Audio Engineering Society

Convention Paper 10263

Presented at the 160th Convention

2026 May 28-30, Copenhagen, Denmark

*This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.*

# The Perception and Measurement of Nonlinear Distortion in Headphones

Sean E. Olive[1] and Pierre Lelièvre[2]

[1] *Sean Olive Audio Consulting, Oak Park, CA, USA.*

[2] *Rtings, Montreal, PQ, Canada.*

Correspondence should be addressed to Sean Olive (sean.e.olive@gmail.com)

## ABSTRACT

Few studies exist on the perception and measurement of nonlinear distortion in headphones. This paper reports the detection thresholds and perceived sound quality from real distortion in headphones. Five different distortion measurements were made on the headphones to determine how well they predict audibility and quality. Music samples were binaurally recorded on six headphones at playback levels ranging from 85 to +110 dBA at 3 dB increments. The recordings were reproduced at a normal playback level (83 dBA) through a reference headphone with low distortion. The headphone recordings were post-processed to remove both level and frequency response differences so only nonlinear distortions and residual noise remained. In a second test, listeners rated the similarity in quality of headphones relative to an undistorted reference and a hidden version of it. The results provide evidence audible distortion in headphones with music occurs at significantly higher playback levels (104 to 112 dBA SPL) than what is considered typical and safe. The percentage of measured THD in the headphone had the highest correlation with the detection thresholds while the non-coherent distortion with music best predicted the similarity ratings. We discuss the results and the practical implications they might have on future headphone design, testing and measurement.

## 1 Introduction

Few perceptual studies exist on the perception and measurement of nonlinear distortion in headphones [1],[2]. So far, the consensus among researchers is that headphone distortion occurs mostly in the bass (below 100 Hz) and is not an issue at normal playback levels, except for cheaper, poorly designed models [2]. Linear distortions related to the frequency response are the main determinant in how good the headphone sounds [3],[4].

While nonlinear distortion may not be the dominant factor in how good a headphone sounds, it is useful to know the levels at which it becomes audible and affects its sound quality, and which distortion measurement best predicts these perceptions. Total harmonic distortion (THD) is the most common metric reported in loudspeaker and headphone measurements, yet it has been reported to be poor at predicting audibility due to the simplicity of the test signal compared to music and it doesn't account for human auditory masking [4]-[6]. Establishing distortion thresholds and perceptually meaningful measurements would provide useful guidance in design, testing and purchase of headphones.

To this end, we conducted a study to measure the detection thresholds of nonlinear distortion produced in six headphones at SPLs between 85 to +110 dBA in 3 dB increments. Binaural recordings of the headphones were post-processed to remove level and frequency/phase differences so that only the nonlinear distortion artifacts and any residual noise remained. Listeners also rated the similarity in sound quality of the headphones relative to the undistorted reference. Five different distortion measurements were made at each SPL so that correlations could be calculated between the measurement and perception of the distortion.

## 2 Previous Work

Two previous studies specifically focused on the perception of nonlinear distortion in headphones. Temme et al. [1] binaurally recorded music samples through five headphones at 90 dB SPL and had listeners rate them based on preference. The binaural recordings and measurements were made at different playback levels ranging from 82 to 100 dB SPL. The 94 dB recording was reproduced through a high quality-low distortion reference headphone at normal playback levels, 80 dB (B-weighted, slow). The headphones were equalized to the same target curve so that mostly nonlinear differences remained. Trained listeners could not reliably discriminate among the headphones in a preference rating task with the exception of one model that had notably higher harmonic distortion below 100 Hz and higher non-coherent distortion measured with music than the other models The other headphones had higher amounts intermodulation (IM) and multitone distortion than the least preferred headphone providing evidence that these metrics are less predictive of audibility and effect on sound quality.

Gutierrez-Parera and Lopez [2] used a diagonal Volterra model to simulate the nonlinear distortion for nine different headphones at six playback level from 70 to 100 dBA. The simulated headphones were reproduced through a reference headphone at a fixed playback level of 85 dBA using 10 expert listeners and one music sample. The recording levels at which the distortion was detected was measured using a two-alternative-forced choice (2AFC) ABX test. The price of the headphone was positively correlated with the SPL it could produce before distortion was audible. Headphones costing <$20 could only reach 80 dBA, models costing $20-$199 reached 83 to 89 dBA. Two models costing >$200 were able to produce at least 100 dBA or more before audible distortion occurred. They did not provide different distortion measurements for correlation analysis except increased levels of second and third harmonic distortion seemed to visually correlate with lower detection thresholds (see Fig. 1 of [2]).

The majority of nonlinear distortion studies have focused on loudspeakers. Voishvillo provides a good overview of the history, present and future of nonlinear measurements for loudspeakers [5],[6]. He advocates the need using complex multitone signals because they better stimulate the nonlinearities produced by music signals compared to simple test signals used to measure harmonic and IM distortion. He argues for more perceptually meaningful distortion metrics that account for human auditory masking which can lead to better predictions of the perceived effects on quality. One such example is the model proposed by Moore et al. [7] that compares the cross correlation values of undistorted versus distorted speech or music signal fed across a bank of 40 gammatone/ERB filters that simulate human critical bands. The weighted sum of values are calculated over 30 msec time frames to calculate a metric $\mathbf{R}_{\text{nonlin}}$ that represents the perceived effect of the distortion on perceived sound quality. For distorted music and speech signals it produced high correlations ($r$ = 0.98). An interesting research question is whether the model could be adapted for predicting distortions found in headphones assuming the subjective data were available to train it.

