OPEN ACCESS Freely available online
Journal of the Audio Engineering Society
Vol. 65, No. 5, May 2017 (© 2017)
DOI: https://doi.org/10.17743/jaes.2017.0006

Modeling Perceptual Characteristics of Loudspeaker Reproduction in a Stereo Setup

CHRISTER P. VOLK,[1, 2] *AES Student Member*, SØREN BECH,[2, 3] *AES Fellow*, TORBEN H. PEDERSEN[1],
(cvo@delta.dk)
AND FLEMMING CHRISTENSEN[2]

[1]*DELTA SenseLab, Venlighedsvej 4, 2790 Hørsholm, Denmark*
[2]*Aalborg University, Department of Electronic Systems, Frederik Bajers Vej 7B, 9220 Aalborg, Denmark*
[3]*Bang & Olufsen, Peter Bangs Vej 15, 7600 Struer, Denmark*

In this study the characteristics of compact loudspeakers in a stereo setup positioned in a standardized listening room were investigated. Perceptual evaluations of eleven loudspeakers were conducted on the basis of six selected sensory descriptors, chosen by experienced listeners during consensus meetings. Based on an analysis of the perceptual evaluation data, four of the descriptors were found suited for modeling, with the purpose of developing metrics for prediction of Bass depth, Punch, Brilliance, and Dark-Bright respectively. Bass depth and Punch were modeled as one due to high correlation between the two. The experimental setup included loudspeaker spinners, enabling fast positioning of loudspeakers. The prediction models were based on binaural recordings, processed using a loudness model, and developed on the basis of previous work on headphone modeling [1, 2]. They were trained on a subset of the data (66%) and validated on the rest. The resulting metrics had high correlations with the perceptual ratings of the validation dataset ($r = 0.85$-0.96).

1 INTRODUCTION

Loudspeaker specifications have traditionally described the physical properties and characteristics of loudspeakers: frequency response, dimensions and volume of the cabinet, diameter of drivers, impedance, total harmonic distortion, sensitivity, etc. Few of these directly describe the sound reproduction and none directly describe perception of the reproduction, i.e., takes into account that the human auditory system is highly non-linear in terms of spectral-, temporal-, and sound level processing (see, e.g., [3]). This disconnect between specifications and perception have made it challenging for acousticians and engineers (and consumers) to predict how a loudspeaker will sound on the basis of these specifications.

Perceptual audio evaluations have long been a reliable method of characterizing the reproduction of loudspeakers, headphones, codecs, etc. The requirements for making reliable listening tests are, however, many, both in terms of facilities, equipment, handling of listeners, etc. (see, e.g., [4]). Additionally, numerous potential biases [5, 6] must be avoided in the listening test design, making the conduction of listening tests a task for experts only. One way of making perceptual characterization more accessible (and readily available) have been to develop metrics for predicting perception from various (more easily obtainable) physical measurements of the sound reproduction. The efforts can be divided in two categories: (1) hedonic predictions of, e.g., Basic Audio Quality [7], Mean Opinion Score [8–11], Preference [12], or spatial quality [13]; and (2) predictions of reproduction characteristics such as Punch [14], Width (sound image) and Bass tightness [15], Stereo image width [16], Discoloration, Treble stressing, General bass emphasis, Low bass emphasis, Brightness, Bass clearness, and Feeling of space [17] and Brightness [18, 19].

While earlier studies focused on making predictions on the basis of the aforementioned specifications (e.g., frequency responses in [20, 21]), more recent modeling efforts have relied more on measurements closer related to the human hearing, e.g., by using binaural recordings as a representation of the physical domain (see, e.g., [15, 16, 22]) and by processing the modeling input using auditory models (see, e.g., [14, 15, 17, 16, 22]). In [22] by Beerends et al., this approach was used successfully as a means to obtain predictions of the subjective overall sound quality of loudspeakers in a stereo setup (average correlation of $r = 0.85$). The recording-based modeling approach (also used in the present study) eliminates the need for technical measures [22]: “This approach thus does not need any technical measurements on the loudspeakers, it only uses recordings of musical fragments played over the loudspeakers.” Furthermore, [22] introduces an approach to mimic the internal reference of listeners that they use to evaluate the sound quality. The approach is based on making recordings of a high-end loudspeaker ("best available"), with a head-and-torso-simulator positioned ideally. In the present paper another approach was taken, which is discussed in detail in Sec. 5 and compared to a third approach by Klippel [17].

In the present study the sensory descriptors describing the dominating perceptual differences between compact loudspeakers in a stereo setup were found by consensus meetings with experienced listeners, and predictive models[1] are designed on the basis of listening tests on loudspeakers in a listening room and analysis of binaural recordings made in the listening position. These tests included evaluations of five sensory descriptors[2], representing identified differences on 11 stereo sets of loudspeakers. The loudspeakers were placed in two positions: eight on loudspeaker spinners and three in corner positions. The loudspeaker spinners allowed evaluations of loudspeakers in identical positions with a minimum of switching time, i.e., strain on the limited auditory memory of humans (see review in [24]).

The present study presents a modeling methodology based on binaural recordings being processed using a loudness model. This methodology has been tested for modeling of headphones (sound reproduction without room influences) in a previous study [2]. The present study thereby tests both the suitability of using the proposed methodology for modeling of loudspeakers in a stereo setup and tests the modeling strategy on a different set of sensory descriptors than previously investigated.

2 LOUDSPEAKERS IN A STEREO-SETUP

The listening test comprised two sessions; each with evaluation of seven stereo sets of loudspeakers, of which three sets were in both sessions. The test consisted of reproductions of two musical excerpts evaluated on six sensory descriptors and rated twice by each listener. One session thereby consisted of 168 ratings and had a duration of no more than two hours including breaks, which listeners were encouraged to take whenever needed. The test software automatically and regularly reminded listeners to take these breaks. One "screen" in the test software consisted of evaluation of each of the seven sets of loudspeakers for one sensory descriptor with one musical excerpt, e.g., bass depth. A "screen" had seven horizontal rating scales, representing each loudspeaker set, presented in a randomized order. The experimental design within one session was a block design with each block consisting of one repetition. Within a block the musical excerpts and sensory descriptors were presented in a randomized order as well. Listeners started both sessions with a familiarization part that included presentation of all stimuli. In this part they were allowed to make small adjustments to the overall sound level and instructed to keep that level for the main test.

In the following subsections the details of the setup, the loudspeakers, the stimuli, and the listeners are presented.

