Audio Engineering Society
Conference Paper
Presented at the Conference on
Headphone Technology
2016 Aug 24–26, Aalborg, Denmark

A Comparison of Sensory Profiles of Headphones Using Real Devices and HATS Recordings

Tore Stegenborg-Andersen[1]

[1]*DELTA SenseLab, Venlighedsvej 4, 2970 Hørsholm, Denmark*

Correspondence should be addressed to Tore Stegenborg-Andersen (tos@delta.dk)

ABSTRACT

This study compares two sets of sensory profiles of eight headphones, obtained in two experiments, with the intent of revealing the differences and or limitations of these methods: The first experiment used a double blind approach with headphone auralizations and in the second experiment assessors listened to the actual headphones, as a non-blind experiment. The results of each experiment are analyzed and compared to reveal the differences, and causes for these differences, for each attribute.

1 Introduction

Sensory testing poses interesting problems with assessor bias regardless of the domain. The number of biases are many: visual bias, tactile bias, observer bias etc. [1]. Double blind testing is generally a solution to this problem, but can be very hard to achieve with some applications.

Within the audio domain, visual- and brand influences are often times unwanted. Solving the visual bias is relatively easy when working with loudspeakers - the classic solution is hiding the loudspeakers under test behind a curtain. This option is not applicable to the headphones domain though, and thus makes visual-, brand- and tactile bias quite a challenge.

Some studies have attempted different solutions to these problems [2], [3] mounting neutral handles on different headphone sets is one solution, it does not remove tactile bias though. Auralizing one set of headphones on another is another solution. This practice has been employed by SenseLab, but few studies exist, which investigate the limitations and consequences of headphone auralization.

This study seeks to shed light on some of the consequences of said auralization, by comparing the sensory profiles from a test where the subjects had access to the actual products, with the sensory profiles from a test with auralized headphones.

An audio design company has recently launched a new headphone product where the concept is a highly customizable modular headphone. The customer is able to customize the headphones upon and after purchase. The customization options include ear-cushions, loudspeaker units, cables, headbands and loudspeaker enclosures. Some options influence only the appearance, others influence the sonic characteristics of the product. DELTA SenseLab was kindly lent several sets of these as prototypes to be able to perform the study presented in this paper.

Given the customization options, it was possible to create different headphones with the same brand, minimizing the brand bias between tested headphones. Tactile and visual biases were still present, as some headphones were circum-aural, some were supra-aural, and the cushion material varied. It is the author's presumption that the introduced biases were as small as possible without using identical headphones, meaning that the differences which are discovered between the two experiments are presumed to be as small as possible.

2 Listening tests

The listening experiments were preceded by attribute development sessions, and attribute training sessions.

Seven different versions of the customizable headphone, and a reference headphone were compared.
The customizable headphones are all closed back. The headphones referred to as A, C, D and F are supra-aural. B, E and G are circum-aural.

The two experiments (auralized and real devices) took place with ca. 2 months in between, in an attempt to reduce training bias. In the ideal situation, 50% of assessors would have performed one experiment first, and 50% the other experiment first. This was not possible due to practical considerations.

2.1 Selection of music samples

Music material was selected by two SenseLab employees, who are both expert assessors. The samples were selected to: represent different musical styles, be critical in terms of revealing the perceptual differences between products, and one was selected to reflect the musical genre of the target group for the headphone manufacturer (and meet the same criterions of criticality as the other samples).

The selected samples were:

Artist / Composer	Track title
Jennifer Warnes	Bird on a wire
George Druschetzky, Ensemble Zefiro	Serenata for winds & strings in E flat major
Helge Lien Trio	Natsukashii
Todd Terje	Delorean Dynamite

