; Phasey highs	Harsh, distorted	Clean, focused
413	MK	Colored mids	Clean, Thin	Sounds phasey; sounds compressed	Harsh, bright	Least coloration
413	RP	Wider sound stage, some Sibilance; brighter, phasey	Clean	Clean	Sounds more spacious, added ambience, phase differences? sibilant	Sibilant; Harsh
413	SB	Phase issues, bass heavy; best, bright	Sounds louder, Bass heavier, phasey; sibilant and harsh	Sounds like data compression; phasey, boomy bass	Sibilant, harsh; highs are distorted, sounds like compression	Least artifacts; highs are less clear