Audio Engineering Society

Conference Paper

Presented at the Conference on
Headphone Technology
2016 Aug 24–26, Aalborg, Denmark

Modelling perceptual characteristics of prototype headphones

Christer P. Volk[1,2], Torben H. Pedersen[1], Søren Bech[2,3], and Flemming Christensen[2]

[1]*DELTA SenseLab, Venlighedsvej 4, 2970 Hørsholm, Denmark*
[2]*Department of Electronic Systems, Aalborg University, Fredrik Bajers Vej 7, 9220 Aalborg East, Denmark*
[3]*Bang & Olufsen A/S, Peter Bangs Vej 15, 7600 Struer, Denmark*

Correspondence should be addressed to Christer P. Volk (cvo@delta.dk)

ABSTRACT

This study tested a framework for modelling of sensory descriptors (words) differentiating headphones. Six descriptors were included in a listening test with recordings of the sound reproductions of seven prototype headphones. A comprehensive data quality analysis investigated both the performance of the listeners and the suitability of the descriptors for modelling. Additionally, two strategies were investigated for modelling metrics describing these descriptors, both relying on specific loudness estimations of the test stimuli. The stability of the initially found metrics was tested with a bootstrap procedure to quantify the potential of the metrics for future predictions within the perceptual space spanned by the headphones. The most promising results were metrics for Bass, Clean and Dark-Bright with correlations values of $r^2 = 0.62$, $r^2 = 0.58$, and $r^2 = 0.90$ respectively.

1 Introduction

The development of headphones can be a process involving many prototypes while exploring the potential of new technologies, constructions, or materials. At some point in the process, the many paths explored must be narrowed down to a single track. This study explores the possibility of modelling the link between the physical sound reproduction and the perceptual characteristics of headphones, thereby potentially providing an indication of which path will lead to the desired design target.

Perceptual characterisation of sound reproduction offers evaluation of performance based on human perception as an alternative to traditional electro-acoustical measurements, such as frequency- and time responses. It can provide valuable insight into the dominating perceptual characteristics of e.g. a set of headphones. An issue with perceptual evaluations is however the time and resources required to collect data. Consequently, a number of mathematical models based on less resource-heavy electro-acoustic measurements have previously been proposed. They were able to predict preference [1], mean opinion score [2, 3] or stereo width [4], thereby providing insight as to sound quality performance. Earlier efforts with prediction of loudspeaker preferences from visual inspection of frequency responses were also attempted by Toole in [5], who concluded that: *"Listeners, it seems, like the sound of loudspekaers with a flat, smooth wideband on-axis amplitude response that is maintained at substantial angles off axis"*.

Within external preference mapping [6], the underlying decision process leading to a listener's preference is assumed to be a weighted sum of perceived auditory characteristics. These weights are based on a personal reference consisting of desired features and can be influenced by prior experience, context, mood etc. [7]. The concept can be described by Eq. 1 for a product i. $S_X(i)$ represents the salience of the characteristics described by a sensory descriptor X and the constants α to ω are an individual's weightings of the characteristics' importance for preference. ε denotes the residual. Note that the relationship between S_X terms may be non-linear, although a linear case is illustrated in Eq. 1.

$$Preference(i) = \alpha S_1(i) + \beta S_2(i) + \cdots + \omega S_N(i) + \varepsilon \quad (1)$$

The weights are individual and subjective, while the salience of an auditory characteristic, S_X, is considered objective and depending only on the auditory acuity of listeners. Leaving out the subjective weights, not all S_X-terms are relevant for product characterisation of a given subset of audio reproduction products being evaluated (compared). A limited number of terms are likely to dominate the overall sensation, but which terms that is will depend on the products being evaluated. Differences in dominating characteristics are for instance seen between the headphones study [8] by Gabrielsson and Sjögren from 1979 and the headphones study [9] by Olive and Welti from 2012. In the older study, the analysis led to characteristics with emphasis on artefacts, while the newer study led to characteristics with emphasis on spectral differences.