Table 1: Selected samples

2.2 Recording procedure

Each headphone was recorded using a B&K HATS 4128C, a B&K 5935L microphone power supply and an RME Fireface800 soundcard, controlled by Adobe Audition.
Fit / leakage was checked by observing a pink noise spectrum, and correcting headphone placement. In some cases, a perfect fit on both ears was not possible, due to sensitivity differences between divers, or poor match between HATS physique and headphone design.
The recording level was adjusted to a loud but comfortable level. The level was adjusted if necessary to avoid clipping on the sound card input.
The recordings were compensated for the HATS DRP-ERP response, so that the effect of the IEC60318-4 [4] coupler was removed (so that assessors did not both listen to their own, and the artificial ear canal). Sennheiser HD650 was chosen as the playback headphone, based on its high quality, open back (no problems with sealing), and the relatively low influence on the overall compensation. A measurement of HD650 and the HATS ear including coupler was performed. A smoothed filter was generated based on this, to create the final compensation filter in 1/3'rd octave bands.
The compensation was added using Adobe Audition's 30-band graphical equalizer. This equalizer uses an FIR filter. No compensation was performed over 10kHz.

The recordings were cut into samples of ca. 20 seconds, and loudness equalized across all samples and systems (headphones) using DELTA's Loudness Equalizer software. This software is implemented in Labview using a stationary Zwicker loudness model [5].
The loudness level at 67 Phon was selected for the equalization.

2.3 Attribute development

An attribute development and consensus session was held before test start. Eight trained expert assessors participated in the attribute development session.

The session occurred as follows:

- Introduction to the project – presentation of the headphone concept.
- Listening and word elicitation based on the Sound wheel from [6]. Both recordings and real devices were available for listening.

Each assessor was given a copy of the Sound wheel, and pen and paper to write down attributes, which they felt were relevant to cover the perceived sonic differences between all the products.

- Discussion of perceived differences and selected attributes. This step includes further listening if needed.

Following the session, the selected attributes for each assessor were collected and counted. For attributes which were deemed too similar/overlapping, the most frequently selected was chosen. Finally, the seven most frequently mentioned attributes were chosen:

- Bass Strength
- Treble Strength
- Midrange Strength
- Tonal balance: Dark-Bright
- Externalization
- Punch
- Clean

Attribute descriptions from [6] were used in the instructions and test-software.

2.4 Experiment with HATS recordings / headphone auralizations

The test with auralized headphones was performed in single-walled listening booths, with background noise rating of NR15 or lower, using Sennheiser HD650 headphones. The listening test was presented in SenseLabOnline, which also handled randomization and data collection automatically. A screenshot from the test is shown in Fig. 1.

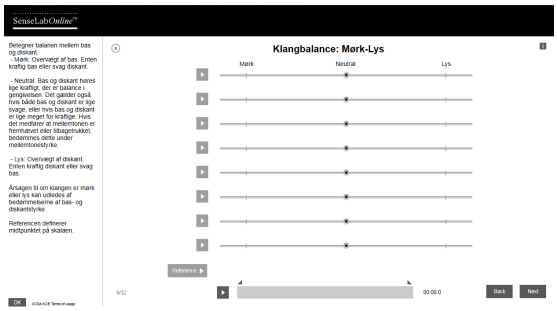

Fig. 1 SenseLab Online user interface

The sound level was set to most comfortable, and measured after the experiment to be ~80dB(A), small adjustments were made for some assessors.
Each trial in the listening test presented eight systems to be rated and a reference. The reference in this case was the original sample, played over the presentation headphone with no compensation. The 8'th system was a hidden reference. Instructions were given in the attribute descriptions on where to place the reference on the scale, meaning that the hidden reference should receive the ratings given in the attribute descriptions – if it was detected.
All systems were rated on one attribute and one sample pr. trial.
A full replicate was included in the test to check assessor performance. Most assessors did the full test in one 2-hour session. A few took two sessions to finish. A few did not finish the repetition.

2.5 Experiment w. real devices

For the experiment with real devices, a Labview software implementation was used to present and manage the listening test, as SenseLab Online did not support real device testing at this time. A screenshot is shown in Fig. 2.