## 3 Methodology

This section describes the methodology used to capture and reproduce the headphone distortions, as well as the types of distortion measurements that were made.

### 3.1 Selection of Headphones

Table 1 lists the seven headphones used in this study including the brand/model, transducer type/enclosure design, and the current retail price in USD based on prices listed on Amazon.com. The sample covers a wide price range ($18 to $799) and includes both dynamic and planar magnetic transducers, open and closed back models and one in-ear model. The intention was not to be exhaustive given our study's constraints in time and cost. Instead, we focused on testing a popular range of popular headphones in terms of design and cost.

The Audeze LCD-X served as our reference playback headphone having the lowest measured distortion among the headphones tested. Fig. 1 shows the normalized percentage of harmonic distortion as a function of frequency at 94 dB for the seven headphones listed in Table 1. The Audeze LCD-X has the lowest amount of distortion (< 0.1 %) over most of its frequency range relative to the other headphones.

| Brand / Model | Type | Price (USD) |
|---|---|---|
| AKG / K361 | Dynamic /Closed | $99 |
| Audeze/ LCD S20 | Planar Magnetic / Closed | $499 |

| | | |
|---|---|---|
| Audeze/ LCD-X | Planar Magnetic / Open | $799 |
| Beyerdynamic/DT7 70 Pro | Dynamic / Open | $199 |
| Loreila/X6 | Dynamic/ Closed | $15 |
| Sennheiser/ HD650 | Dynamic /Open | $370 |
| Truthear x/ Crinacle Zero: Red | Dynamic/ In-ear | $58 |

Table 1: Details on headphones used in this study.

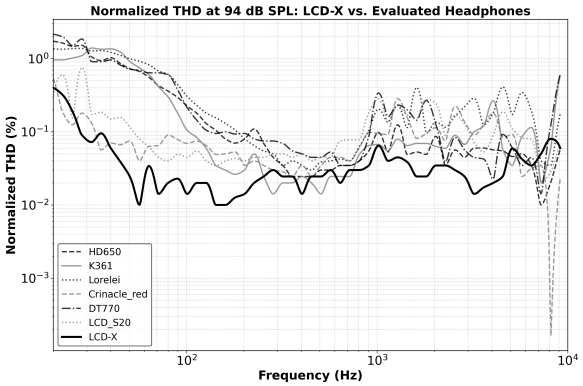



Figure 1. The normalized harmonic distortion of the headphones used in this study measured at 94 dBA SPL.

Below 300 Hz there is monotonic increase in distortion and significant divergence among the headphones in their performance. The $15 Loreila has the highest amount of distortion and the LCD-20S and Crinacle_Red have the least amount.

### 3.2 Binaural Recordings

Binaural recordings and measurements of the headphones were made using the equipment shown in Fig. 2. A computer running Listen Inc. SoundCheck (version 23) measurement software [8] generated the test sequences, stored the captured data and provided analysis. Connected to the computer was an audio interface (RME Fireface UCX 2) that distributed the test signals to the headphone amplifier (Drop + THX AAA 789) that drove the headphones on a Brüel & Kjær Type 5128-B. Signal conditioning was via the Brüel & Kjær Type 1704-C-102.

Recordings and measurements repeated from 80 dBA to +110 dBA SPL in 3 dB increments were made in a quiet audiometry booth [9]. The booth's level of background measured in 1/3-octave bands was -10 dB SPL above 100 Hz and below 100 Hz rising gradually to 40 dB at 20 Hz. A quiet noise floor was essential to minimize errors in the objective and subjective distortion measurements.

The SPL reproduced by the headphone was calibrated by adjusting the level of a pink noise signal at -16 LUF until it matched measurement target SPL. The music samples were also adjusted to -16 LUF so that the pink noise calibration signal matched the loudness of music samples when reproduced through the headphones.

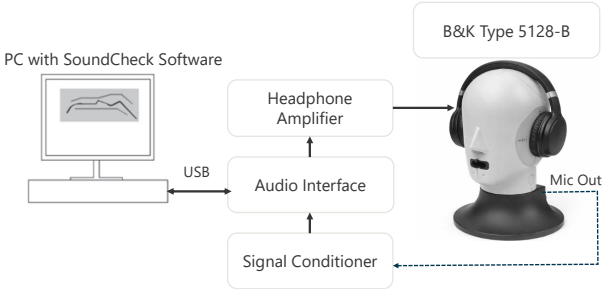



Figure 2 Block diagram of measurement and recording setup.