2.1 Stereo-Setup

The listening test was conducted in a listening room compliant with the ITU-R BS.1116-3 [25] recommendation. The loudspeakers were evaluated in the stereo setup depicted in Fig. 1 (not to scale). Four sets of loudspeakers (spot 1–4) were secured on loudspeaker spinners (DELTA Low Noise Rapid Speaker Spinners), which could move a requested loudspeaker set into the ideal position of the equilateral triangle in about a second no matter the previous position. The figure shows two situations:

Scenario 1 (left) A set of loudspeakers (1–4) on the loudspeaker spinners are playing after being moved into the ideal positions of the equilateral triangle (playback positions).

Scenario 2 (right) A set of loudspeakers in the corners (C1-C3) are playing and the loudspeakers on the spinners are moved to other positions.

Note that the two spinners were always in mirrored positions of each other (not depicted) with two loudspeakers playing in stereo. The loudspeakers on spinners were individually positioned to point towards the listening position (when in the playback positions) and with their acoustical center, as specified by the manufacturers, at 110 ± 0.5 cm above the floor (approximately the height from the floor to the ear canal entrance of an average seated listener). The center of each loudspeaker spinner was positioned 0.85 m from the side wall and 1.05 m from the back wall. They allowed four sets of loudspeakers to be correctly positioned in an ideal stereo setup (an equilateral triangle) when evaluated by the listeners. Additionally, they were programmed to rotate the least possible (left or right) when moving loudspeakers into the playback positions to minimize switching time. Three additional sets of loudspeakers (C1–C3) were positioned in the corners of the room. These were included in the test design as special sound sources with desired features to provide low-, mid-, and high anchors, which could stabiize the scale usage across the two listening sessions and increase the range of perceptible differences. A set of Genelec 8020C (C1 positions) were stacked on top of a set of "SenseLab Low Anchor" (SLA) custom-built loudspeakers (C2 positions) with Genelec 8050A positioned beside the two (C3 positions). Their acoustic centers were at a height of 145, 122.5, and 133 cm respectively, i.e., higher than the loudspeakers on the spinners. The set of loudspeakers in the corners were programmed to have individual virtual positions on the loudspeaker spinners. This had two purposes: (1) it rotated the spinners to a position as shown in Fig. 1 on the right, where the sound emitted was the least obstructed by the loudspeakers on the spinners, and (2) it gave listeners the impression that all loudspeakers were placed on the

[1] In this paper the term "metric" is used to describe the end result of the modeling efforts, while "prediction model" is used to refer to the development stages of a "metric."

[2] 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 loudspeaker reproduction. This definition is adapted from [23].

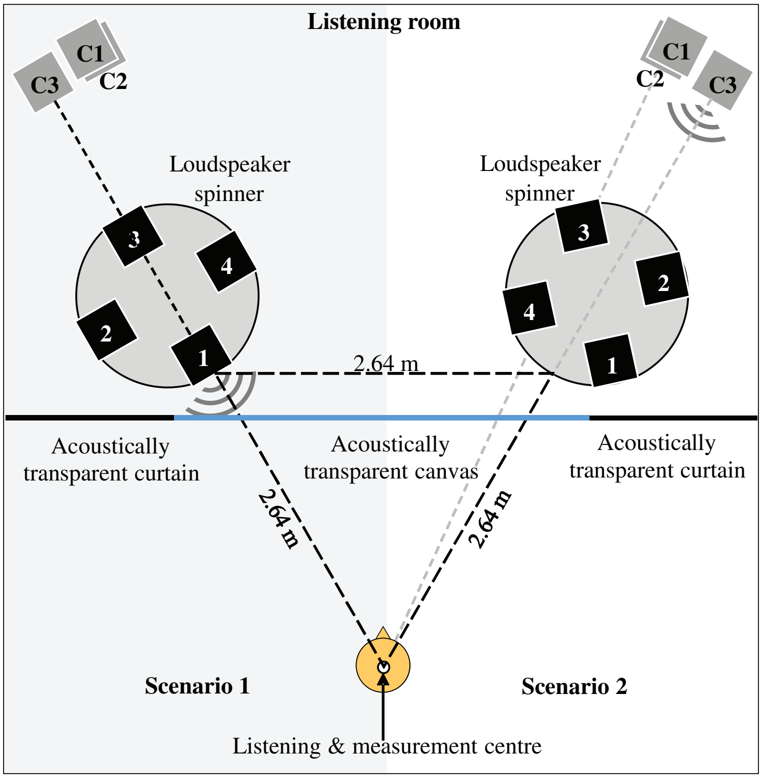
Listening room
C1
C2
C3
C1
C2
C3
Loudspeaker spinner
Loudspeaker spinner
1
2
3
4
1
2
3
4
2.64 m
2.64 m
2.64 m
Acoustically transparent curtain
Acoustically transparent canvas
Acoustically transparent curtain
Scenario 1
Scenario 2
Listening & measurement centre

Fig. 1. Experimental setup. Four sets of loudspeakers were positioned in a standard stereo setup (equilateral triangle) and three other sets in the corners of the room. The loudspeaker spinners move a set of loudspeakers (selected in the test interface) into the playback positions prior to stimuli presentation (Scenario 1). If a set of loudspeakers positioned in the corners (C1–C3) were selected, the spinners instead moved to a position with less influence from the loudspeakers on the spinners (Scenario 2). *Note: The diagram is not to scale.*

spinners (important to reduce system identification, which can lead to listener expectation bias [6])

The listener was seated in a chair positioned in the center of the width-dimension in the room and view of the loudspeakers were blocked by two layers of thin curtains (sides) and an acoustically transparent canvas (center, damping <1 dB below 16 kHz at an 30° incident angle) displaying the test interfaces.

2.2 Loudspeakers and Calibration

A perceptual evaluation was made of 11 models of compact loudspeakers. Eight were chosen as representative of loudspeakers in the consumer segment (price range of 60–554 USD, median of 329 USD) and three were loudspeakers used as anchors as described in the previous section. Except for the custom-made loudspeaker (SLA), all loudspeakers had two drivers (tweeter and midrange). The volume of the loudspeaker cabinets were in the range 3–21 l (median 10.4 l), with the exception of the large Genelec 8050A (36 l). The in-situ frequency response of the eight ideally-positioned loudspeakers are depicted in Fig. 2. The measurements stem from recordings at the ear-entrance-point (EEP) of a B&K 4100 head-and-torso simulator with a blocked ear canal using a 35-second pink noise test signal.

The loudspeakers were evaluated in two separate listening sessions to accommodate the space limitations on the loudspeaker spinners. For each session four models were paired to span a wide range of differences between products, i.e., by mixing brands, sizes, and price ranges. In the following a *loudspeaker set* refers to two identical loudspeakers used for stereo reproductions.