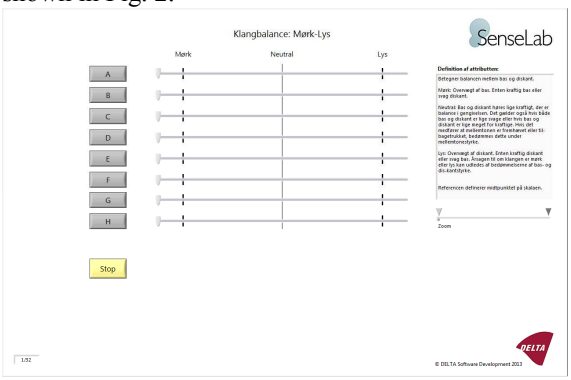

Fig. 2: Labview user interface

Each headphone was recorded, and loudness in Phone was measured using DELTA Loudness EQ software. A gain or attenuation was calculated, to achieve identical loudness, and applied to a copy of the original sample, corresponding to each headphone.
Eighteen persons participated in the experiment with real devices. The test was performed in a quiet meeting room.
The test design was identical to the HATS experiment, with the exception of the hidden reference, which could not be included in the non-blind experiment.
The reference headphone was included, and rated, in the same way as the remaining seven.
The assessors completed two repetitions in two sessions of 2 hours each. A few did not complete the repetition.

2.6 Listeners

Listeners were all participants in DELTA SenseLab's trained expert listener panel [7]. Eighteen people completed the tests, 2 females, 16 males, ranging in age from 20 to 54 years, with a median age of 29.

3 Investigation of data quality

Following the experiments, statistical analyses was performed on the data. The automatic stats analysis of SenseLab Online was utilized for this purpose.

The following investigation is based on a univariate two-way type III ANOVA. The factors of the model are Assessor, Sample, System, Condition. Two-way interactions between all factors are included as well.
In this case "Condition" describes the two experiments – using HATS recordings or real devices.

Data quality was examined in terms of the ANOVA assumptions. The normality of the residuals is checked by visually inspecting the distribution and comparing to a normal distribution, this is done for each attribute.

The homogeneity of the variance of the residuals was investigated by examining boxplots of the residuals for each attribute.

If the assumption of homogeneity of variance of the residuals is fulfilled, the boxplots will look similar. (i.e. similar interquartile differences). For all but the assessor factor, this was the case. However, for the assessors, a difference in accuracy is expected, as they do not contribute the same error.
The reference system contributes less error than the other systems. This is also expected as the reference ratings were given in the HATS experiment.

4 Test results

To remain within the scope of this study, the results are examined from the perspective of finding differences between the two experiments.

4.1 ANOVA

The condition effect is statistically significance for 4 of 7 attributes:

Attribute	F-Value	Pr(>F)
Bass strength	3.19	0.07
Treble strength	0.16	0.69
Externalization	0.44	0.51
Tonal balance	7.52	0.01
Midrange strength	61.5	<0.0001
Punch	70	<0.0001
Clean	440	<0.0001

Table 2: p and F-values for factor "Condition" for each attribute. Grey cells mean the factor is statistically significant.

The System:Condition interaction is statistically significant for 6 of 7 attributes

Attribute	F-Value	Pr(>F)
Bass strength	43.6	<0.0001
Treble strength	18.5	<0.0001
Externalization	1.98	0.05
Tonal balance	24.2	<0.0001
Midrange strength	27.1	<0.0001
Punch	37.1	<0.0001
Clean	19.9	<0.0001

Table 3: p and F-values for interaction between factors "System" and "Condition". Grey cells mean the interaction is statistically significant.

The Sample:Condition interaction is statistically significant for 2 of 7 attributes

Attribute	F-Value	Pr(>F)
Bass strength	0.68	0.56
Treble strength	4.29	0.01
Extnalization	0.56	0.64
Tonal balance	1.75	0.16
Midrange strength	1.4	0.24
Punch	1.5	0.21
Clean	3.38	0.02

Table 4: p and F-values for interaction between factors "Sample" and "Condition". Grey cells mean the interaction is statistically significant.

