OPEN ACCESS Freely available online

Journal of the Audio Engineering Society
Vol. 64, No. 1/2, January/February 2016 (© 2016)
DOI: http://dx.doi.org/10.17743/jaes.2015.0094

On Some Biases Encountered in Modern Audio Quality Listening Tests (Part 2): Selected Graphical Examples and Discussion

SŁAWOMIR ZIELIŃSKI, *AES Member*
(s.zielinski@pb.edu.pl)

Faculty of Computer Science, Białystok University of Technology, Poland

This paper provides complementary data to the review of biases in audio quality listening tests by Zieliński et al. (2008) [1]. The paper presents selected illustrations of range equalizing bias, centering bias, stimulus spacing bias, contraction bias, and bias due to nonlinear properties of assessment scale. The illustrations are given in graphical form and respective discussions of biases using empirical data obtained by various researchers over the period of the past 15 years. The presented collection of illustrations along with the discussion may help the experimenters to identify potential biases affecting their data and avoid typical pitfalls in reporting the outcomes of the listening tests.

0 INTRODUCTION

This paper is an extension of the previous paper on biases encountered in audio quality listening tests by Zieliński et al. [1]. It provides some new as well as old but newly analyzed material. While the previous paper covered a broad range of biases, this paper focuses only on five types of systematic errors potentially affecting quantifying judgments. Namely, it provides graphical examples of stimulus spacing bias, centering bias, range equalizing bias, contraction bias, and bias due to nonlinear properties of an assessment scale, using Poulton's classification of biases [2]. The five types of biases presented in this paper were selected due to their prevalence in audio and speech quality evaluation experiments as assessed by this author. The paper does not expand much on the theory of bias modeling, already covered by the previous paper, but it serves for illustrative purposes hopefully clarifying the material presented in our previous review.

In contrast to the previous paper where the five aforementioned types of biases were illustrated using predominantly abstract or theoretical graphs, this paper is based on empirical data. The figures illustrating the biases were plotted using the data obtained by this author, directly acquired from other researchers or were extracted from the literature from the past 15 years. The rationale for their selection was to gather a set of visual illustrations representing the most typical biases in the listening tests, demonstrating that they can have a large influence on the experimental data.

Some of the graphs included in this paper were already presented at various AES Conventions [3], [4], [5] but they have never been published in a *Journal* article. This paper gathers the previously scattered examples together in one place and presents them in a succinct form. In addition, the paper provides an example of the range equalizing bias that was never published in print before but was presented at the ITU-T workshop [6]. Moreover, the paper presents new examples illustrating how, in our opinion, the range equalizing bias propagated and perpetuated itself in the most recent standards for the objective assessment of speech quality. It also presents a new example illustrating the application of the MUSHRA test [7] to the quality assessment of speech codecs where a systematic shift of data was exhibited.

The paper includes an example of a bias encountered in the quality assessment of broadcast video signals. Although the example comes from a different discipline, it was included in the paper due to its uniqueness as it demonstrates variation across quality assessment methods in terms of their resilience to biases. A classical example of a bias encountered in loudspeakers assessment was also included in the paper due to its distinctiveness.

Although much effort was taken by the author to present the most illustrative examples available, demonstrating the five aforementioned biases in isolation from other effects, caution must be exercised by the reader when visually inspecting the graphs. Some data were taken from the experiments that were not designed to investigate the bias effects. Hence, the presented results might have been influenced not only by the discussed biases but, to some extent, by

some other factors. Nevertheless, it is assumed by the author, based on the data origin and the theoretical models of the biases reviewed in our previous paper [1] that the presented examples were predominantly affected by each of the discussed biases and therefore can serve as valid illustrations of the biases in question.

Classical studies on semantic scaling of the verbal terms used in speech, audio or video quality assessment indicated that the presence of verbal quality descriptors along the scales might introduce substantial non-linear warping of the assessment scale. In our previous paper we quoted some contradictory findings suggesting that the above departure from linearity, if any, was less than inferred from the semantic scaling experiments [1]. In this paper we took this point further by asserting that the semantic scaling experiments themselves might have been at flaw. We also hypothesized, based on the presented graphs, that the major factor causing a potential departure of an assessment scale from linearity could be due to the stimulus spacing bias, not due to the presence of the labels, as commonly assumed.