Eight of the 11 loudspeaker sets were positioned ideally, while three were positioned differently and included in both tests to obtain similar scale usage across the two sessions (as previously discussed). To reduce the influence of the corner positions the two Genelec loudspeakers (intended as mid- and high anchors) had their frequency response adjusted using 1/3 octave-band filters to be flat within ±3.5 dB in the range 80 Hz–8 kHz for the Genelec 8020C (mid anchor) and 31.5 Hz–16 kHz for the Genelec 8050A (high anchor). The resulting frequency responses were measured with a single microphone in the listening position and with four loudspeaker-sets on the loudspeaker spinners. The Genelec 8020C additionally had 1/3-octave bands above 10 kHz damped 12 dB to differentiate it from the larger 8050A with regards to both the low- and high frequency extension. The set of SLA loudspeakers were included in the test to expand the range of perceivable characteristics downwards and were thus not equalized. To slightly reduce the difference to the loudspeakers on the spinners, two strong resonances at 500 Hz and 800 Hz were, however, dampened 6 dB (using 1/3-octave equalizers). The frequency responses of these three special-positioned loudspeakers are depicted in Fig. 3. The

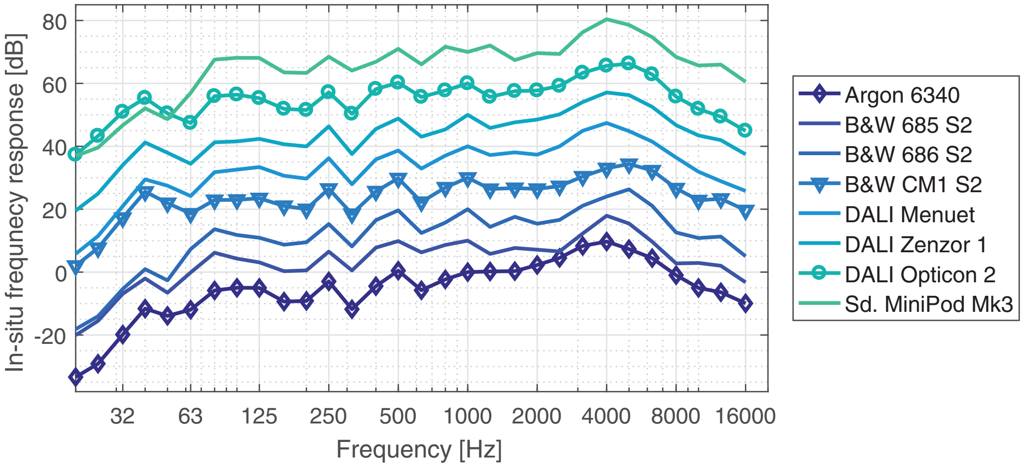

Fig. 2. In-situ frequency response of the eight ideally-positioned loudspeakers measured at the ear-entrance-point of a B&K 4100 head-and-torso-simulator. The responses are plotted in the same order as listed in the legend (highlighted by symbols on three of the curves). The lowest response is normalized to have 0 dB at 1 kHz and the others are offset by multiples of 10 dB.

frequency responses of the two equalized loudspeaker sets were not completely identical between sessions as depicted in Fig. 3. Part of the reason being that different loudspeakers were positioned in front of them on the speaker spinners. In addition, the cut-off frequency for the Genelec 8020C seem to have shifted by 1/3-octave band towards the higher frequencies. They nevertheless received ratings without significant differences between sessions (also indicating that there was no significant session effect in the experimental design).

All loudspeakers were level calibrated to produce 70 ± 0.5 dB(A) in the listening position (measured with a single measurement microphone). The calibration signal had a pink noise spectrum and was band-pass filtered to a frequency range of 80 Hz–14 kHz. After the calibration two of the authors and a colleague checked that no perceptual level differences were noticeable for the chosen stimuli.

2.3 Stimuli and Sensory Descriptors

Two musical excerpts were chosen for reproduction over the loudspeakers: a 15-second soft pop excerpt (“Bird on a Wire” by Jennifer Warnes) and a 24-second oriental excerpt (“Moonlight on Spring River” by Zhao Cong). Both excerpts were cut to maintain the rhythm during looping. Frequency content of the two excerpts are shown in Fig. 4. The Jennifer Warnes excerpt is dominated by a female vocal and a drum beat but also includes a variety of other instruments. The frequency content is smooth in a wide range. The Zhao Cong excerpt is a calm instrumental composition dominated by very deep bass drums and a melody played on pipa (Chinese “lute”). The mix includes many additional instruments as well. The frequency content is broad but with a lower level in the high bass/low midrange. These samples were selected on the basis of the authors’ subjective perception of clearly separable sources in the

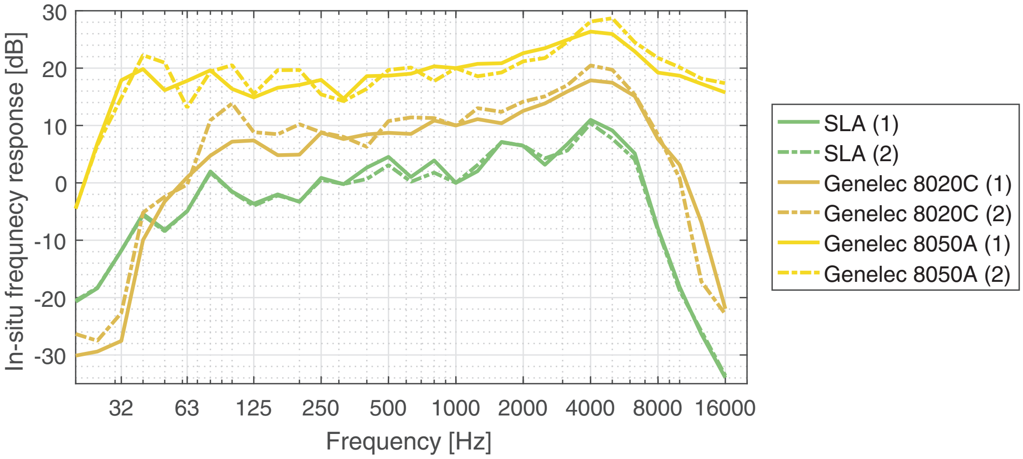

Fig. 3. In-situ frequency response of the three corner loudspeakers measured at the ear-entrance-point of a B&K 4100 head-and-torso-simulator. The responses are plotted in the same order as listed in the legend. Measurements from both sessions are depicted (with session 2 measurements dash-dotted) and denoted in the legend by the number in parentheses. The lowest response is normalized to have 0 dB at 1 kHz and the others are offset by multiples of 10 dB.