In the present study a framework was established and tested for finding metrics able to predict the perceptual characteristics of headphones sound reproductions. Recordings were made of seven prototype headphones' reproduction of a selection of musical excerpts. The recordings were used as stimuli in a listening test as well as the basis on which proposed metrics were calculated. The metrics were thereby developed directly on the basis of what listeners perceived. This approach was also used in e.g. [10] to study the dominating perceptual dimensions differentiating monophonic loudspeaker reproduction in a room. The listening test of the present study consisted of evaluation of a number of sensory descriptors[1] by expert listeners, with the purpose of characterising the perceptual space spanned by the headphones. The perceptual ratings were modelled on the basis of estimated stimuli loudness spectra, and consequently the non-linearities of the human auditory processing are incorporated in the proposed metrics.

[1] A sensory descriptor is defined here as a word or phrase that describes, identifies, or labels a perceptual characteristic of a system, e.g. a headphone reproduction. This definition is adapted from [11].

2 Listening test

2.1 Headphones

A total of eight electrodynamic headphones were included in this study. Seven were prototype models from one manufacturer, and one additional set of high-end headphones was included as a reference. The seven prototypes consisted of four supraaural and three circumaural. All the prototypes were closed-back headphones, while the reference headphone model was circumaural and open-back.

2.2 Stimuli

Four musical excerpts were played over the eight headphones. The reproduced audio was recorded, post-processed, and presented to listeners over a pair of Sennheiser HD 650 headphones (playback headphones). In the recording process the headphones were placed on a Brüel & Kjær 4128C head- and torso simulator (B&K HATS). To minimize (asymmetrical) leakage, the headphone positioning was checked by recording pink noise and comparing the right-left input level balance. The amplifier gain was adjusted for each set of headphones to record at approximately the same sound level across all the headphones (mean $L_{eq} = 69.9\,dB\,(A)$, $\sigma = 5.4\,dB$, variation dominantly due to musical excerpt differences). The binaural recordings were captured using a RME Fireface 800 soundcard with a 24 bit A/D converter and saved at sample rate of $48\,kHz$.

The recordings were post-processed to compensate for the influence of the ear-canals of the B&K HATS as well as for the frequency response of the playback headphones. The compensation was performed by means of a equalizer with 1/3-octave band minimum-phase FIR filters on both channels, i.e. without compensation for left-right imbalance in sensitivity. The filter had a dip in the range $80-400\,Hz$ with a minimum of $-1.4\,dB$ at $125\,Hz$ and another in the range $0.5-12.5\,kHz$ with a minimum of $-12\,dB$ at $3.15\,kHz$. The post-processed recordings were loudness normalized using an automated process, where loudness was estimated using

a stationary loudness model [12] and iteratively level-adjusted to reach a target (channel-averaged) of equal loudness at a playback level of $67 \pm 0.01\,Phon$. Finally, the stimuli were converted to 16 bit WAV files for reasons of compatibility with the test software. The post-processed headphone recordings are referred to as auralised headphones in the following.

A total of ten musical excerpts were originally recorded. During a session of informal listening by two experienced listeners, four were selected for the listening test, as they were perceived to facilitate the largest discrimination between headphones: Jennifer Warnes ('Bird on a Wire', Famous Blue Raincoat, 1987), Todd Terje ('Delorean Dynamite', It's album time, 2014), Helge Lien Trio ('Natsukashii', Natsukashii, 2011), and George Druschetzky Ensemble Zefiro ('Serenata for winds & strings in E flat major: Maestoso, Allegro', Druschetzky: Quartetto; Serenata; Quintetto, 2002). These four excerpts represent the musical genres: Pop, Electronic, Jazz, and Classical respectively.

2.3 Test procedure

The listening test comprised of evaluations with six sensory descriptors selected as suited for discriminating between the headphones. The selection was based on consensus meetings with trained listeners [13], specifically trained in the sensory descriptors described in the DELTA-developed Sound wheel [11] and all descriptors were consequently selected among these. The chosen descriptors were: Bass strength, Midrange strength, Treble strength, Dark-bright, Clean, and Punch. They comprised sensory descriptors from three main groups: Dynamics, Timbre and Transparency. Danish names and definitions were used, as all listeners were native Danish speakers.