It is assumed that the reader is already familiar with the commonly used methods for the assessment of speech or audio quality. It is also assumed that the reader is acquainted with the theoretical models of the typical biases occurring in quantifying judgments, such as those reviewed by Zieliński et al. [1]. An uninitiated reader is referred to the basic textbooks on audio and speech quality assessment, e.g., by Bech and Zacharov [8] or by Raake [9]. The standard procedures commonly used for subjective assessment of audio and speech quality are outlined in the following recommendations: ITU-R Rec. BS.1534 (MUSHRA) [7], ITU-R Rec. BS.1116 [10], and ITU-T Rec. P.800 [11].

There are ongoing efforts to reduce systematic errors encountered in the internationally standardized methods. For example, the work of the ITU-R group 6C culminated in the recent revision of the BS.1534 standard (MUSHRA) [7] [12]. Although some improvements were achieved with respect to the standard clarity, the selection of assessors or the guidelines for statistical analysis, no fundamental improvements were demonstrated in relation to the reduction of biases or development of new diagnostic tools.[1] This highlights the need for further studies into the methodologies of quality assessment.

It is hoped that the presented collection of the graphical examples of biases, taken from the empirical studies, will prompt the researchers to take every precaution to avoid or to reduce biases in their listening tests.

1 CONFUSION AROUND THE TERM "CONTEXT"

Despite the existence of the formal classification system of the systematic errors in quantifying judgments [2], many researchers from the field of quality assessment across various disciplines (speech, audio, multimedia) use the term "context" or its derivatives such as "contextual effects" to describe any source of systematic errors observed in their experimental data. A broad meaning of these terms gives rise to some ambiguity in the reported results. Such imprecise terminology can even be misleading as there are at least three different meanings for which the term "context" or its derivatives is normally used:

1) Stimulus context—the range or the distribution of stimuli [13]–[19],
2) Scaling context—the scaling method used [20],
3) Environmental or situational context [5] [21].

Each of the three aforementioned categories is still very broad. Even within the first category there are at least several possible mechanisms potentially giving rise to bias, which will be demonstrated in the paper.

The reason for using such broad terms in reporting experimental biases by the researchers is unknown. It is possible that the researchers want to exercise caution when giving precise names to the experimental errors observed. Although for given experimental data one of the biases is normally dominant, taking precedence from other biases, it might be challenging for researchers to identify and name it. It might also be possible that there is no sufficient knowledge among the researchers to allow them to identify and name the experimental errors. Regardless of the answer, the aforementioned vagueness exemplifies the need for disambiguating the terminology used in listening tests reports.

2 RANGE EQUALIZING BIAS

In this section and throughout the paper two approaches will be used in reporting a magnitude of a bias. The first one, which we feel is more intuitive, is to report an observed deviation as a percentage of the whole range of the rating scale used, be it a 5-point mean opinion score (MOS) scale or a 100-point quality scale. Considering that in some disciplines, in particular those where measurement units have no intrinsic meaning, the magnitude of the reported effects size is often normalized to a pooled variance [22], Cohen's d values will also be reported.

The range equalizing bias, in its extreme manifestation, causes the scores obtained in listening tests to span the whole range of the assessment scale regardless of the actual range of the evaluated stimuli. A graphical model of the range equalizing bias and other biases presented in this paper can be found in [1]. Its magnitude depends on the number of stimuli under assessment. As it will be demonstrated below, the range equalizing bias causes the assessment scale to self-calibrate in order to encompass the whole range of stimuli, potentially causing a substantial systematic shift in data. This is why some authors use the term "rubber ruler" effect to describe the self-calibrating property of the assessment scale forced by the range equalizing bias [23].

A systematic shift of listening test scores, which could be attributed to the range equalizing bias, is illustrated in Fig. 1. The data presented here come from the two separate experiments described in [24] and [25]. The listening

[1] Discussions on improved anchors for MUSHRA are currently underway in ITU-R Workgroup 6C-1A.