4.2 Mean and 95% confidence intervals

Mean and 95% confidence intervals were calculated using SenseLabOnline, and are shown in Fig. 3.
Each plot shows ratings for one attribute, averaged across 18 assessors and 4 samples. The repetition is not included.
The square shows the mean value, and the confidence bars show the 95% confidence intervals. Along the horizontal axis are systems, and along the vertical axis are ratings. Numerical values and axis labels are shown on the side of each plot. Numerical values were not visible to assessors.

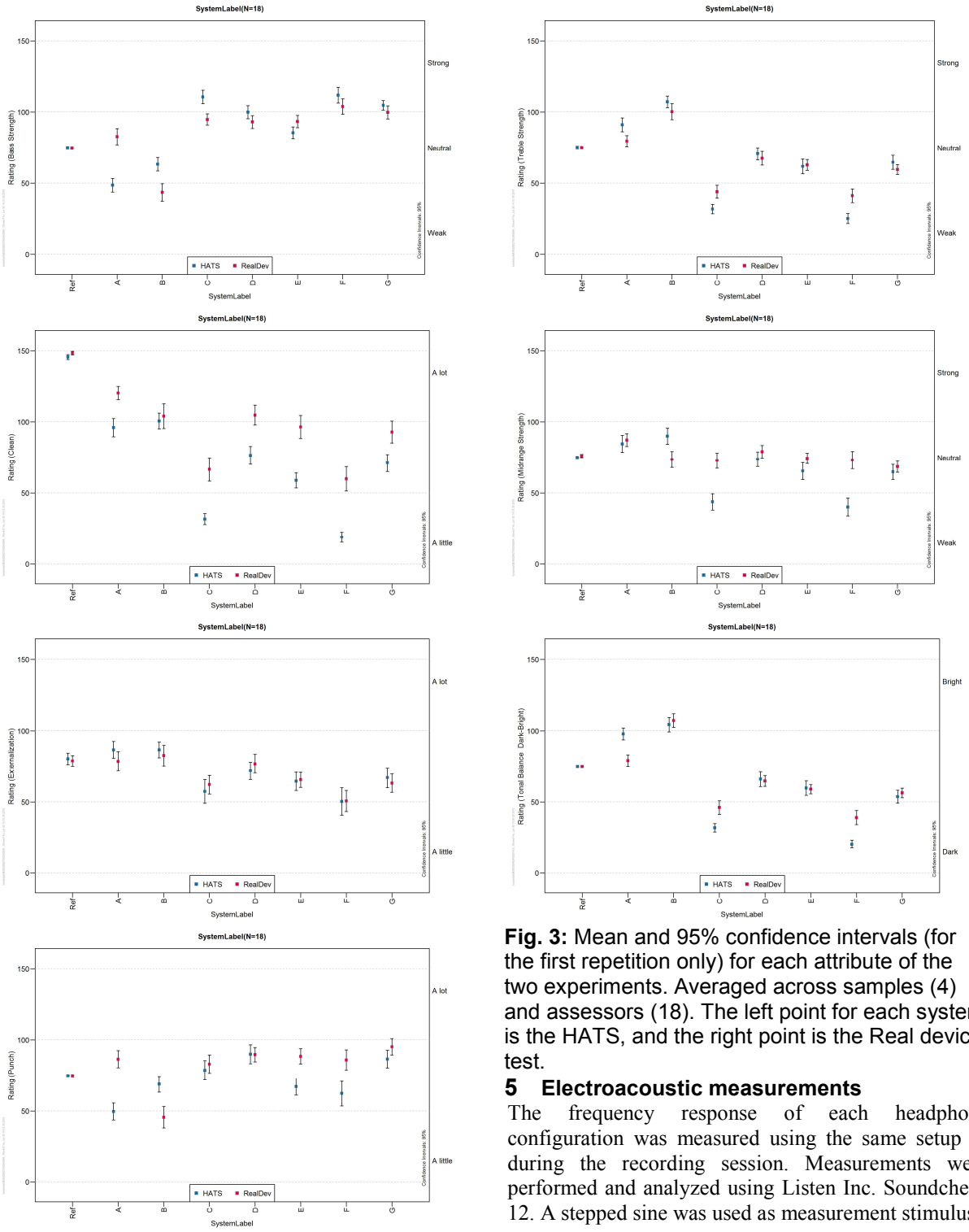

Fig. 3: Mean and 95% confidence intervals (for the first repetition only) for each attribute of the two experiments. Averaged across samples (4) and assessors (18). The left point for each system is the HATS, and the right point is the Real device test.