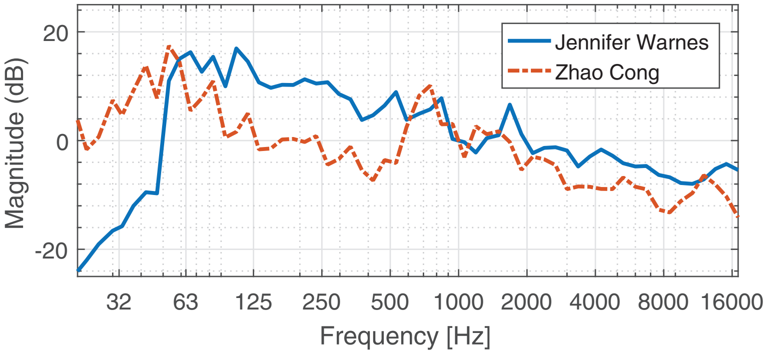

Fig. 4. Frequency content, L_{EQ}, of the raw musical excerpts depicted in 1/6-octave bands. Normalized to 0 dB at 1 kHz.

stereo image and a sense of clarity (as defined in the Sound wheel [23]) in the recordings.

Reproduction of the two excerpts were evaluated by listeners on six perceptual characteristics defined by the following sensory descriptors: (1) Punch, (2) Bass depth, (3) Brilliance, (4) Dark-Bright, (5) Natural, and (6) Spatial precision. The descriptors were all from a Sound wheel for audio reproduction [23] that each have a definition as well as a low- and high verbal scale anchor, e.g., “a little” and “a lot” for Brilliance. Note that Punch is defined differently in the Sound wheel compared to that of [14], where it is referred to as something that: *“characterize music or sound sources that convey a sense of dynamic power or weight to the listener.”* Their idea of Punch seemed to have more in common with the descriptors referred to as Bass precision and Attack in the Sound wheel [23]. Punch as defined in the Sound wheel [23] is *“ability to effortlessly handle large volume excursions without compression.”*

2.4 Listeners

Ten listeners participated in the listening test. They were all experienced and trained listeners with normal hearing and ranged in age from 20 to 46 with a median of 30 years. Nine of the 10 were trained specially in perceptual loudspeaker evaluation. Their performance was evaluated using a combination of eGauge [26] and Tucker-1 plots. This performance evaluation was described in detail in [2] and includes criteria for removing listeners performing below specified requirements with regards to discrimination and reproducibility.

3 PERCEPTUAL MODELING AND RESULTS

3.1 Data Basis for Perceptual Modeling

Out of the six evaluated sensory descriptors, listeners were not able to discriminate between the loudspeakers for Spatial precision and Natural, i.e., none of the loudspeakers on the loudspeaker spinners were rated significantly different from each other on an $\alpha = 0.05$ level. Furthermore, Bass depth and Punch were highly correlated ($r^2 = 0.85$). Consequently, only “BassPunch” (treating Bass depth and Punch as replicates of the same descriptor), Brilliance, and Dark-Bright were modelled. For each of these metrics are proposed on the basis of listening test data and corresponding binaural recordings made in the listening position.

The recordings captured the two musical excerpts when reproduced over the loudspeakers. A Brüel & Kjær 4100 head- and torso simulator without ear canals was placed in a chair in the listening position with the microphones centered in a height of 110 ± 0.5 cm above the floor.

The dataset of perceptual ratings and corresponding recordings was split up into a training- and validation set. The selection of loudspeakers for each set was chosen separately for each sensory descriptor following this strategy: A first step was to discard perceptual data of the loudspeakers in the corners that had the largest confidence intervals, i.e., session 1 or session 2 data of the same loudspeaker set. This was needed as the loudspeakers were too similar to treat as separate data points and would have led to overly optimistic evaluation of the prediction models. A second step was to sort the perceptual data per loudspeaker set in ascending order of mean rating and select loudspeakers ranked 2, 5, 8, and 10 for validation (≈36 %) and the rest for training. This selecting scheme ensured that validation ratings were within the range spanned by the training ratings (of which the scope of the prediction models are limited). Additionally, it ensured a wide spread of ratings in both the training and the validation set[3].

3.2 Modeling Methodology: BassPunch and Brilliance

Bass depth and Brilliance were defined to describe similar concepts: the bass extension and the treble extension respectively. From a perceptual viewpoint the strongest cue in identifying the bass- or treble extension is loudness at the lowest or highest range of frequencies. Punch is considered related to both the spectral and the temporal response of loudspeakers, with the temporal response determined by its time constant (also referred to as onset or rise time). In the Sound wheel definitions [23], Punch is described as related

[3] Note that division of datasets into training and validation subsets is normally done using random draw, i.e., randomly assigning data to one or the other subset, which minimizes the risk of biased/boosted result. With a small dataset this approach isn’t suitable as the random subsets risk only spanning a small fraction of the rating range.

to the reproduction of bass and drums and defined as *"ability to effortlessly handle large volume excursions without compression,"* i.e., to fully reproduce the (relative) level of the low-frequency content. A high Punch rating thus requires good deep bass reproduction, besides the assumed good temporal characteristics, and some level of correlation between Punch and Bass Depth is therefore expected. Since it was very high is this study, spectral characteristics was assumed to be the dominating aspect of Punch here. Consequently, both the combined descriptor BassPunch and Brilliance were modeled using a generic methodology based on spectral characteristics, proposed in two previous papers for use with perceptual modeling of headphone differences [1] and characteristics [2].

The methodology is based on specific loudness estimations of binaural recordings, here calculated using the time-varying model by Glasberg and Moore [27]. Briefly described, the loudness model corrects for outer- and middle ear influences, calculates the excitation pattern of the basilar membrane, and estimates the specific instantaneous loudness for each millisecond in a frequency resolution of 0.25 equivalent rectangular bands (ERBs). In a final step in the original model short- and long term loudness is estimated (taking temporal masking into account). This processing step was, however, not relevant for this purpose as specific loudness was of interest, i.e., averaging over time instead of frequency.

Prediction models were trained using an optimization routine (also described in [1, 2]) that optimized the variables of an equation on the form described by Eq. (1), such that *metric* correlates the most with the ratings of the sensory descriptors. $Dens_m(f)$ is the temporal mean of the instantaneous specific loudness, while A and B denotes the frequency limits of an AB range. The optimization routine searches for the optimum AB range in steps of 0.25 ERBs for A and B independently, but limited to a minimum AB range of 2 ERBs. This limitation was added to reduce the risk of finding spurious high correlations in narrow AB ranges, unlikely to have significantly affected perception and rating of any of the sensory descriptors.

$$metric = \frac{\text{AB range}}{\text{Full range}} = \frac{\sum_{f=A}^{B} Dens_m(f)}{\sum Dens_m} \tag{1}$$

In [1], where the methodology was first described, an additional equation was suggested, which had a limited range in the denominator "CD" as well, as opposed to the full-range of Eq. (1). This equation was also tested in the present study for modeling BassPunch and Brilliance but did not lead to as high correlations as the simpler equation in Eq. (1), and results are consequently not reported.