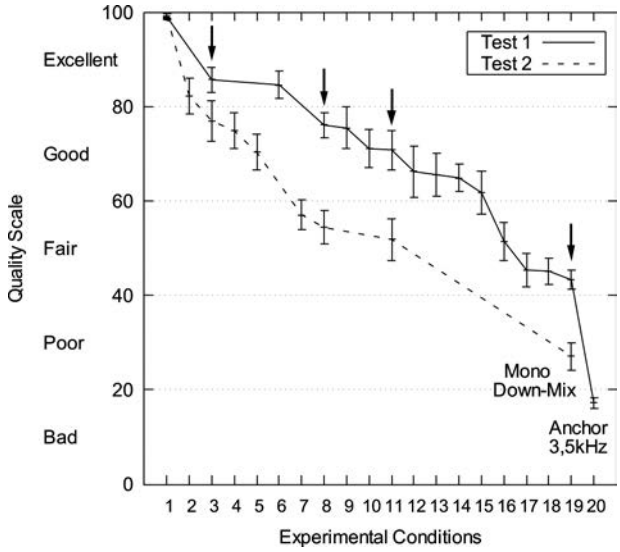

Fig. 1. Example of the range equalizing bias in the MUSHRA test. Graph shows mean values and 95% confidence intervals (CI). Data extracted from the experiments undertaken by Zieliński et al. and published in [24] and [25]. Arrows indicate identical experimental conditions in both tests.

panel in these two experiments consisted of 21 and 10 experienced listeners, respectively. Both experiments were designed according to the MUSHRA test; however, in the second experiment the method was modified by discarding the low-quality 3.5 kHz anchor. The data obtained from that latter experiment is referred to as Test 2 (see the dashed line). Only the data obtained for the same program material and for the same listening position were extracted from both experiments in order to provide a consistent basis for comparison. As it is shown in Fig. 1, the range of the scale used by the listeners, calculated as the distance between the maximum and minimum values, in both experiments is similar and equals 73 and 83 points respectively, with a difference of only 10% of the range of the rating scale (Cohen's $d = 0.81$). Considering that the perceptual range of the stimuli was different in both experiments, this may be an indication of the range equalizing bias. The highest quality conditions were the same in both experiments (hidden reference) and are represented by the scores placed at the top of the scale. However, the lowest quality conditions were different. In Test 1 a 3.5 kHz low-pass filtered anchor was of the lowest quality and is represented by condition No. 20. For Test 2 the lowest level of quality was exhibited by a mono down-mix (condition No. 19). Low-pass filtering causes much more detrimental effects to the audio quality than down-mixing [24], which to some extent is confirmed by a relatively large distance between the scores obtained for down-mix to mono and for the 3.5 kHz anchor in Test 1, being equal to 26% of the range of the scale with a Cohen's d value of 1.97 (see conditions 19 and 20 respectively). Consequently, one would expect a bigger difference between the minimum values obtained from the two experiments. No experimental data is currently available to prove this claim, and respective additional tests are for future work.

Table 1. Magnitude of the bias observed in Fig. 1.

Condition No	3	8	11	19
Difference	9%	22%	19%	16%
Cohen's d	0.43	1.22	0.86	1.00

Four experimental conditions presented in Fig. 1, apart from the hidden reference, were identical in both experiments and hence they can be used as control conditions to test for the presence of any potential biases. The results obtained for these four common experimental conditions are indicated with the arrows (conditions 3, 8, 11, and 19). As can be seen, the results obtained for the control conditions are different between the experiments. The magnitude of these differences is quite large as it ranges up to 22% of the range of the scale ($d = 1.22$), which corresponds to a whole category interval used for the verbal labels in the MUSHRA test. Consequently, the meaning of the results, based on the scale labels, varies between the experiments. For example, the experimental condition No. 11 was assessed as "good" in Test 1 and as "fair" in Test 2. The overview of the effect sizes observed in Fig. 1 is provided in Table 1.