### 3.3 Normalization of Recordings

The recordings were post-processed to remove the linear transfer function of the headphone and equipment in Fig. 2 so that any differences among the headphone recordings were strictly traceable to nonlinear distortion and noise. The levels of the recordings were normalized to -16 LUF so that there were no level or loudness differences among the headphones and music programs during the listening tests.

### 3.4 Distortion Measurements

Five different distortion measurements made at each headphone playback level are briefly described below.

#### 3.4.1 Total Harmonic Distortion (THD)

Traditional THD measurements report the increased in level(s) of the harmonics as a percentage of the level of the fundamental of the excitation tone. THD is the most commonly reported distortion metric in the headphone industry, often with the levels of the 2nd and 3rd harmonic reported [3]. While the metric is conceptually easy to measure and understand, it been shown repeatedly to be poor at predicting

audibility of distortion using complex test signals like music [3]-[5].

#### 3.4.2 Frequency Normalized THD

Frequency normalized THD accounts for a device having a non-flat frequency response. Unlike conventional unnormalized THD discussed above, normalized THD adjusts the reference for each harmonic. The second harmonic (e.g. 200 Hz) for an excitation signal at 100 Hz is referenced to the frequency response level at 200 Hz - not 100 Hz as in the case in unnormalized THD. It is generally believed this is more accurate representation of the distortion in the measured device and its audibility.

#### 3.4.3 Intermodulation Distortion (IMD)

Two types of IMD distortion were measured as described in the IEC-60268-7 standard [10]. One uses a composite signal of 70 Hz and 600 Hz with an amplitude ratio of 4:1 to measure intermodulation (IM) products. The standard give two values Ld2 (2nd order distortion produces at 530 Hz and 670 Hz) and Lc3 (third order distortion products at 460 Hz and 740 Hz). The levels of those products are expressed as a percentage in level relative to the stimulus.

#### 3.4.4 Difference Frequency Distortion (DFD)

DFD, also part of IEC-60268-7, uses two sinusoidal signals, separated in frequency by 80 Hz to measure difference frequency distortion across the frequency band. We measured DFD from 100 Hz to 20 kHz with 12 log-spaced points per octave.

#### 3.4.5 Non-Coherent Distortion

Non-coherent distortion compares the power spectrums of the input/output signals and considers anything unrelated to the linear part to be the non-coherent part including any residual noise [1]. It can be measured with any test signal including music making it potentially a good metric for the audibility of headphone distortion reproducing music [1]. We calculated the non-coherent distortion using two music samples used for the listening tests, “Fast Car” and “Spanish Harlem”. We also measured non-coherent using pink noise.

## 4 Listening Tests

Two listening tests were designed to measure 1) the detection threshold of distortion and 2) the perceived similarity of the distortion in the headphone relative to the undistorted reference. Each test is described below.

### 4.1 Detection Threshold Test

This test measured the 67% threshold of audibility of headphone distortion using a 2ACF adaptive staircase method implemented as an ABX test [2]. In the ABX test the listener completes a series of trials where they compare A and B, where the undistorted (clean) and distorted music treatments are randomly assigned to A and B, and one assigned to X. The listener’s task is to decide if X is A or B.

An adaptative staircase rule based on Kaernbach [11] adjusts the level of the distortion up or down based on the listeners’ response. The method is more efficient than conventional up-down transform methods by adjusting the step-size of the distortion level based on up-down reversals rather than increasing the number of required consecutive correct responses. Fewer trials are required to estimate the threshold. To measure the 67% threshold of detection the step-size is doubled when an incorrect answer is registered and halved when a reversal occurs.

Fig. 3 shows an example of a sequence of trials where the level of distortion is adjusted based on the listeners’ response. The test begins at the highest distortion level and is reduced 3 dB when correct answers are recorded and increased 6 dB after incorrect responses are registered. The transition from correct to incorrect responses, and vice versa, is called a reversal (indicated by circles). The 67% threshold is calculated based on the average number *n* of specified reversals where n is an even number.




Figure 3. An example of an ABX test where the adaptive step size is adjusted based on whether the response is correct or incorrect.

A custom web-based ABX application was written to administer the test. The graphical user interface included controls to play and pause music, enter responses and create custom playback loops for the music. Feedback was provided after the listener entered their responses. The detection thresholds were measured in separate tests for three different

headphones: AKG K361, Crinacle Red and Audeze LCD S20.

#### 4.1.1 Listener Selection and Training

A total of 14 listeners participated ranging in age from 29 to 51 years old ($M$ = 34.4 years, SD = 5.7 years; 13 males, 1 female), all employees of Rtings. Ten of the listeners were trained completing 30 hours of training over period of 3 months, identifying loudspeaker distortions using the online Kippel test [15]. This test uses a similar 2AFC adaptive staircase rule but requires the listener to identify which of the pairs is distorted, whereas our test required them to identify X, which could be either distorted or clean. Our task was arguably easier since it requires no judgment about quality.