The listening test consisted of evaluation of the eight headphones with regards to each sensory descriptor on a $15\,cm$ rating scale anchored by two words specific for each descriptor. The reference headphones were included both as a labelled reference and as a hidden anchor system. The definitions included instructions on which rating to give the hidden anchor (if identified). Consequently the ratings of the reference headphones were excluded from the data analysis. The SenseLabOnline test software (senselabonline.com), allowed listeners to listen to each stimulus as many times as needed and switch between stimuli almost instantaneously. One "screen" in the user interface included stimuli for one musical excerpt and evaluation on one sensory descriptor with all auralized headphones. The full test comprised 48 "screens" (six descriptors, four musical excerpts and 2 repetitions) presented in randomised order and evaluated during one 2-hour session per listener. Listeners were encouraged to take breaks on a regular basis. The playback level during the test was set to approximately $80\,dB(A)$ (measured with the playback headphones positioned on a B&K HATS at the default level setting). The listeners did however have the option to ask the test leader for small level adjustments ($\pm 4\,dB$) during familiarisation to accommodate a comfortable listening level for the individual. Level adjustments affect the perception of the spectral balance of the stimuli, but makes the long sessions more comfortable for listeners and have a small influence in comparison to the natural variation in hearing between listeners.

2.4 Listeners

Eighteen listeners participated in the listening test. All were trained listeners from DELTA SenseLab's expert panel. Among the 18 listeners eight were trained specifically in the sensory descriptors of the Sound wheel [11]. The listeners ranged in age from 20 to 54 with a median of 29, and all had their hearing tested both prior to joining the expert panel as well as periodically afterwards.

2.5 Listening test results

2.5.1 Listener performance

The performance of the participating listeners was evaluated per sensory descriptor by means of two statistical measures, which will be briefly explained: the eGauge metrics Discrimination and Reproducibility [14] as well as Tucker-1 analysis [15]. Discrimination describes a listener's ability to statistically discriminate between systems (headphones), i.e. a measure of how big an influence the systems have on the ratings - as opposed to other factors, such as the influence of musical excerpts, conditions, etc. Reproducibility describes how consistent a listener rate the same stimuli (a headphone and musical excerpt combination in this case) between repetitions. The Tucker-1 analysis is based on a Principal Component Analysis (PCA) and was used here to gauge listener performance in terms of consistency and agreement with the panel average/consensus.

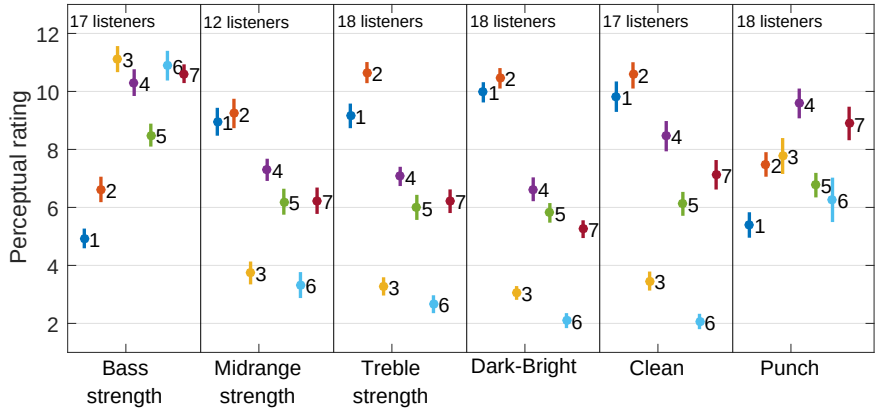

Fig. 1: [Colors online] Mean and 95% confidence intervals of the six sensory descriptor ratings. Each point and number represents the rating of one auralized headphone averaged over selected listeners, musical excerpts, and repetitions. The number of listeners included for each descriptor is displayed above the ratings.

Listeners below the noise floor (performance of random evaluations), with regards to Discrimination and Reproducibility, were removed from the dataset prior to further analysis, as was listeners within the inner ring (less than 50% explained variance) of the Tucker-1 loading scores plot. This meant removing six listeners from Midrange strength, and one listener from both Bass strength and Clean. Midrange strength was thus a difficult sensory descriptor to evaluate for the listeners on the presented stimuli. In addition, the Tucker-1 analysis of the sensory descriptors, showed a wide spread of listeners' ratings within the two circles for the Punch sensory descriptor, which signifies lack of agreement between listeners. This could be caused by e.g. listeners' needing more training or sensory descriptors which are not well-defined or ill-suited for the purpose.