Note that the results presented using a dashed line in Fig. 1 (Test 2) were substantially biased because they were obtained using the modified MUSHRA standard without a mandatory 3.5 kHz low-pass filtered anchor. The MUSHRA guideline explicitly demands a 3.5 kHz anchor to be included in the tests in order to minimize the bias. Such a large bias is not likely to occur when using the anchor as required by the guideline. However, it needs to be emphasized here that a choice of a standard 3.5 kHz low-pass filtered anchor may not always be conducive for minimizing the bias. Anchors can play a stabilizing role in quality assessment methods and can also be used as diagnostic tools, provided that their characteristics are perceptually similar to those exhibited by stimuli under assessment (see "Requirements for Optimum Anchor Behaviours" in the recent version of the MUSRHRA standard [7]). Since the perceptual effects of low-pass filtering is different to that evoked by modern audio codecs, the utility of the 3.5 kHz anchor may be questionable. As mentioned before, the discussions are underway in ITU-R 6C-1A group to identify an appropriate set of anchors.

Another example of data affected by the range equalizing bias is provided in Fig. 2. The graphs were plotted using the data obtained by Cheer [26]. In his experiment Cheer asked listeners to evaluate audio quality of three sets of band-pass filtered speech recordings. The sets contained narrowband, wideband, and fullband stimuli respectively and were assessed by three separate groups according to the Absolute Category Rating (ACR) method [11]. Each group consisted of 20 naïve listeners. The exact values of cut-off frequencies of a band-pass filter applied by Cheer to process speech stimuli are presented in the graph along its horizontal axis. The minimum level of quality was identical in all three groups of stimuli. However, the maximum level of quality varied between the groups and was the highest for the fullband set, medium for the wideband set, and the

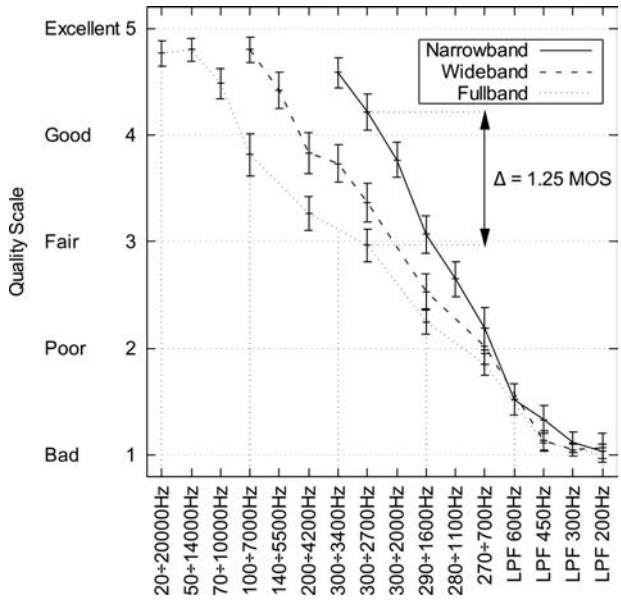

Fig. 2. Example of the range equalizing bias in the Absolute Category Rating (ACR) test. Graph shows mean values and 95% CI. Data obtained from Cheer [26] and formerly presented at the ITU-T workshop [6].

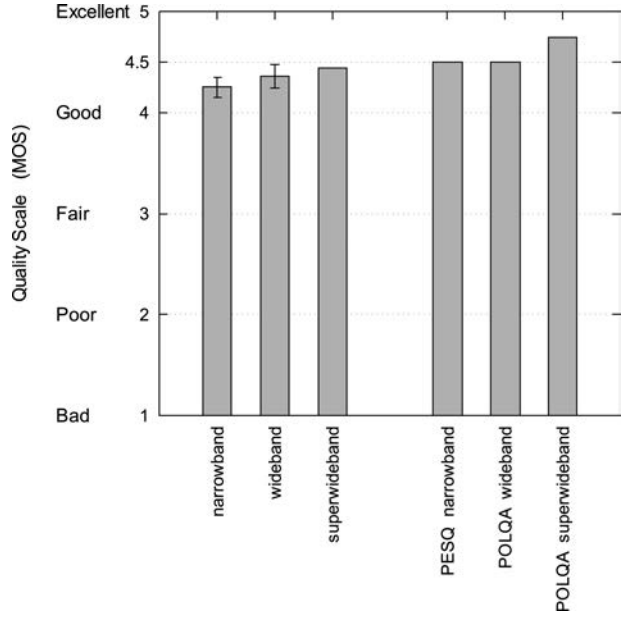

Fig. 3. Example of the range equalizing effect in subjective and objective assessment of non-coded clean telephone speech. Error bars indicate 95% CI.