#### 4.1.2 Selection of Program

Short excerpts (< 15 s) from two music tracks (Tracy Chapman's "Fast Car" and Rebecca Pidgeon's "Spanish Harlem" were selected for the listening tests. Both tracks were used in previous distortion studies provide a means for comparison. Fast Car has a wideband spectrum with considerable bass that can excite nonlinearities in the transducer and modulate the female vocal [12]. Spanish Harlem is instrumentally sparser and provides less masking of distortions generated by the bass and produced the larger differences in listeners' headphone preferences [1]. Including these signals allows for a comparison in results across studies.

### 4.2 Similarity Listening Test

The purpose of this listening test was to study how distortion affects the perceived sound quality of music relative to the original undistorted or clean music file.

A modified MUSHRA [13] test required listeners to compare five headphones to the original undistorted reference and rate the similarity in sound quality of each on a 100-point scale. They also rated a hidden version of the reference as means to estimate their discrimination and reliability. Fig. 4 shows the graphical user interface of the custom application used to administer the test. The similarity scale has five semantic differentials (*identical, very similar, similar, somewhat different and very different)* anchored at equal increments from 90 to 10 on the scale. Five headphones (K361, HD650, DT770, Loreli and Crinacle Red) were rated using "Fast Car".

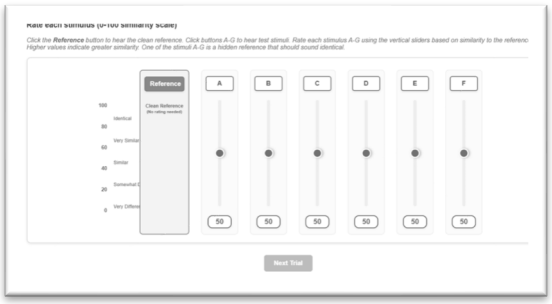



Figure 4. A screenshot of the Similarity Test listening test application.

The headphone recordings were made at level of 94 dBA SPL and reproduced through reference headphone (Audeze LCD-X) at 83 dBA. Each headphone was rated in five trials with the order of the headphones randomized. A total of nine listeners participated.

## 5 Results

### 5.1 Detection Threshold Test

An analysis of variance (ANOVA) test was performed with the independent variables including headphone (3 levels), program (2 levels) and training (2 levels) and listener to determine their influence on detection thresholds. The main factors that influenced the detection thresholds were Headphone; $F$ (2,209) = 11.80, $p < 0.001$, $\eta^2p = 0.101$, with significant interactions between Headphone and Program; $F$ (2,209) = 4.13, $p = 0.02$, $\eta^2p = 0.038$. The main effects and interactions are discussed below.

#### 5.1.1 Mean Detection Thresholds

Fig. 5 shows the distribution of detection thresholds for each headphone averaged across all listeners and programs. The boxplot shows the median, upper and lower quartiles of the thresholds with the whiskers denoting the range of values with outliers represented as circles. Also shown near the median value is the standard estimated mean with error bars. The mean detection thresholds expressed in dBA SPL for each headphone were Crinacle_Red ($M$ = 110.51, SD = 6.97), K361 ($M$ =109.64, SD = 9.372) and LCD_S20 ($M$ = 105.68, SD = 6.99). A Tukey post-hoc test indicated the difference in thresholds for Crinacle_Red and LCD_S20 were significantly different ($p$ = 0.031).

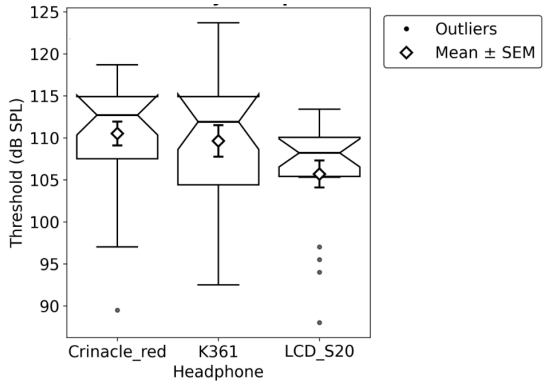



Figure 5. Boxplots showing the distribution of detection thresholds for three headphones based on all listeners and two programs.

### 5.1.2 Interaction Effect between Headphone and Program

Fig. 6 shows the interaction effect between Headphone and Program. Fast Car produced lower detection thresholds than Spanish_Harlem for headphones LCD_S20 and K361, but the opposite effect was true for Crinacle_Red.

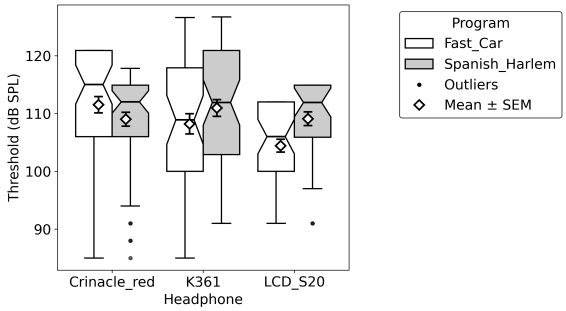



Figure 6. Boxplots show the distribution of detection thresholds for the interaction effect between headphone and program.