2.5.2 Sensory descriptor assessment

In Fig. 1 the mean ratings for each sensory descriptor are depicted[2], based on ratings of the listeners that was not removed as a result of the performance criteria described in the previous section.

In terms of modelling perspectives, the most important prerequisite, was considered to be the sensory descriptors ability to discriminate between the headphones, e.g. to model Bass strength, the ratings of

[2]The interaction between factors 'Headphone' and 'Musical excerpt' is significant for most sensory descriptors (see Table 1), but its influence (F-value) is at least one order of magnitude lower than the main 'Headphone' factor. Consequently, Fig. 1 shows the average over excerpts.

Variable	MS	F	Pr > F
Sys (**Bass str.**)	79000	356	< 0.0001
Sys/sample	1630	7.33	< 0.0001
Sys (**Midrange str.**)	51300	233	< 0.0001
Sys/sample	187	1.58	0.06
Sys (**Treble str.**)	120000	626	< 0.0001
Sys/sample	1220	6.36	< 0.0001
Sys (**Dark-Bright**)	145000	701	< 0.0001
Sys/sample	1950	9.45	< 0.0001
Sys (**Clean**)	137000	448	< 0.0001
Sys/sample	507	1.67	0.04
Sys (**Punch**)	30800	82.3	< 0.0001
Sys/sample	2880	7.68	< 0.0001

Table 1: System effect (6 DF) and system/sample interaction effect (18 DF) from 4-ways ANOVA tables for each sensory descriptor. Sys/sample is the interaction between headphone and musical excerpt. Mean square (MS) and F-values are rounded to three significant digits.

the headphones was required to be significantly different from each other and preferably span the majority of the rating scale. A 4-ways (Headphone×Musical excerpt×Listener×Repetition) fixed-effect analysis of variances (ANOVA) was conducted for each descriptor to test its ability to discriminate between headphones. The results are presented in Table 1 and showed that all sensory descriptors were able to discriminate between two or more of the headphones. Punch, however, had the lowest F-value (discriminatory power) and combined with poor agreement in the Tucker-1 analysis, no efforts were done to model this descriptor. The interaction between headphones and musical excerpt was significant for all descriptors with the exception of Midrange strength ($p \leq 0.05$). A similar p-value was however seen for Clean as well.

r	Bass	Dark-Bright	Mid	Treble	Clean	Punch
Bass	**1.00**	-0.87	-0.81	-0.79	-0.72	0.58
Dark-Bright	-0.87	**1.00**	**0.97**	**0.97**	**0.94**	-0.17
Mid	-0.81	**0.97**	**1.00**	**0.97**	**0.98**	-0.05
Treble	-0.79	**0.97**	**0.97**	**1.00**	**0.96**	-0.03
Clean	-0.72	**0.94**	**0.98**	**0.96**	**1.00**	0.08
Punch	0.58	-0.17	-0.05	-0.03	0.08	**1.00**

Table 2: [Colors online] Pearson correlation coefficients for the sensory descriptor ratings. The descriptors are sorted in descending order of their absolute correlation with bass. The intensity of the background color represent the degree of correlation. Bold numbers have $r > |0.90|$.

A correlation analysis of the sensory descriptor ratings, based on mean values from well-performing listeners, are shown in Table 2. The correlations seen to be high between all descriptors - with the exception of Punch. Due to the listeners' disagreement in the rating of Punch it was however not possible to know whether the descriptor comprised an important perceptual dimension or simply a dimension of noise. It is noteworthy that Dark-Bright had a higher correlation with Treble strength than with Bass strength. Furthermore Clean was highly correlated with Midrange strength, Treble strength as well as Dark-Bright. While the ratings of four sensory descriptors were highly correlated it remained of interest to model all as each potentially represented a separate coupling to the physical world. Consequently they had varying performance potentials, making it of interest to model all and select the most promising in terms of prediction capabilities. Note, however, that correlations between sensory descriptors, does not imply that the sensory descriptors refers to the same percept in general, as similarity between this subset of headphones would lead to high correlations as well.