lowest for the narrowband set of stimuli. Thus, the three sets of stimuli represented three different ranges of quality levels. Under bias-free conditions this should be reflected in three different distributions of quality scores produced by listeners. However, regardless of the perceptual differences in quality between the stimuli groups, the scores obtained in the listening tests spanned almost the same range of the scale, as it is demonstrated in Fig. 2. This effect constitutes a typical manifestation of the range equalizing bias. It is also interesting to note that the bias caused a substantial systematic shift of data for some control stimuli. This effect was most pronounced for 300–2700 Hz stimuli, causing a difference of 1.25 Mean-Opinion-Score (MOS) points between the narrowband and wideband experiments (31% of the range of the rating scale with a Cohen's *d* value being equal to 2.00), as indicated in the figure.

Another example of the range equalizing bias is presented in Fig. 3. The first three boxes in the figure illustrate the typical results of the subjective assessments obtained for non-coded clean speech recordings for the three following telephone bandwidths accordingly: narrowband (300–3400 Hz), wideband (50–7000 Hz), and superwideband (50–14000 Hz). The figure shows that the results for the non-coded clean speech are typically equalized to approximately 4.5 MOS, regardless of the actual bandwidth and hence regardless of actual speech quality. The plotted results were extracted from ITU Technical Paper [27] (Fig. 2 and 30) and from the paper by Xie et al. [28] (Tab. 4, confidence intervals not available). The remaining three boxes in the figure show the results of the objective quality assessment of the non-coded speech. They were plotted based on the results presented by Pomy [29]. Since the algorithms for the objective evaluation of speech quality were calibrated to the data obtained using subjective tests, they exhibit similar effects to the ones observed in subjective tests, including the range equalizing bias, as shown in Fig. 3. The quality of the narrowband non-coded speech as assessed by the PESQ algorithm [30] is equal to the MOS level of the quality of the wideband speech measured by POLQA algorithm [31]. The quality measurement of the non-coded superwideband speech is, according to POLQA algorithm, equal to 4.75 MOS. Hence, regardless of the bandwidth of non-coded clean speech, the results for the three examples presented in the right-hand side of the figure are "equalized" (narrowed downed) to the range between 4.5 and 4.75 MOS points.

Note that the results presented in Fig. 3 come from the experiments that were not designed to be compared. One of the recent ITU recommendations (P.800.2) states that such comparisons should not be made as they may not be "meaningful" [32]. The results presented in Fig. 3 should not be viewed as a departure from the recommended practice but as a deliberate exemplification of the point made by that recommendation regarding the relative properties of the MOS scores.

The algorithm for the perceptual evaluation of speech quality algorithm (PESQ), as defined in ITU-T Rec. P.862 [30], was originally calibrated to the dataset obtained using narrowband speech stimuli. To accommodate for the growing popularity and the need to assess quality of wideband telephony the extension of the PESQ algorithm was developed [33]. The basic PESQ algorithm and its wideband extension produce different scores for the same control stimuli, which can be attributed to the range equalizing bias. This measurement mismatch is acknowledged in the ITU-T P.862 recommendation. The document explains that "direct comparisons between scores produced by the wideband extension and scores produced by baseline ITU-T