Additionally, the search ranges (investigated AB ranges) were limited to sensible ranges in relation to the general meanings/definitions of bass and treble, namely 20–500 Hz for the BassPunch prediction model and 6.0–14.7 kHz for the Brilliance (14.7 kHz being the highest center frequency of the loudness model output).

3.3 Modeling Methodology: Dark-Bright

In a previous study [2] we described a metric for prediction of Dark-Bright ratings. This metric was based on finding the spectral centroid of the stimuli. While this had been done previously for a descriptor referred to as "Brightness" (similar in description to Dark-Bright) the novelty was to base the metric on specific loudness estimates instead of frequency content. The metric thereby constitutes the center frequency at which the loudness in the low- and high frequencies are equal (or in practise have minimum difference). Eq. (2) describes the solution to the minimization problem of finding the perceptual centroid[4]. $Dens_m(f)$ is again the temporal mean of the instantaneous specific loudness and f is the frequency. f_{MIN}, f_{CEN}, f_{MAX} are the minimum, centroid, and maximum center frequencies respectively. f_{CEN} thereby represents the point of equal loudness, i.e., the perceptual spectral centroid.

$$\begin{aligned} &f_{CEN}: \\ &\min_{f_{CEN}\in\mathbb{Z}} \left| \sum_{f=f_{MIN}}^{f_{CEN}} Dens_m(f) - \sum_{f=f_{CEN}+1}^{f_{MAX}} Dens_m(f) \right| \\ &\text{subject to} \\ &\qquad f_{MIN} \le f_{CEN} \le f_{MAX} \end{aligned} \tag{2}$$

As it was hypothesized that the loudness of the mid-frequencies may not influence the perception of the Dark-Bright balance to the same extent as the loudness in the bass- or treble frequency ranges, an alternative prediction model is proposed here. Loudness in the midrange frequencies—defined here to be the range 400–4000 Hz—was reduced in steps of one percent point from $p = 100\,\%$ to $p = 0\,\%$ of the original loudness level $Dens_m(f)$ to investigate the effect on the correlation level with the ratings of Dark-Bright. The optimum value of p, leading to the highest correlation with the perceptual data, was found on the training data and tested on the validation data. Note, that this alternative can be viewed as applying a weighted upside-down rectangular window to the specific loudness spectrum, which is unlikely to occur in the human auditory processing. This is, however, a method of testing whether the hypothesis of a weighting function might be part of listeners auditory processing when evaluating spectral balance. Eq. (2) can be reused for this alternative approach simply by replacing $Dens_m$ with $Dens_{mw}$, defined in Eq. (3). Results for the two proposals are reported in Sec. 4.

$$\begin{aligned} &Dens_{mw}(f) = Dens_m(f) \cdot w(f) \\ &\text{where } w(f) = \begin{cases} p & \text{for } 0.4 < f < 4.0\,kHz \\ 1 & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

3.4 Modeling Methodology: Logistic Transformation

In an effort to obtain models with meaningful predictions in the entire rating interval, i.e., outside the interval

[4] Note that the frequency resolution is 0.25 ERBs and that the total number of frequency bins, 153, was uneven.

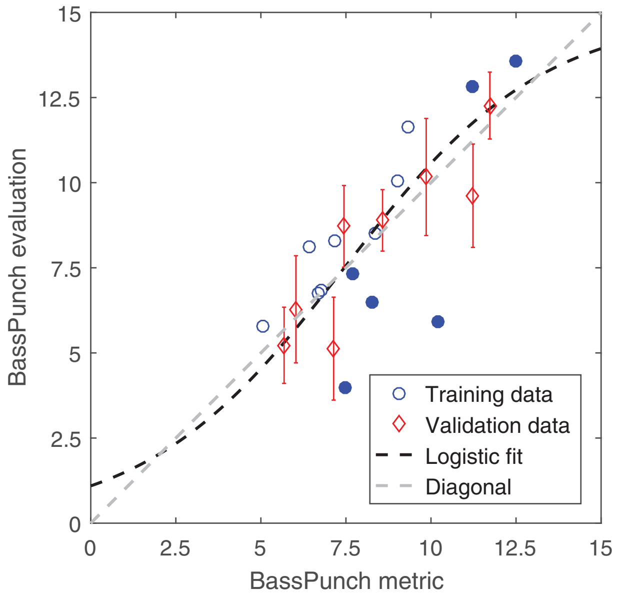

Fig. 5. Perceptual ratings of BassPunch vs. the proposed prediction model (prior to logistic transformation). Each set of loudspeakers is represented by two data points—one for each musical excerpt. The filled symbols represent loudspeakers from corner positions. The vertical bars represent the 95 %-confidence intervals.

of the currently collected data, all prediction models presented so far were transformed using a logistic (s-curve) fit. This ensures that the prediction models can never be outside the range of the scale in the listening test, i.e., 0–15. A logistic transformation ensures a saturation of the prediction value at the lowest and highest end of the scale. The transformation was done in five steps. First, the output of the prediction models was standardized. This was done separately for each of the two musical excerpts to remove excerpt-specific shift and scaling effects. Second, a linear fit was used to convert the output to the original rating scale (0–15). This was needed because the third step required strictly positive values. Third, a logit transformation was applied (Eq. (4)). This step transforms the data such that the output and the perceptual data have an approximately linear relationship. Fourth, a linear fit was found for the logit transformed data. Finally, the linear fit coefficients, $c1$ and $c2$ from step four, were transformed back to the original scale using the logistic transformation (the inverse of the logit transform) in Eq. (5). The two linear fits (steps 2 and 4) were made with perceptual ratings in the training subset and the coefficients were used for both the training and the validation subsets. The results prior to the final step of logistic fitting are depicted in Fig. 5 to Fig. 7 with the logistic transformation curve (step 5) plotted.