3 Modelling methodology

Metrics were developed to investigate the link between the sensory descriptor ratings and the listening test stimuli. They were all based on a stationary loudness model [12], where specific loudness was estimated for each stimulus in steps of 0.1 Bark from Bark number 0.1 ($20\,Hz$) to 24.0 ($15.5\,kHz$). The resulting loudness metrics (presented later on in Table 3) are the summed loudness in a frequency range AB divided by the sum of the full loudness spectrum, as described by Eq. 2. $Dens_m(f)$ is the temporal mean of the time-varying specific loudness, while A and B denotes the frequency limits of the AB range.

$$metric = \frac{\text{AB range}}{\text{Full range}} = \frac{\sum_{f=A}^{B} Dens_m(f)}{\sum Dens_m} \tag{2}$$

An optimisation routine was implemented to find the frequency range AB, at which the metrics had maximum correlation with the sensory descriptor ratings. As it was not of interest to get metrics related specifically to individual musical excerpts, the routine found the frequency range where the maximum average correlation across excerpts were located, as described by Eq. 3 and 4. $Metric$ is a matrix with $Dens_m(AB)$ loudnesses for all tested combinations of AB frequency ranges. P_{ex} is the average perceptual ratings for each set of headphones for the musical excerpt ex. R_{ex} is the Pearson correlation matrix for all tested combinations of AB frequency ranges. R_{max} is the maximum Pearson correlation averaged over all P_{ex}. A metric's AB frequency range was consequently the range where R_{max} was found.

$$R_{ex} = corr\,(Metric, P_{ex}) \tag{3}$$

$$R_{max} = max\left(\frac{\sum_{ex=1}^{4} R_{ex}}{4}\right) \tag{4}$$

All combinations of search ranges in the full loudness spectrum and all search positions were tested with the 0.1 Bark resolution with one constraint: The search range was restricted to AB ranges with a minimum width of two Barks, to avoid getting peak correlations in narrow ranges unlikely to be the cause of the perceptual rating. In the remaining part of this paper, the output of the optimisation routine, when analysed with all seven prototype headphones, are referred to as the 'baseline'.

The output of the optimisation routine was likely to lead to several peaks due to the wide search area. This ensured a search unbiased by the authors' theories, but complicated the analysis. Even if one peak had a (significantly) higher correlation coefficient, this may not have been the case with slightly different prototypes. This issue is commonly dealt with in the literature by training the metric on one set of data and validating the results with a separate dataset. This is however not desired with small datasets, such as a limited number of prototype headphones. Therefore, a bootstrap method was used to get a better representation of the sample space spanned by all prototype headphones. A total of 500 bootstrap iterations (sampling with replacement) of the optimisation routine were processed for each descriptor, followed by a classification task: For each iteration the optimisation routine output's maximum peak was classified either as matching one of the peaks from the baseline or an unclassified peak ("Other"). A match was defined as: A and B from the AB range being within $\pm 2\,Bark$ of a peak from the baseline. Due to the limited number of headphones, this process was likely not to have a high hit rate, i.e. maxima's coinciding with the baseline peaks, but still allowed comparison of hit rates across peaks as well as providing estimated correlation coefficient confidence intervals.

3.1 Dark-Bright metric

For modelling of the Dark-Bright descriptor another approach was chosen. In the literature a metric commonly referred to as the spectral centroid (see e.g. [16, 17]) is reported to be a good predictor of brightness (a sensory descriptor similar to Dark-Bright). It is the balancing point in a spectrum, where an equal amount of energy is located below and above the point. In contrast to the cited papers the spectral centroid was, in the present study, calculated on the basis of loudness spectra rather than frequency spectra, as it was hypothesised to be a better predictor, due to the closer relation with the perception of listeners. The proposed Dark-Bright metric was therefore calculated as described by Eq. 5. Since the output of the loudness model was in discrete 0.1 Bark bins, the solution became a minimization problem. $Dens_m(b)$ is the temporal mean of the time-varying specific loudness and b is the 0.1 Bark bin number. b_{MIN}, b_{CEN}, b_{MAX} are the minimum, centroid, and maximum bin numbers respectively. b_{CEN} thereby represents the point of equal loudness, i.e. the perceptual spectral centroid.