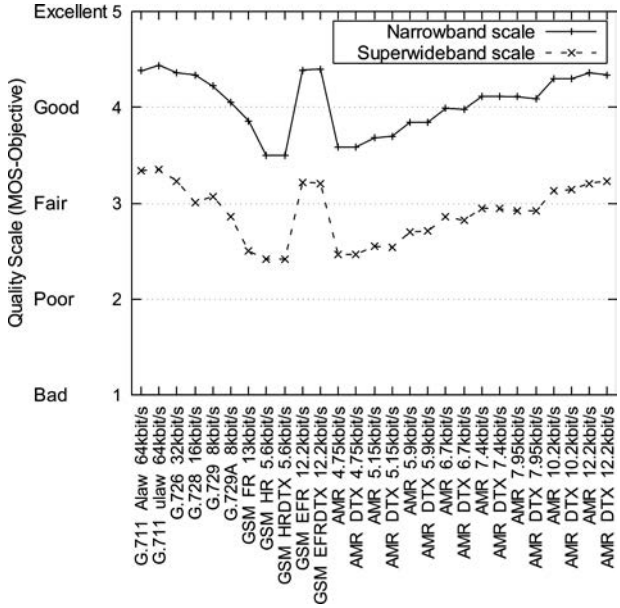

Fig. 4. Example illustrating the use of two different metric scales by the POLQA algorithm. Data taken from Appendix II of ITU-T Rec. P.863.1 [34] (typical average expected scores after level pre-alignment).

Rec. P.862 or ITU-T Rec. P.862.1 are not possible, due to the different experimental context" [33]. Consequently, the PESQ users are left with the algorithm producing scores using two different MOS scales, one for narrowband stimuli and another for wideband stimuli. This situation is far from optimal, particularly in view of progressing convergence of speech and audio technologies where ultimately one full bandwidth quality scale is needed for both speech and audio applications.

A similar "double-scale" scenario perpetuated to the most recent standard for the objective speech quality assessment known under the acronym of POLQA [31]. The algorithm can operate in one of two modes, effectively producing scores using two different MOS scales, one for narrowband stimuli (300–3400 Hz) and another for superwideband stimuli (50–14000 Hz). This difference is illustrated in Fig. 4. It shows the expected objective scores for the same codecs under the two modes of operation of the POLQA algorithm: narrowband and superwideband. The quality of codecs is overestimated when the algorithm operates in narrowband mode compared to the results obtained in the superwideband mode. The discrepancy between the scores is considerable as it reaches approximately 1 MOS point (25% of the scale).

The magnitude of the range equalizing bias can be controlled by using the direct or indirect anchoring techniques, as we already explained in detail in our previous paper [1]. The most important points regarding the anchoring techniques will be reiterated here. In the indirect anchoring technique the listeners are not informed about the inclusion of the anchor stimuli (the fact of the inclusion of the anchors in an experiment is hidden from assessors). In contrast, in the direct anchoring technique the assessors are informed about the anchor recordings and also instructed in the way how to assess them. In the simplest case of direct anchoring, two stimuli are used to determine some characteristic points on the scale, typically associated with the end points of the scale (or near the ends, in order to leave a margin for some more extreme judgments). In this way the assessment scale gets standardized according to the anchor stimuli as they set a permanent "yardstick" [36]. Examples of the studies where this technique was used for sound quality assessment could be found in [37] [38]. In a more sophisticated version of this method, extra anchors to define intermediate points along the scale can also be used.

If the low and high quality anchors are selected in such a way that they encompass the range of stimuli under assessment, the magnitude of the range equalizing should be kept constant and independent of the stimuli under assessment. In the MUHSRA standard the hidden reference by definition fulfills the requirements for the high-quality direct anchor. However, the low-quality 3.5 kHz indirect anchor may not always meet the above requirements. If, for example, a group of stimuli under assessment contains at least one stimulus exhibiting a lower quality than that of the 3.5 kHz anchor, the above requirement would be violated (this could be checked using a pilot experiment). In such case experimenters may need to define their own low-quality anchor. The requirements for optimum anchor behaviors can be found in Attachment 5 to Annex 1 of the recent release of the MUSHRA recommendation [7].

3 CENTERING BIAS

While the previously discussed range equalizing bias can be said that it "stretches the scale," the centering bias causes the scores to "float" along the scale, rendering the assessment scores relative rather than absolute. The mechanism of the bias can be explained as follows. If a given stimulus is assessed in the presence of lower quality items, it is rated as having better quality than it actually does and consequently gets overestimated scores. On the other hand, if a stimulus is assessed together with higher quality stimuli it tends to get underestimated scores. According to Poulton [2], the centering bias does not affect the relative distances between the judgments but it determines the way in which the judgments are projected onto the grading scale.