$$x_2 = \log\left(\frac{x}{15 - x}\right) \tag{4}$$

$$x_3 = \frac{15}{1 + e^{-c1 \cdot x_2 - c2}} \tag{5}$$

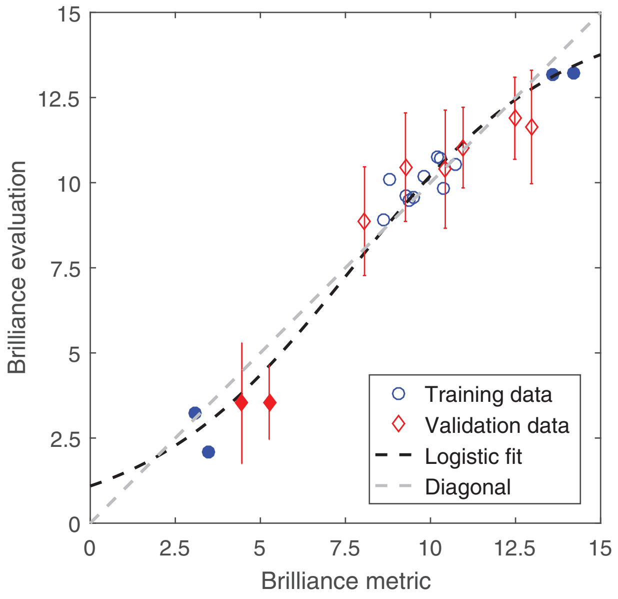

Fig. 6. Same as Fig. 5 but for Brilliance.

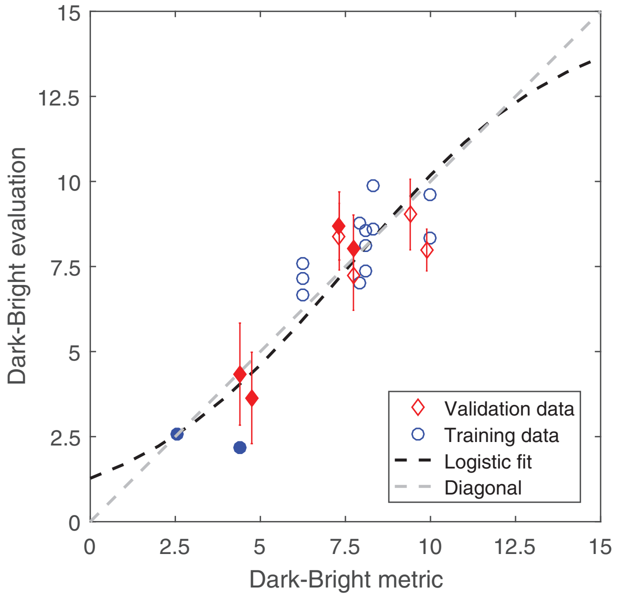

Fig. 7. Same as Fig. 5 but for Dark-Bright (R).

4 MODELING RESULTS

The modeling led to metrics for prediction of the sensory descriptors BassPunch, Brilliance, and Dark-Bright respectively. The performance of these are presented in Table 1 with parameters specified in the Details column. The numbers presented are the Pearson correlation coefficients, r, of the logistic transformed metrics and the AB-ranges are described by their 0.25 ERB center frequencies (f_c). Scatterplots are depicted in Fig. 5 to Fig. 7. Note, that for clarity only confidence intervals (CIs) for the validation data set are depicted. The CIs for the training data are similar in size (as both are based on ratings by the same number of listeners).

The big difference between the training and validation coefficients for BassPunch are caused by two sets of

Table 1. Performance of metrics describing sensory descriptor ratings. Training (Train.) and Validation (Val.) values are Pearson correlation coefficients, r. RMSE is the root-mean-square error (on the scale from 0–15). For the Dark-Bright metrics (R) denotes the alternative version with a weighted upside-down rectangular window.

Metric	Train.	Val.	RMSE	Details
BassPunch	0.70	0.90	1.06	AB: 20–72 *Hz*
Brilliance	0.99	0.96	1.00	AB: 8.3–10 *kHz*
Dark-Bright	0.61	0.17	2.19	
Dark-Bright (R)	0.88	0.85	1.08	$p = 7\%$

loudspeakers being outliers. The fit between perceptual ratings and prediction model prior to logistic transformation is depicted in Fig. 5. The two outliers (lowest filled circles) are SLA and Genelec 8020C from the corner positions evaluated on the same musical excerpt (Jennifer Warnes). The low correlation coefficient for Dark-Bright (without window) was also caused by two of the loudspeakers in the corners being outliers. The proposed alternative Dark-Bright metric with a weighted upside-down rectangular window, Dark-Bright (R), led to better correlations than the original Dark-Bright metric for both the training- and validation sets.

5 DISCUSSION

In general, the three proposed metrics—BassPunch, Brilliance, and Dark-Bright (R)—performed well having correlations coefficients $r \geq 0.85$ for the validation data sets and root-mean-square errors of $RMSE \approx 1$ (on the scale from 0–15). Furthermore, the AB-range for BassPunch seems intuitively reasonable, while for the Brilliance metric the 10 kHz upper limit seems low and cannot be explained by lack of frequency content in the musical excerpts. The BassPunch metric, however, had a lower correlation coefficient for the training set ($r = 0.70$) than the validation set ($r = 0.90$), and consequently $r = 0.70$ may be the most realistic estimate of its prediction performance level. The outliers in the validation training set are likely to be a consequence of the lack of spectral content at the lowest frequencies (<50 Hz) of the Jennifer Warnes excerpt (see Fig. 4), which constitutes a significant part of the metric's AB range (20–70 Hz).

Due to the understanding of Punch as a characteristic linked to temporal as well as spectral properties of the reproduction, several alternative modeling schemes were tried in an effort to include this aspect in either a separate Punch metric or in a combined BassPunch metric, but none led to consistent predictions. Examples of efforts included modifications of Eq. (1), where $Dens_m$ was replaced by either the temporal *maximum* of the instantaneous specific loudness, $Dens_{max}$ or the mean of an upper percentile of the instantaneous specific loudness, $Dens_{mp}$.

The higher performance of the Dark-Bright (R) metric in comparison to the simpler Dark-Bright metric proposed in [2] suggests that listeners put less emphasis on the midrange frequencies when evaluating the spectral balance of sound reproduction. It is, however, important to reiterate that the proposed weighting function is unlikely to correspond to that of a listener, as it has two strong discontinuity points at the start- and end frequencies of the function. A smoother function is expected to better represent this step in the auditory processing. Additional research is also required to establish the best-fit frequency limits as the current range, 400–4000 Hz, were chosen only as an initial estimate. Furthermore, it is of interest to establish whether these limits are similar for all listeners or whether clusters exist. Investigation of these improvements are planned to be the subject of a future study.

In terms of performance of the proposed metrics, it was of further interest whether all confidence intervals of the validation data points overlap the curve of the logistic fit in Fig. 5 to Fig. 7, in which case the best possible fit is reached within the uncertainty of the data and more data would be needed to verify further improvements. The Brilliance metric reached this prediction performance level, while small improvements are still possible within the statistical uncertainties of the current data set for both the BassPunch and the Dark-Bright metrics. Before further optimization is done it might, however, be of more value to validate the metrics with more data: more loudspeakers, different musical stimuli or different (higher) rooms with different reverberation times.