5 Electroacoustic measurements

The frequency response of each headphone configuration was measured using the same setup as during the recording session. Measurements were performed and analyzed using Listen Inc. Soundcheck 12. A stepped sine was used as measurement stimulus.

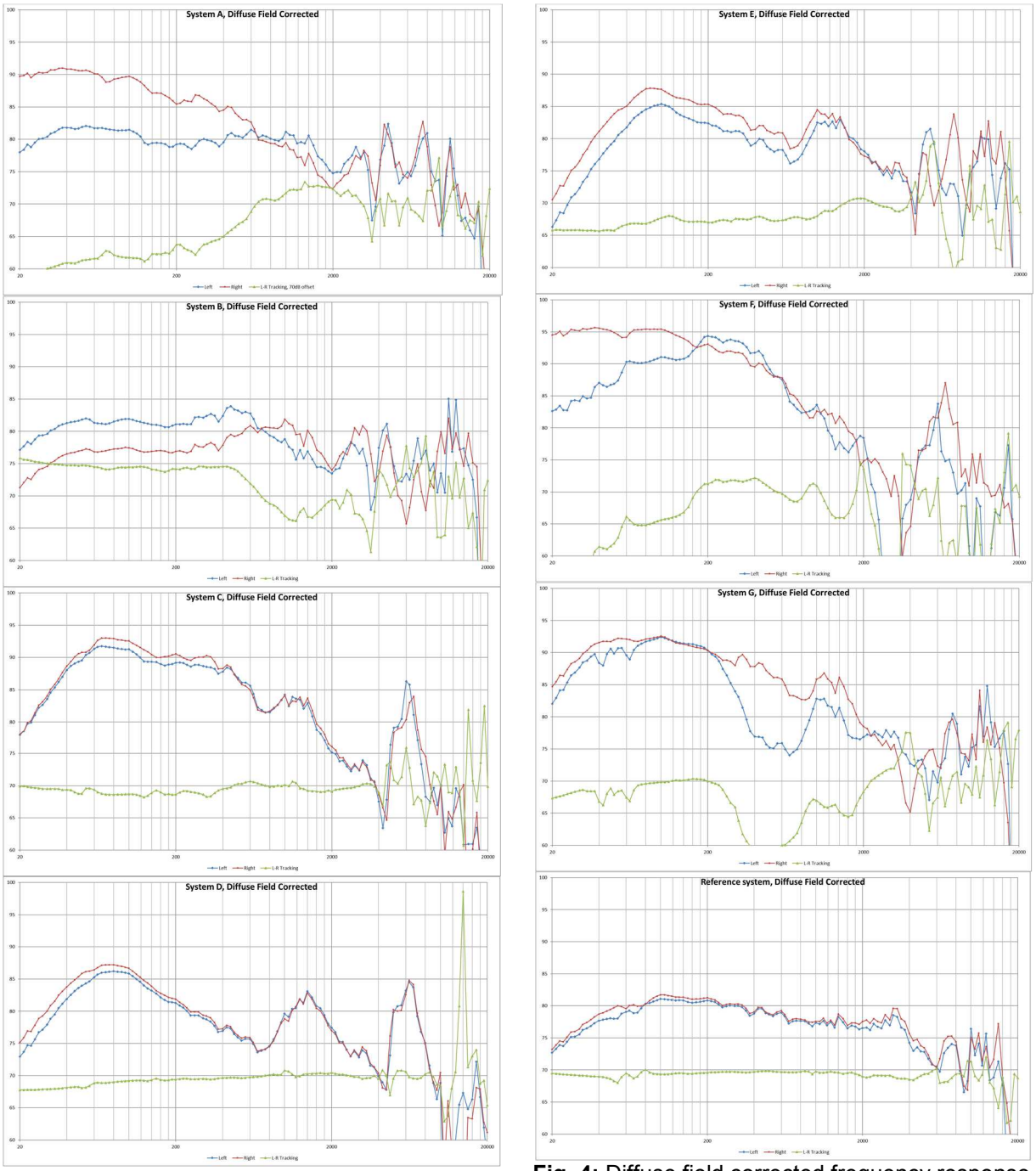

Fig. 4: Diffuse field corrected frequency response measurements of each headphone configuration. L-R tracking is offset to 70dB.

Unfortunately, electroacoustic measurements are not available for the auralizations.

6 Discussion

Overall there are statistically significant differences between the two experiments. This is clear from the ANOVA tables, and Fig. 3.

6.1 Bass strength

For the attribute "Bass strength", the condition alone does not have a significant effect on the results however the combination of condition and system does, meaning that *some* systems were perceived differently between the two experiments. Looking at Fig. 3, to see what causes this, it is clear that systems A, B and C were rated differently in the two experiments. As all headphones, except the reference, are closed, the explanation for these differences are likely in leaks, either to the HATS during recording, or to the assessors during the test. Headphone B is circum-aural and quite large, so it *should* be easy to seal to the head. It was discovered that the cushion was loose on one side of this headphone during the real device test, and therefore a rubber gasket presumably did not seal properly to the loudspeaker enclosure. Headphone A did not seal well to the HATS, and thus has significantly less bass on the recording than it does in practice.

For headphone C, the explanation for the difference is not as simple. A possible explanation is that a compression of the extreme ends of the scale is happening. It can also be noted that headphones C, D and E are not rated significantly different. This also hints that the scale is compressed, likely because the headphones cannot be switched between instantly, as they can for the HATS condition. A similar scale compression was observed in an almost identical experiment in [2].

Comparing perceptual measurements with electroacoustic measurements (Fig. 4), it can be observed that systems C-G have significant bass boosts, and are perceived as having more bass than the reference and systems A and B. It can also observe a large difference between right and left channel for some systems. This is assumed to be caused by a general problem with loose cushions/the rubber gasket (For the bass region).