$$\min_{b_{CEN}\in\mathbb{Z}} \left| \sum_{b=b_{MIN}}^{b_{CEN}} Dens_m(b) - \sum_{b=b_{CEN}+1}^{b_{MAX}} Dens_m(b) \right| \quad \text{s.t.} \quad b_{MIN} \geq b_{CEN} \leq b_{MAX} \tag{5}$$

3.2 Metrics results

Numeric results of the optimization routine and bootstrap classification are shown in Table 3. The routine output's a map showing correlation values for all processed AB ranges. The first column shows the multiple (competing) peaks for each sensory descriptor, e.g. three for Bass strength. The **Bass strength** P1 and P2 AB ranges pointed to the same conclusion: that the low-frequency range up to $210\,Hz$ was important for the perception of bass strength, i.e. P1 and P2 could be considered as equivalent. The 'Peak R' column displays the Pearson correlation coefficient peaks from the baseline with the perceptual data. The third peak, P3, was related to treble, implying that the level of high-frequencies may affect the perception of bass strength. The 'Hit rate' column shows that 26.2% of the bootstrap iterations led to P1 or P2 having the best correlation with the sensory descriptor Bass strength, with a median correlation coefficient for P1 of $r = -0.79$ and 95% CI's of $[-0.57$ to $-0.99]$. For **Midrange strength**, the peak with the highest correlation had an AB range within bass and low-midrange frequencies (again P2 was equivalent), while P4 was the only peak with a maxima coinciding with the baseline peaks in 500 of the bootstrap iterations. The correlation CI's of these bootstrap maxima was however inconsistent and spanned both positive and negative values, implying instability. For **Treble strength**, P2-P4 all led to high hit rates, with P2 having the highest hit rate and a narrow CI. P5 had a low hit rate, but covered the area traditionally considered the treble range. For **Clean**, P4 and P5 resulted in a combined hit rate of 31.6%. The bootstrap CI's for P4 spanned the smallest range of values.

Metric	Peak R	AB Range	Hit rate (%)	Bootstrap R	95% CI's
Bass strength P1	-0.97	210-15000	25.8	-0.79	-0.99; -0.57
P2	0.96	20-210	0.4	0	
P3	-0.82	8900-14000	11.2	-0.82	-0.90; -0.74
Midrange strength P1	-0.98	20-610	0		
P2	0.98	640-15000	0		
P3	0.97	690-5900	0		
P4	0.91	8900-15000	6.2	0.80	**-0.99; 0.89**
Treble strength P1	-0.95	20-650	2.8	-0.97	-1.00; -0.95
P2	0.95	680-15000	29.6	0.97	0.95; 0.99
P3	0.95	730-5700	26.6	0.96	0.93; 1.00
P4	0.93	2500-4500	20.0	0.98	0.95; 0.99
P5	0.93	8700-15000	2.8	0.99	0.94; 1.00
Clean P1	0.99	800-4800	0		
P2	-0.98	20-720	0.8	-0.99	-1.00; -0.98
P3	0.98	730-15000	0		
P4	-0.90	20-8200	24.6	0.76	0.58; 0.97
P5	0.90	8200-15000	7.0	0.77	**-1.00; 0.92**

Table 3: Potential metrics describing sensory descriptors. The first column displays a metric, e.g. bass strength, and peaks (P1, P2, etc.) in the output map of the optimization routine (baseline). The two next columns display the Pearson correlation coefficient value of the baseline peaks and their AB ranges. The 'Hit rate' column displays percentage bootstrap iterations having maximum at the baseline peak. The last two columns display the median correlation of the iterations with a match as well as their 95% confidence intervals (CI) calculated from bootstrap percentiles. CI's in **bold** have a range spanning both positive and negative *r* values. All numbers are rounded to two significant digits.