### 5.1.3 Individual Listeners Thresholds

While Training produced no statistically significant effect on the detection thresholds there were significant differences among the 13 individual listeners. Fig. 7 summarizes their mean detection thresholds covering a 22 dB range from 92.5 dB SPL (L12) to 114.3 dB SPL (L4). It is noteworthy that the confidence intervals are not dramatically different. There appears to be only a small benefit training listeners for this task perhaps because of its simplicity and it provides training performing the task through feedback and an adaptation to the listeners' performance.

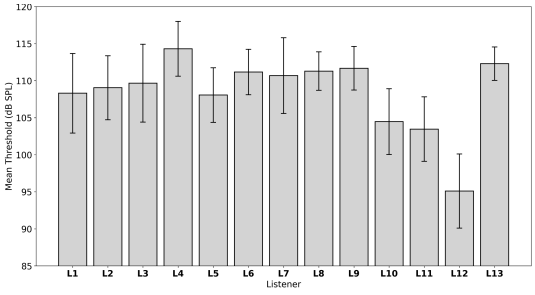



Figure 7. Individual listener detection thresholds averaged across three headphones and two music programs.

### 5.2 Similarity Listening Test

A two-way repeated measures ANOVA was performed to determine the influence of Headphones and Training on the similarity ratings. The main effect was Headphones; $F$ (5,39) = 5.82, $p$ <0.001, $\eta^2 p$ = 0.13. Tukey HSD post-hoc tests revealed that there were three significant differences in the ratings of Crinacle_Red ($M$ = 40.16, $SD$ =26.16 ) versus Clean ($M$ = 69 , $SD$ = 25.06, $p$ = <0.001 ) and two other headphones: K361 ($M$= 60.93, $SD$ =27.34, p < 0.001) and HD650 ($M$ = 57.56, $SD$ = 26.85, $p$= 0.02).

### 5.2.1 Headphone Similarity Ratings

Fig. 8 is a boxplot summarizing the main effect Headphone. The hidden reference (Clean) was most similar ($M$ = 69) to the reference. The small differences in ratings among the headphones and the large error bars suggest that the sound quality differences are small and difficult to reliably report. The one exception was the Crinacle_Red ($M$ = 40.16) which sounded the least similar to the Clean and, to a lesser extent, K361and HD650. This will be discussed in more detail in section 6.

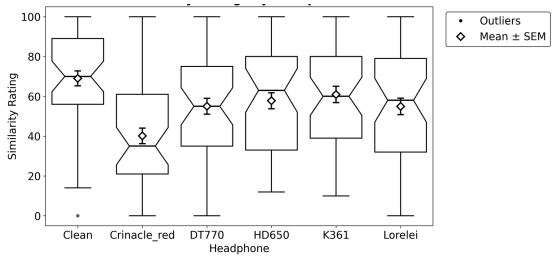



Figure 8. Boxplots showing the distribution of similarity ratings for the hidden reference (Clean) and five other headphones.

### 5.2.2 Individual Listeners

There were signficiant differences among individual listeners in how they rated the headphones. Fig. 9

shows the mean ratings and 95% confidence tervals averaged across all treatments for each listener. This graph only tells us that some listeners on average used a different part of the scale, but nothing else. Smaller confidence intervals indicates they were less discriminating and gave similar ratings to each headphone.

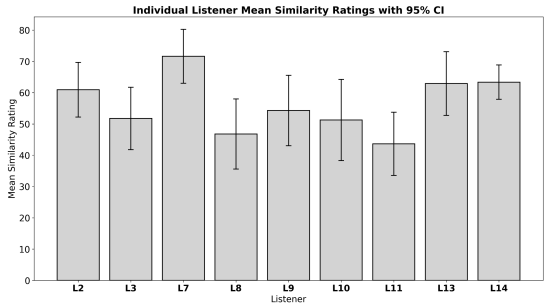



Figure 9. Mean similarity ratings and 95% confidence intervals averaged across all headphones or individual listeners.

The individual listener ratings for the headphones give more detailed information (see Fig. 10). Only one listener (L10) identified the hidden reference 100% of the time, clearly a gifted listener. In contrast, other (L2, L13, L14) had difficulty discriminating between the headphones including the Clean reference.The most discriminating listeners (L3,L9, L10, L11) could reliably discriminate between Crinacle_Red and the other headphones and disproporiately contributed to its lower ratings.

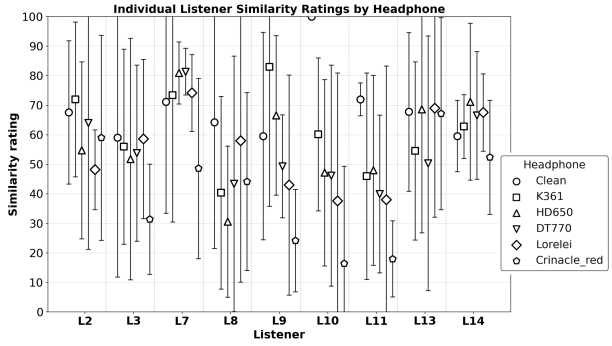



Figure 10. Individual listeners mean headphone similarity ratings and 95% confidence intervals.