An example of the effect that could be explained using the centering bias model is presented in Fig. 5. It contains combined results from the two MUSHRA tests of coded English speech performed by Skoglund [39]. The listening panel in these two tests consisted of 19 and 18 listeners respectively. Note that this is one of the examples of applying the MUHSRA standard to the application that was beyond its original scope of applications. The MUSHRA method was originally intended solely for audio applications, not for evaluation of transmitted speech quality. The two tests contained a common condition in the form of the 3.5 kHz anchor. For clarity the results for the hidden reference were omitted in the figure. In Test 3 the anchor was assessed in the presence of lower quality items and as a result its score was overestimated. In contrast, the same anchor was assessed

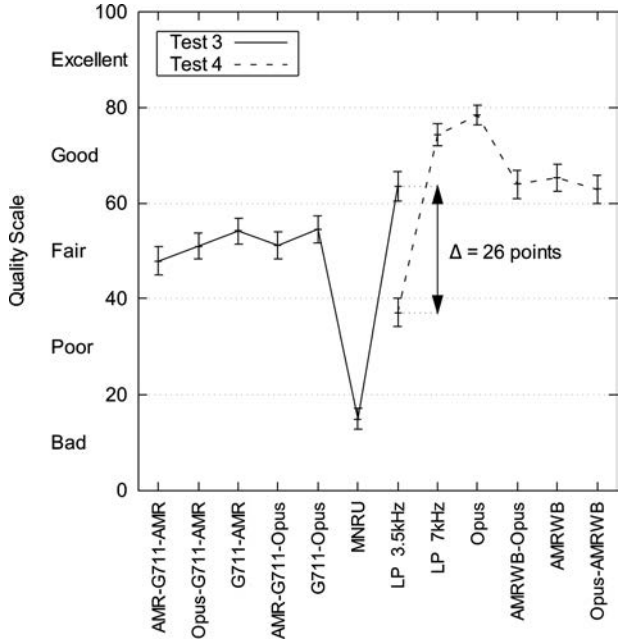

Fig. 5. Example of a shift in scores for a 3.5 MUSHRA anchor. Graph shows mean values and 95% CI. Data extracted from the experiment undertaken by Skoglund [39]. Results obtained for the hidden reference were omitted.

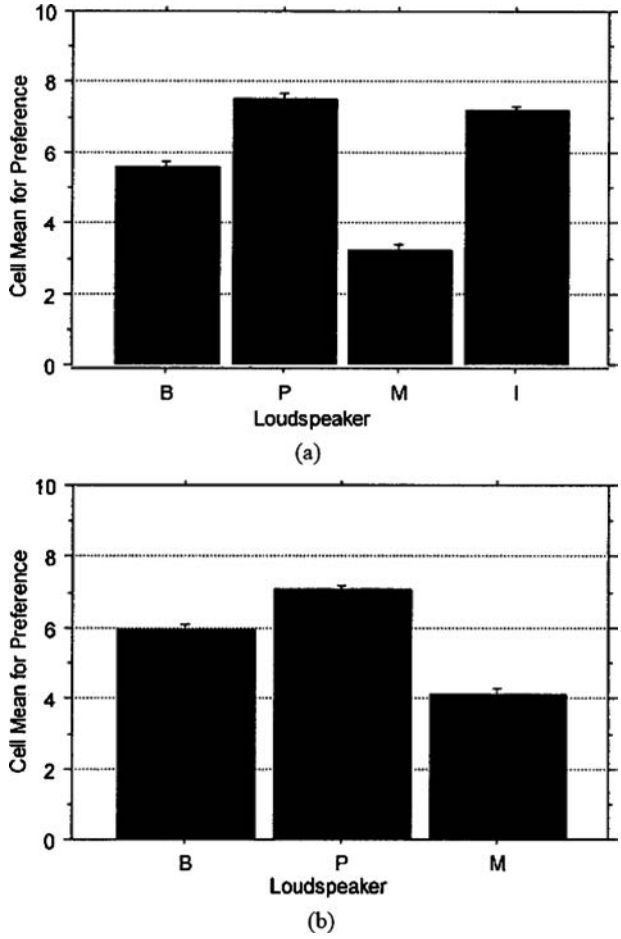

Fig. 6. Mean loudspeaker ratings and 95% CI. (a) Four-way test. (b) Three-way test. The graphs were originally published by S. E. Olive in *J. Audio Eng. Soc.,* vol. 51 (2003) [40] (used with permission).