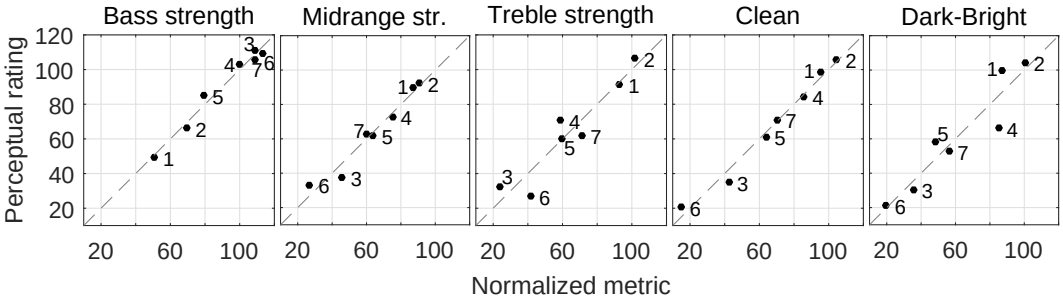

Fig. 2: Scatterplot of perceptual ratings vs. fitted metric values. The metric values where normalised to the scale of the perceptual ratings (0-150). The metric values are based on P1 for Bass strength, P3 for Midrange strength, P5 for Treble strength, and P4 for Clean respectively.

For the Dark-Bright sensory descriptor a bootstrap method was again employed to test the stability of the proposed metric. Pearson correlation coefficients for 500 bootstrap iterations had a median value of $r = 0.95$ and 95% CI's of [0.81;0.99].

The relation between the proposed metrics and the perceptual ratings are shown in Fig. 2.

4 Discussion

For the sensory descriptors modelled using Eq. 2, the output of the optimization routine led to multiple peaks with correlation coefficients $r^2 \geq 0.67$ ($r \geq |0.82|$), when all headphones were analysed. This approach was deemed appropriate for modelling of the four sensory descriptors presented in Table 3. In the case of Bass-, Midrange- and Treble strength, they all had a peak, which logically seemed probable: An AB-range of $20-200\,Hz$ (P2) for Bass strength, an AB-range of $690-5900\,Hz$ (P3) for Midrange strength, and an AB range of $8.7-15\,kHz$ (P5) for Treble strength. The bootstrap categorisation method showed the peak P1 (and the equivalent P2) to have a promising hit rate (26.2%), while Midrange strength, in contrast to logic, showed the highest hit rate for an AB-range in the high-frequency region. For Treble strength three peaks got high hit rates ($20.0-29.6\%$), none of which seem logically related to the perception of treble. In the case of Midrange- and Treble strength the bootstrapping process uncovered uncertainties in the data, but did not point to peaks likely to have a causal relation with perception. For Bass strength, the method showed that peak P1 may be a better predictor of bass strength, than the equivalent but more logical choice peak P2, due to numerical stability. For Clean peak P4 and the equivalent P5 got a combined hit rate of 31.6%, with P4 being more stable with regards to CI's. Here, the bootstrapping process revealed peak P4 as a potential better predictor of the descriptor Clean, than P1, although P1 had the largest r-value in the output from the optimization routine when all headphones were analysed (baseline). In general, the many equivalent peaks may points to a problem in structure of the metrics described by Eq. 2. For prediction of bass strength, a better metric could be ${}^{AB}/_{CD}$, i.e. with the denominator covering a limited frequency range CD rather than the full range. This approach was investigated in another study [18].

For Dark-bright the loudness spectral centroid correlated well with the perceptual data (median $r = 0.95$). Closer inspection of the relation between the metric's output and the perceptual ratings showed the auralized headphones '4' as a slight outlier for two of the four musical excerpts. Compared with the other headphones, these had an increased midrange within the frequency range $\approx 700-1400\,Hz$. This may indicate that wide resonances in the sound reproduction affect the perception of Dark-Bright slightly different than predicted by the proposed metric.

The three most promising metrics Bass strength, Dark-Bright and Clean models the sensory descriptors with the least correlation between them, thereby constituting a strong set of metrics for characterisation of the perceptual space spanned by the evaluated prototype headphones.

5 Summary

This paper presented a framework for modelling of sensory descriptors related to timbre of seven prototype headphones. Three metrics deemed stable was proposed for Bass strength ($r^2 = 0.62$), Clean ($r^2 = 0.58$), and Dark-Bright ($r^2 = 0.90$) respectively. All of them were based on loudness estimates of listening test stimuli. The first two were modelled from a simple equation (Eq. 2) with specific loudness in an AB frequency range (found by optimisation) divided by loudness in the full spectrum. The metric proposed for Dark-Bright prediction was based on a spectral centroid calculation of specific loudness. For two other sensory descriptors, Midrange- and Treble strength, stability investigations based on a bootstrapping process revealed inconsistencies in the sign of the correlations or multiple competing local maxima.