## 5.2. Discrimination Test

The similarity test provided an efficient way to screen multiple headphones for audible distortion and estimate the effect on sound quality. However, the noisy results in Fig. 10 hinted that the task was difficult and the audible differences were likely above threshold. A more sensitive test was conducted to assess whether the differences were audible. Seven of the listeners performed an ABX test where their highest and lowest rated headphones in the similarity test were compared in 10 trials using the same Fast Car loop recorded at 94 dBA. Fig. 11 summarizes the number of correct identifications out of ten trials for different pairs of headphones.

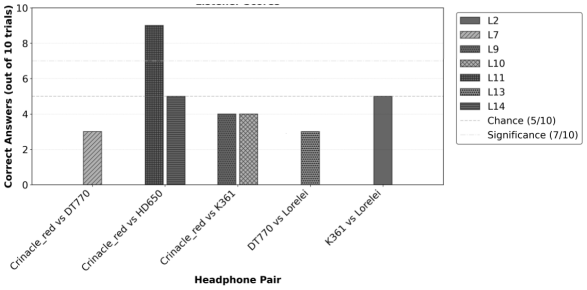



Figure 11. Individual listener results in an ABX discrimination showing the number of correct times they could identify different pairs of headphones in 10 trials.

Only listener (L11) out of seven performed better than chance (50%) identifying Crinacle_Red and HD650 in 9 of 10 trials. In summary, all but one listener could not reliably discriminate between the headphone pairs of headphones in the ABX tests even though their similarity ratings hinted otherwise. This stands as a reminder why MUSHRA tests shouldn't be used when the sound quality impairments are this small.

## 5.3 Correlation with Measurements

In this section we discuss the correlation between the headphone distortion measurements and the listeners' detection thresholds and similarity ratings.

### 5.3.1 Correlation between Measurements and Detection Thresholds

Fig. 12 plots the detection thresholds for each headphone and program (y-axis) versus the percentage of distortion (x-axis) using different distortion metrics based on their mean value. A best-fit regression line is shown along with the Pearson correlation coefficient $r$ and p-value. The graphs shown in descending order from most to least correlated are: THD ($r = 0.85$, $p = 0.03$ ) normalized THD ($r = 0.84$, $p = 0.03$ ), DLD ($r = 0.729$, $p = 0.01$ ), Non-coherent with music ($r = 0.391$, $p = 0.44$ ) and IM ($r = 0.03$, $p = 0.95$).

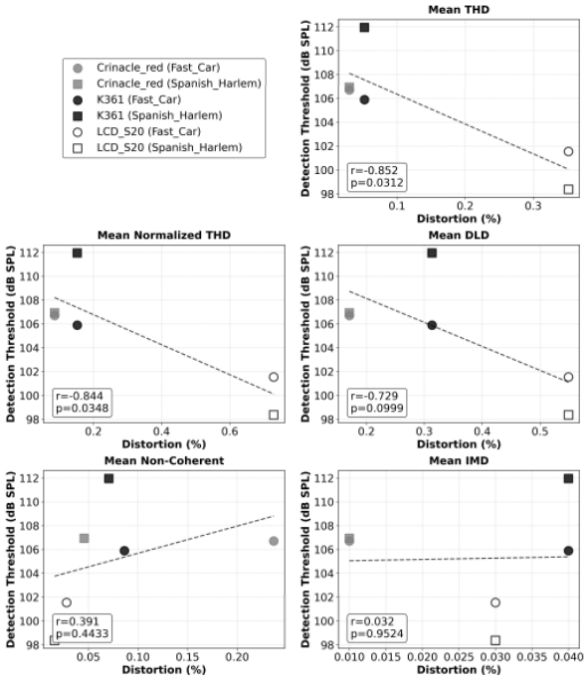



Figure 12. Correlation between detection thresholds and five distortion measurements.

Both THD metrics produce similarly high correlation values with the detection thresholds and indicate that normalized THD above mean level of 0.1 to 0.2 % may be audible in music. The detection thresholds of LCD_S20 occur at significantly lower SPLs and are associated with higher THD levels (0.7%). This headphone is a special case or outlier. After some investigation we discovered the headphone was not properly sealed on the test fixture and this affected the signal-noise below 40 Hz.

### 5.3.2 Correlation between Measurements and Similarity Ratings

Fig. 13 plots the mean headphone similarity ratings against the percentage of distortion measured at 94 dBA SPL using four different distortion metrics. The shaded areas are 95% confidence limits. Shown in ascending order from least to most correlated are: (a) DFD ($r$ = 0.209, $p$ = 0.736 ), (b) total IMD at 600 Hz ($r$ = 0.271, $p$ = 0.659), (c) normalized THD ($r$ = 0.789, $p$ = 0.113), and (d) non-coherent distortion with Fast Car ($r$ = 0.929, $p$ = 0.022). To summarize, both DFD and IMD had low correlation with the similarity ratings, normalized THD had medium correlation and the non-coherent distortion using "Fast Car" – the signal used for the listening tests - produced the strongest correlation values.