In comparison with previous modeling effort in the literature, one important difference in this study is the definition of the listeners "internal reference." The traditional view is, for instance, seen in [17] where Klippel proposed seven metrics for describing loudspeaker performance. The basis of his metrics was a calculation of "discoloration," which were defined as stated in Eq. (6), where $N'_{test}(z)$ and $N'_{ref}(z)$ are the specific loudness of the test- and reference stimuli respectively.

$$N'(z) = N'_{test}(z) - N'_{ref}(z) \tag{6}$$

Eq. (6) implies that the listeners know the recorded reference and are able to use this as an "internal reference" for assessment of loudspeakers by the deviations from this reference. The weakness here is that the listener does not know the recorded reference, as it cannot be presented to the listeners without being affected by the reproduction system. This approach is also used in prediction models involving codecs, e.g., P.863 POLQA [10, 11] and QESTRAL [13], but here the discoloration of the reproduction system is included in both the reference and the compressed systems under test. In [22], the approach was to include a "Best available" high-end stereo set of loudspeaker as a modeled "internal reference." This approach is similar to that by Klippel [17], assuming that the loudspeaker setup is reproducing the stimuli approximately ideally but with the added influence of the room. Adding the room influence likely improves the correspondence between a prediction model and a listener's perception. However, depending on how well the stimuli is known by the listeners and how degraded the reproduction of the loudspeakers being evaluated are, the listeners might not be able to extract a near-perfect reference from the presented stimuli. For example: How would a listener form an internal reference of the deep bass of a musical excerpt, if the loudspeakers under evaluation are tiny PC speakers. In the present study the perceptual sound reproduction characteristics were defined as: *"The perceived changes to the envisioned original sound."* So we assumed that the listener creates an internal reference of the original sound on the basis of what is heard and assess the loudspeaker characteristics as the deviation from this reference. The weakness here is that the internal reference dependent of both the characteristics of the musical excerpts and the loudspeakers under evaluation. We try to overcome this weakness by letting the assessors listen to all systems with different musical excerpts (familiarization, see Sec. 2) before the listening test, such that the internal reference should be an average over excerpts and thereby be a tool for assessing the loudspeakers with limited influence of the specific excerpts. In the processing of data, the internal reference in the present study was approximated by averaging over stimuli available in the training set and used for standardization as described in Sec. 3.4 (step 1).

Besides the different definitions of listener reference, the study by Klippel [17] showed many similarities supporting the findings of the current study. His metrics were based on seven sensory descriptors identified by comparisons of loudspeakers using a combination of ratio- and multidimensional scaling methods. They were analyzed using factor analysis and thereby comprise a list of dominating perceptual differences between the sound reproduction of loudspeakers. Four of these are similar to the set used in the present study, i.e., (1) Treble Stressing ≈ Brilliance, (2) Low bass emphasis ≈ Bass depth/Punch, (3) Brightness ≈ Dark-Bright, and (4) Feeling of space ≈ Spatial precision. Note, however, that Klippel's Treble Stressing was linked to the perception of sharpness or shrillness, where Brilliance is defined as treble extension. Klippel's proposed metric Low bass emphasis describes the ratio between the discolouration below $f_c = 60$ Hz and all critical bands above, with discoloration defined as spectral deviation from the original stimulus (discussed above). This is comparable to the AB-range found for BassPunch of 20–72 Hz (see Table 1).

Klippel's proposed Brightness (Dark-bright) metric is shown in Eq. (7), where S is Treble stressing and B is General bass emphasis.

$$H = 0.7S - 0.3B \tag{7}$$

B is calculated from the same equation as Low bass emphasis but with a pivot point at $f_c = 150$ Hz. S is based on discoloration as well, but multiplied by a weighting function increasing at the higher frequencies. Consequently, Klippel's Brightness metric puts higher emphasis on bass and treble than on midrange frequencies as well but additionally puts higher weight on treble than bass, which may be a consequence of a low pivot point at 150 Hz, which does not encompass the full bass frequency range.

6 SUMMARY

In this study three metrics were developed for prediction of the perceived characteristics of loudspeakers' sound reproduction in a stereo setup evaluated in a standardized listening room with regard to BassPunch, Brilliance, and Dark-Bright. The metrics were developed with the intention of finding specifications of loudspeakers' sound reproduction with perceptual relevance. They were based on binaural recordings made in the same setup as was used for perceptual evaluations of 11 stereo sets of loudspeakers. The recordings, made using a head- and torso simulator, were processed using a loudness model and led to metrics describing spectral characteristics of the reproduction. Two were based on the relative specific loudness of a limiting frequency range (AB) and one was based on a weighted specific loudness centroid. The prediction models were trained on a training subset with seven sets of loudspeakers and validated on four others. The range of correlation coefficients were $r = 0.85$–0.96 (details in Table 1, page 24). All metrics thus showed potential for prediction of a comparable loudspeaker segment and with a root-mean-square-error of $RMSE \approx 1$ on a 0–15 rating scale for the validation set. This RMSE level was largely comparable to the statistical 95 % confidence intervals of the perceptual evaluations.

7 ACKNOWLEDGMENTS

This work was funded by DELTA and the Danish Agency for Science, Technology and Innovation (Case number: 1355-00061). The authors wish to thank Hi-Fi Klubben for making eight of the compact loudspeakers available for this study as well as Tore Stegenborg-Andersen for assistance in conducting the listening test.

8 REFERENCES

[1] C. P. Volk, M. Lavandier, S. Bech, and F. Christensen, "Identifying the Dominating Perceptual Differences in Headphone Reproduction," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3664–3674 (2016). https://dx.doi.org/10.1121/1.4967225

[2] C. P. Volk, T. H. Pedersen, S. Bech, and F. Christensen, "Modelling Perceptual Characteristics of Prototype Headphones," presented at the *2016 AES International Conference on Headphone Technology* (2016 Aug.), conference paper 5-2.

[3] C. J. Plack, *The Sense of Hearing* (Mahwah, N.J: Lawrence Erlbaum Associates, 2005).

[4] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application* (Chichester, England; Hoboken, NJ: John Wiley & Sons, 2006).

[5] S. Zielinski, "On Some Biases Encountered in Modern Audio Quality Listening Tests (Part 2): Selected Graphical Examples and Discussion," *J. Audio Eng. Soc*, vol. 64, pp. 55–74 (2016 Jan./Feb.). https://dx.doi.org/10.17743/jaes.2015.0094

[6] S. Zielinski, F. Rumsey, and S. Bech, "On Some Biases Encountered in Modern Audio Quality Listening Tests—A Review," *J. Audio Eng. Soc*, vol. 56, pp. 427–451 (2008 Jun.).