6.2 Treble strength

The condition factor is not statistically significant for treble strength. Factor interactions sample:condition and system:condition are though. In this case, looking at Fig. 3 gives the impression that a compression of the scale took place in the real device experiment, causing the extremes to be slightly less extreme in the real device experiment (systems C and F). This could be caused by the inability to switch instantly between systems as was possible in the HATS condition [2].

System A is perceived differently in treble strength. This is most likely explained by the large difference in bass strength, as the two *can* influence each other.

The conclusion is that treble strength differences might be exaggerated when the option to crossfade between headphones is available or vice versa.

Electro acoustic measurements are difficult to interpret in the treble domain, but systems C and F do seem to be the most extreme in their lack of treble.

6.3 Externalization

Externalization does not show statistical significance in the condition factor or the interactions with it. After the experiment, it is evident that forcing the reference to define the center of the scale for this attribute, more or less compressed the scale into 50% of the original size. This makes the likelihood of significant differences lower, and thus more or less breaks the attribute, as it was not an easy attribute to discriminate in the first place. The expectation from the author was that some systems may be perceived as having more externalization than the reference. This was apparently not the case.

6.4 Tonal balance: Dark-Bright

The condition factor and the system:condition interaction shows significance for this attribute. It can be seen in Fig. 3 that the differences between experiments match well with the differences, which are also evident in the other spectral attributes – bass strength and treble strength. Systems A, C and F are significantly different. System A is influenced by the same issue as influences it in other attributes. Systems C and F reflect the same scale compression issue that is observed for treble strength.

Electro acoustic measurements make sense in the case of this attribute, as the overall tilt of the curve somewhat matches the perceptual results.