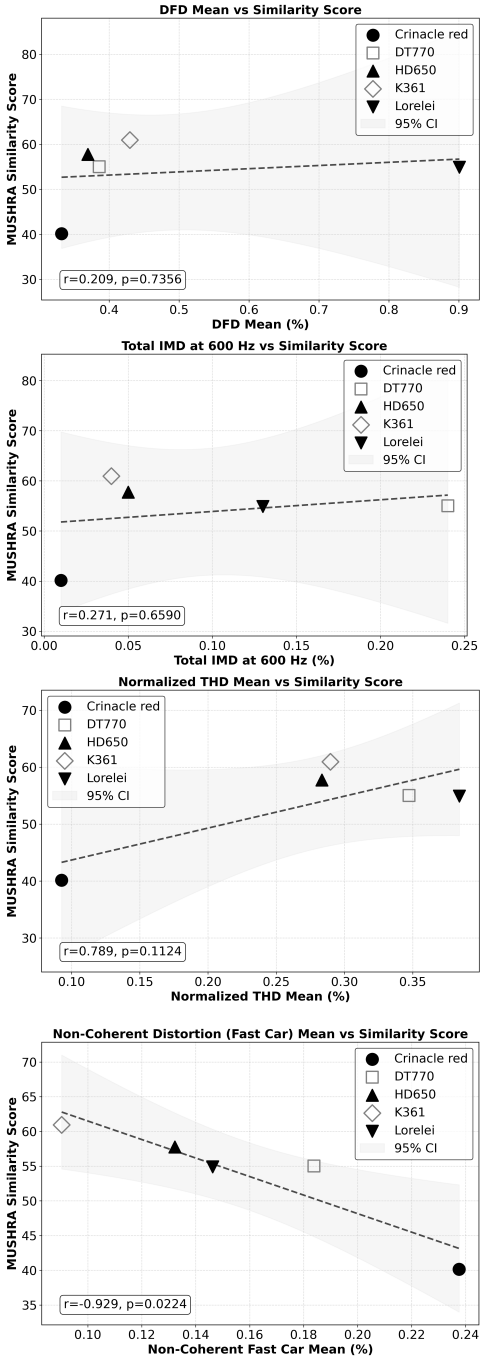



Figure 13. The correlation listeners' sound quality ratings of five headphones and their percentage of measured distortion using five different metrics: (a) DFD (b) total IMD at 600 Hz (c) normalized THD and (d) non-coherency using the test signal Fast Car.

# 6 Discussion

## 6.1 Headphone Distortion Not Audible at Normal Playback Levels

The distortion thresholds were at 104 to 112 dBA SPL depending on the headphone and music (see Fig. 6). These are 20-30 dB higher than the typical average playback levels reported in the literature that range from 70 to 85 dBA depending on study and background noise [14]-[16]. The recommended safe playback level according to OSHA is 85dBA over an 8-hour period and some audio researchers suggest 75 dBA is a safer target to protect hearing (see Chapter 16 of [4]). The practical reality is that most headphone enthusiasts are unlikely to experience audible distortion in their headphones when listening at typical, safe playback levels using the headphones tested in our study.

## 6.2 Comparison to Other Studies

The detection thresholds reported in this study were significantly higher than those reported in [2]. Seven of their nine headphones produced audible distortion below 90 dBA with five models costing < $20 USD producing distorted detected <80 dBA. Only two models costing over $200 had no audible distortion below 100 dBA. Unlike [2], we found no correlation between headphone price and audible distortion. Even our least expensive $15 headphone (Loreila X6) which had the highest THD, DFD and IM distortion at 94 dBA (see Fig. 13), could not be reliably discriminated from the Clean reference in the Similarity test. The differences in results between studies could be attributed to methodology (distortion simulated via diagonal Volterra series versus recordings) and choices of headphone and music samples.

Our results generally agree with those reported in [1] where they used a similar binaural recording and playback method and reported listener preference ratings and measurements made at 94 dBA. While they reported preference ratings – not similarity ratings – a similar observation was listeners could not reliably discriminate among the headphones in terms of their sound quality. The one exception was Headphone D in [1], which had significantly higher THD (7%) below 100 Hz and the highest non-coherent distortion between 80 Hz to 1kHz (~2%) measured with music. It is not surprising that listeners could easily detect Headphone D given its distortion is significantly higher than the detection thresholds in our study: normalized THD were <1% and <0.3% for non-coherent distortion with music (see Fig. 11). They also concluded that distortion measurements based on IM and multitone were less predictive of audible effects.

## 6.3 Differences in Listener Discrimination

Training had a negligible effect on detection of distortion that was not statistically significant ($F$ = 2.93, $p$ = 0.088, $\eta^2p$ = 0.013). Trained listeners had slightly lower detection thresholds than untrained listeners. However, two listeners were super-discriminators who could detect small amounts of distortion near the lower limits of the recordings (85 dBA). This is evident in Fig. 5 which reveals a number of outliers for headphone S_20 at the lower range of the recorded SPLs. Most likely they were detecting small amounts of background noise below 40 Hz in the recording from an air leak due a gap in the headphone's coupling to the test fixture. This prevented the listeners from completing 6 reversals and their results were not included in the final analysis. Another listener could identify the hidden clean reference in five trials but not identify headphone pairs in the discrimination test. The main takeaway is that there exits significant differences in the abilities of listeners to detect small amounts of distortion (see Fig. 6).