[7] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ—The

ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, pp. 3–29 (2000 Jan./Feb.).

[8] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual Evaluation of Speech Quality (PESQ)—The New ITU Standard for End-to-End Speech Quality Assessment Part I–Time-Delay Compensation," *J. Audio Eng. Soc.*, vol. 50, pp. 755–764 (2002 Oct.).

[9] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ)—The New ITU Standard for End-to-End Speech Quality Assessment Part II: Psychoacoustic Model," *J. Audio Eng. Soc.*, vol. 50, pp. 765–778 (2002 Oct.).

[10] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I—Temporal Alignment," *J. Audio Eng. Soc.*, vol. 61, pp. 366–384 (2013 Jun.).

[11] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part II—Perceptual Model," *J. Audio Eng. Soc.*, vol. 61, pp. 385–402 (2013 Jun.).

[12] S. E. Olive, "A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part II—Development of the Model," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), convention paper 6190.

[13] M. Dewhirst, R. Conetta, F. Rumsey, P. Jackson, S. Zielinski, S. George, S. Bech, and D. Meares, "QESTRAL (Part 4): Test Signals, Combining Metrics, and the Prediction of Overall Spatial Quality," presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), convention paper 7598.

[14] S. Fenton and H. Lee, "Towards a Perceptual Model of 'Punch' in Musical Signals," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), convention paper 9381.

[15] S. Kim and W. L. Martens, "Deriving Physical Predictors for Auditory Attribute Ratings Made in Response to Multichannel Music Reproductions," presented at the *123rd Convention of the Audio Engineering Society* (2007 Oct.), convention paper 7195.

[16] M. Takanen and G. Lorho, "A Binaural Auditory Model for the Evaluation of Reproduced Stereophonic Sound," presented at the *AES 45th International Conference: Applications of Time-Frequency Processing in Audio* (2012 Mar.), conference paper 6-6.

[17] W. Klippel, "Multidimensional Relationship between Subjective Listening Impression and Objective Loudspeaker Parameters," *Acta Acust. united Ac.*, vol. 70, no. 1, pp. 45–54 (1990).

[18] I. B. Labuschagne and J. J. Hanekom, "Preparation of Stimuli for Timbre Perception Studies," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2256–2267 (2013). https://dx.doi.org/10.1121/1.4817877

[19] E. Schubert and J. Wolfe, "Does Timbral Brightness Scale with Frequency and Spectral Centroid?" *Acta Acust. united Ac.*, vol. 92, pp. 820–825 (2006).

[20] A. Gabrielsson, "Perceived Sound Quality of Reproductions with Different Frequency Responses and Sound Levels," *J. Acoust. Soc. Am.*, vol. 88, no. 3, p. 1359 (1990). https://dx.doi.org/10.1121/1.399713

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

[22] J. G. Beerends, K. V. Nieuwenhuizen, and E. L. v. d. Broek, "Quantifying Sound Quality in Loudspeaker Reproduction," *J. Audio Eng. Soc.*, vol. 64, pp. 784–799 (2016 Oct.). https://doi.org/10.17743/jaes.2016.0034

[23] T. H. Pedersen and N. Zacharov, "The Development of a Sound Wheel for Reproduced Sound," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9310.

[24] N. Cowan, "On Short and Long Auditory Stores," *Psychol. Bull.*, vol. 96, no. 2, pp. 341–370 (1984). https://dx.doi.org/10.1037/0033-2909.96.2.341

[25] ITU-R, "Recommendation BS 1116-3, Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," Recommendation ITU-R BS 1116-3, International Telecommunication Union Radiocommunication Assembly (ITU-R), Switzerland, Feb. 2015.

[26] G. Lorho, G. Le Ray, and N. Zacharov, "eGauge—A Measure of Assessor Expertise in Audio Quality Evaluations," presented at the *AES 38th International Conference: Sound Quality Evaluation* (2010 Jun.), conference paper 7-2.

[27] B. R. Glasberg and B. C. J. Moore, "A Model of Loudness Applicable to Time-Varying Sounds," *J. Audio Eng. Soc.*, vol. 50, pp. 331–342 (2002 May).

THE AUTHORS

Christer P. Volk

Søren Bech

Torben H. Pedersen

Flemming Christensen

Christer P. Volk received a B.Eng. degree in electronics and computer engineering from the Engineering College of Copenhagen in 2009. He studied acoustics, focusing on psychoacoustics, at the Technical University of Denmark, where he obtained an M.Eng. degree in engineering acoustics. In January 2017, he defended his Ph.D. thesis entitled: "Prediction of Perceptual Audio Reproduction Characteristics" made in collaboration with DELTA SenseLab and Aalborg University, Denmark.

•

Søren Bech received a M.Sc. and a Ph.D. from the Department of Acoustic Technology (AT) of the Technical University of Denmark. From 1982–92 he was a research fellow at AT studying perception and evaluation of reproduced sound in small rooms. In 1992 he joined Bang & Olufsen where he is Director of Research. In 2011 he was appointed Professor in audio perception at Aalborg University. His research interest includes human perception of reproduced sound in small and medium sized rooms, experimental procedures, and statistical analysis of data from sensory analysis of audio and video quality.

•

Torben H. Pedersen received a M.Sc. from the Department of Acoustics at the Technical University of Denmark in 1977. He has been working with various disciplines within acoustics and psychoacoustics at DELTA since 1978 and in the latest years as a Senior Technology Specialist at DELTA SenseLab. He has been working with perceptual models for noise related metrics for tones, impulses, and low frequency noise. He participated in a research project on sound quality with Aalborg University, Bang & Olufsen, and Brüel & Kjær in 2001–2006. He is currently working with perceptual audio evaluation and has in collaboration with colleagues developed the Sound wheel [23] for reproduced sound.

•

Flemming Christensen was born in Horsens, Denmark, in 1969. In 1993 he received his M.Sc.E.E. degree with specialization in acoustics from Aalborg University, and in 2001 a Ph.D. degree in the area of binaural technology. He is holding a position as Associate Professor at Aalborg University and his main research interests are within electro acoustics, binaural technology, sound quality, and human sound perception. He is teaching acoustics, sound perception and FPGA and microprocessor systems design, and recently also acoustics and signal processing at The Royal Academy of Music (Aarhus/Aalborg). He is appointed technical assessor within acoustics for the Danish Accreditation Fund. In his spare time he has had a long career as musician playing mainly keyboards and guitars in many different musical genres.