6 Acknowledgements

This work was funded by DELTA and the Danish Agency for Science, Technology and Innovation (Case number: 1355-00061). The author wishes to thank the audio design company for supplying the prototype headphones, Tore Stegenborg-Andersen for processing of the stimuli and for conducting the listening test, as well as the two anonymous reviewers for valuable comments and suggestions.

References

[1] Olive, S. E., "A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part II - Development of the Model," in *AES Convention 117*, 2004, Convention paper 6190.

[2] ITU-R, "Method for objective measurements of perceived audio quality," Recommendation ITU-R BS.1387-1, International Telecommunication Union Radiocommunication Assembly (ITU-R), United States, 1998.

[3] ITU-T, "Perceptual objective listening quality assessment," Recommendation ITU-T P.863, ITU Telecommunication Standardization Sector (ITU-T), United States, 2011.

[4] Takanen, M. and Lorho, G., "A Binaural Auditory Model for the Evaluation of Reproduced Stereophonic Sound," in *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*, pp. 1–10, 2012.

[5] Toole, F. E., "Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 2," *J. Audio Eng. Soc*, 34(5), pp. 323–348, 1986.

[6] McEwan, J. A., "Preference Mapping for Product Optimization," in *Multivariate Analysis of Data in Sensory Science*, volume 16 of *Data Handling in Science and Technology*, pp. 71–102, Elsevier, 1st edition, 1996, ISBN 978-0-444-89956-9.

[7] Blauert, J. and Jekosch, U., "A Layer Model of Sound Quality," *J. Audio Eng. Soc*, 60(1/2), pp. 4–12, 2012.

[8] Gabrielsson, A. and Sjögren, H., "Perceived sound quality of sound-reproducing systems," *J. Acoust. Soc. Am.*, 65(4), pp. 1019–1033, 1979, doi:10.1121/1.382579.

[9] Olive, S. and Welti, T., "The Relationship between Perception and Measurement of Headphone Sound Quality," in *Audio Engineering Society Convention 133*, 2012, Convention paper 8744.

[10] Lavandier, M., Herzog, P., and Meunier, S., "Comparative measurements of loudspeakers in a listening situation," *J. Acoust. Soc. Am.*, 123(1), pp. 77–87, 2008.

[11] Pedersen, T. H. and Zacharov, N., "The Development of a Sound Wheel for Reproduced Sound," in *Audio Engineering Society Convention 138*, pp. 1–13, Audio Engineering Society, Warsaw, Poland, 2015, Convention paper 9310.

[12] Zwicker, E. and Scharf, B., "A model of loudness summation." *Psychological Review*, 72(1), pp. 3–26, 1965, doi:10.1037/h0021703.

[13] Stone, H., Sidel, J., Oliver, S., Woolsey, A., and Singleton, R. C., "Sensory Evaluation by Quantitative Descriptive Analysis," *Food Technology*, 28, pp. 24–34, 1974.

[14] Lorho, G., Le Ray, G., and Zacharov, N., "eGauge—A Measure of Assessor Expertise in Audio Quality Evaluations," in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, pp. 1–10, 2010.

[15] Kroonenberg, P. M. and de Leeuw, J., "Principal component analysis of three-mode data by means of alternating least squares algorithms," *Psychometrika*, 45(1), pp. 69–97, 1980, doi:10.1007/BF02293599.

[16] Schubert, E. and Wolfe, J., "Does Timbral Brightness Scale with Frequency and Spectral Centroid?" *Acta Acustica united with Acustica*, 92, pp. 820–825, 2006.

[17] Labuschagne, I. B. and Hanekom, J. J., "Preparation of stimuli for timbre perception studies," *J. Acoust. Soc. Am.*, 134(3), pp. 2256–2267, 2013, doi:10.1121/1.4817877.

[18] Volk, C. P., Lavandier, M., Bech, S., and Christensen, F., "Identifying the dominating perceptual differences in headphone reproduction," *Submitted, J. Acoust. Soc. Am.*, 2016.