# 7 Conclusions

This paper presented results from listening tests where the detection thresholds of nonlinear distortion and its effect on sound quality were measured (section 4). We also reported the correlation between five common distortion measurements and the listening test results (section 5). The following conclusions can be drawn based on the evidence.

1. The distortion detection thresholds of music reproduced through headphones was significantly higher (104 to 112 dBA SPL) than typical playback levels that are considered safe (70 to 85 dBA).
2. At lower levels (94 dBA) listeners could not reliably discriminate between the headphones based on similarity in sound quality. A follow up ABX discrimination test revealed none of the listeners except one reliably identify the most dissimilar headphone pairs.
3. The percentage of THD (frequency normalized and unnormalized is strongly correlated with the detection threshold for music which can be as low as 0.1 to 0.2%. Non-coherent distortion using music signal was weakly correlated but strongly correlated with similarity ratings followed by THD. Other distortion metrics (IM,

DFD were less predictable of audibility and sound quality.

There are several limitations in this study that make prevent generalization beyond the described test conditions. Our subjective and objective measurement results are based on only six headphones and two music samples, an insufficient sample size to generalize to the larger population of headphones available. We didn't test active headphones with ANC that can minimize distortion due to feedback but introduce other types of distortions related to signal processing not tested here [17]. While more research is needed, we suspect headphone nonlinear distortion is not a practical issue at normal, safe playback levels. The frequency response remains the most important indicator in how good it sounds [4].

## 8 Acknowledgements

We would like to thank Steve Temme and the support team at Listen Inc for the loan and support SoundCheck software used for audio measurements in this study. We also thank Jérémie Voix and the team at École de technologie supérieure (ÉTS) for providing the audiometric measurement room free of charge. The facilities were made available through the GRAM (Acoustics Research Group of Montréal) and ICAR (Common Infrastructure for Research in Acoustics) laboratory.

## References

[1] S. Temme, S. Olive, S. Tatarunis, T. Welti, and E. McMullin, "The Correlation between Distortion Audibility in Headphones and Listener Preference," presented at the 137th Audio Eng. Soc. Convention, (2014).

[2] P. Gutierrez-Parera and J. Lopez, " Perception of nonlinear distortion on emulation of frequency responses of headphones," *J. Acoust. Soc. Am.* 143, 2085–2088 (2018).

[3] P. Lelièvre, "Distortion, Audibility, and Preferences," https://www.rtings.com/headphones/learn/research/distortion-audibility (January 22, 2026)

[4] F. Toole with S. Olive and T. Welti, *Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers, Rooms and Headphones,* 4th Edition, Routledge (2025).

[5] A. Voishvillo, "Assessment of Nonlinearity in Transducers and Sound Systems – from THD to Perceptual Models," presented at the 121st Audio Eng. Soc. Convention (2006).

[6] A. Voishvillo, "Measurements and Perception of Nonlinear Distortion – Comparing Numbers and Sound Quality," presented at the 123rd Audio Eng. Soc. Convention (2007).

[7] B.C.J. Moore, C. -T. Tan, N. Zacharov & V. Mattila, "Measuring and predicting the perceived quality of music and speech subjected to combined linear and non-linear distortion," Journal of the Audio Eng. Soc.,52(12), 1228–124. (2004).

[8] Listen Inc., "SoundCheck System Overview," https://www.listeninc.com/products/soundcheck/, January 22, 2026.

[9] GRAM (Acoustics Research of Montreal), https://www.etsmtl.ca/en/labs/gram/infrastructure , January 22, 2026.

[10] International Electrotechnical Commission, IEC 60268-7:2020, "Sound system equipment - Part 7: Headphones and earphones," (2020).

[11] C. Kaernbach, "Simple adaptive testing with the weighted up-down method," Perception and Psychophysics, Volume 4,v(1991).

[12] W. Klippel, "Online Listening Test," https://www.klippel.de/listeningtest/ (last accessed January 30, 2026).

[13] ITU-R BS.1534, "Method for the subjective assessment of intermediate quality levels of coding systems," Geneva, Switzerland: International Telecommunications Union (2014).

[14] S. Ahmed, S. Fallah, B. Garrido, A Gross, M King, T Morrish, D. Pereira, and S. Sharma " Use of portable audio device by university students," Canadian Acoustics (2007).

[15] S. Keith, D. Michaud, and V. Chiu, "Evaluating the maximum playback sound

levels from portable digital audio players," *J. Acoust. Soc. Am.* 123, 4227–4237 (2008).

[16] S. Widén, C. Möller, and K. Kähäri, "Headphone listening habits, hearing thresholds and listening levels in Swedish adolescents with severe to profound HL and adolescents with normal hearing," International Journal of Audiology, Volume 57, Issue 10 (2018).

[17] S. Temme, "Active Noise Cancellation Distortion in Headphones,' presented at the International Audio Eng. Soc. Conference, Helsinki, (2025).