6.5 Midrange strength

For midrange strength the condition factor and the system:condition interaction have statistical significance. Observing Fig. 3 shows that this attribute

did not give good discrimination for the real device test. For the HATS condition there is better discrimination. The explanation is likely again the ability to switch instantly between systems in the HATS condition, making it easier to compare smaller perceptual characteristics.
The electro acoustic measurements do not seem to reveal what exactly causes the differences in the perceptual results.

6.6 Punch

The condition factor and the system:condition interaction showed statistical significance.
As with midrange strength the discrimination for the real device condition is not good, making a comparison of the results difficult. For systems A and B, a result which is very similar to Bass Strength is observed. Based on the definition of Punch as something relating to accuracy of bass, this makes sense. For systems E and F, a significant difference between conditions is observed. It is unclear where this comes from, but similar differences are seen in Clean and Midrange strength, perhaps they influence the Punch rating. From the authors perspective, there is no clear mapping between Punch and frequency response.

6.7 Clean

This is perhaps the most interesting attribute for this study, as it shows a large significance in all 3 factors/interactions relating to condition. It is clear from Fig. 3 that "Clean" does not give the same results with an auralized headphone, as it does with the real headphone. The ordering appears to be the same, with systems A and B not showing statistically significant difference. Sadly, it does not seem to be possible to derive, from the other attributes, why this is the case. The author's best guess is that the dissimilarity between the HATS 4128C ear + coupler and the ears of the assessors are the issues causing this. As no compensation was introduced above 10kHz, there *must* be differences. These differences should be the same across all systems, as all systems were recorded / auralized in the same way. This does seem to be the case, with the exception of headphone B.
The conclusion from this could be that a better approximation to the human ear is necessary for headphone auralizations to be completely viable. However, since human ears are far from identical [8], individually measured DRP-ERP curves might be the only option.
Clean might be related to treble strength, as we see similar results, however we do not observe the same overall difference between conditions for Treble strength, so there must be a difference. Again the electro acoustic measurements are not clearly mapped to the perceptual measurements.

7 Summary

A study comparing sensory profiles of auralized headphones and sensory profiles of real headphones was performed on 8 headphones on 7 attributes and 4 samples, by 18 trained expert assessors. The results showed significant differences for all but one attribute.
It was hypothesized that the cause of these differences can be found in:

- Leaks during the recording process
- A faulty device (leaking) in the real device experiment
- Scale compression in the real device experiment
- Assessors were unable to discriminate between systems for some attributes without the ability to switch instantly between them.
- For at least one attribute the differences between experiments seems to stem from the difference between HATS ear and assessor ear.

8 Acknowledgements

This work was funded by DELTA and the Danish Agency for Science, Technology and Innovation. The author wishes to thank the anonymous audio design company, for supplying the headphones for test, and Christer P. Volk for his invaluable feedback during the paper writing and analysis.

9 References

[1] E. C. Poulton, Bias in Quantifying Judgements, 1989.

[2] S. E. Olive, T. Welti and E. McMullin, "Virtual headphone listening test methodology," in *AES 51'st conference*, Helsinki, 2013.

[3] T. Hirvonen, M. Vaalgamaa, J. Backman and K. Matti, "Listenint test methodology for headphone evaluation," in *AES Convention 114*, Amsterdam, 2003.

[4] International Telecommunication Union, "Recommendation ITU-T P.57 Artificial ears," International Telecommunications Union, 2011.

[5] E. Zwicker and B. Scharf, "A model of loudness summation," in *Psychological review*, 1965, pp. 3-26.

[6] T. H. Pedersen and N. V. Zacharov, "The development of a soundwheel for reproduced sound," in *AES Convention 138*, Warsaw, 2015.

[7] S. V. Legarth and N. V. Zacharov, "Assessor selection process for multisensory applications," in *AES 126'th Convention*, Munich, 2009.

[8] A. T. Christensen, W. Hess, A. Silzle and D. Hammershøi, "Magnitude and phase response measurement of headphones at the eardrum," in *AES 51'st Conference*, Helsinki, 2013.