with higher quality recordings in Test 4 and as a result its score got underestimated. This phenomenon introduced a difference of 26 points between the tests, as indicated in the figure (26% of the range of the rating scale with a Cohen's *d* value being equal to 1.21). In terms of quality labels the anchor was assessed as "good" and as "poor" in Tests 3 and 4 respectively. This indicates that MUSHRA scores exhibit relative, not absolute properties.

Another example of a possible centering bias is illustrated in Fig. 6. It shows two cases of loudspeaker ratings obtained by S. E. Olive [40]. In his study the listening panel consisted of 268 listeners. In case (a), four loudspeakers were assessed, whereas in case (b) three loudspeakers were under assessment. The three loudspeakers were common in both cases. The inclusion of the high quality loudspeaker *I* in case (a) caused a downward shift of the preference scores obtained for loudspeakers *B* and *M* compared to case (b). The magnitude of this shift for loudspeakers *B* and *M* was equal to 0.37 and 0.88 points respectively, which accounts to 4% and 9% of the range of the rating scale. This effect can be explained by the centering bias. It can also be explained by the range equalizing bias as discussed by Olive in his paper [40].

A graphical example of a systematic shift in the results of the listening tests was recently shown in a paper by Lee et al. [41] (not presented in this article). Their publication constitutes one of the unique papers where authors made an attempt to detect and diagnose bias effects. In their experiment, focused on the quality assessment of the commercial digital radio systems, the authors observed and reported a difference between the results obtained in two phases of their listening tests. Not only did they provide plots illustrating the aforementioned differences but they also included a hypothetical graph explaining likely perceptual mechanisms responsible for the observed effect. In view of their explanation, the centering bias might have been a predominant cause of the discrepancy of the results seen in phase 1 of their experiment.

In theory, the magnitude of the centering bias depends on the mid-point between the minimum and maximum level of the stimuli used in the test [2]. Therefore, in order to reduce the magnitude of this bias the low and high-quality anchors should be chosen in such a way that they encompass the range of all stimuli under assessment. In other words, the quality level of the high-quality anchor should be equal or greater than the level of the maximum quality item under evaluation, whereas the quality level of the low-quality anchor should be equal or less than the level of the minimum quality item under assessment, which can be verified by means of a pilot test.

4 STIMULUS SPACING BIAS

The stimulus spacing bias originates from listeners' tendency to equalize the differences between the scores

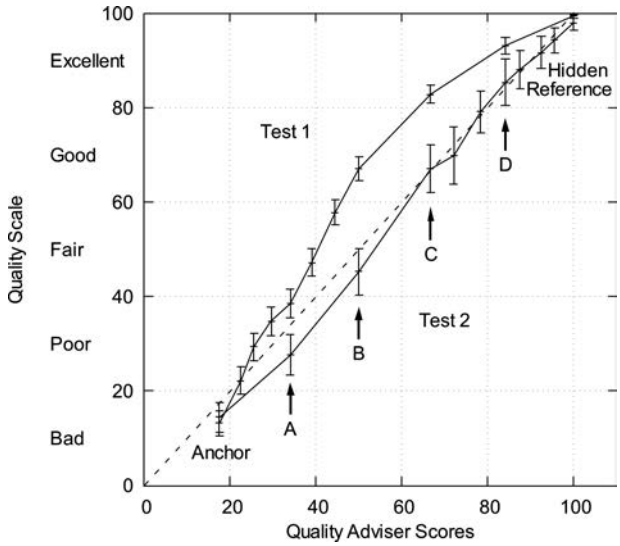

Fig. 7. Example of a nonlinear shift in MUSHRA scores caused by the stimulus spacing bias. Graph shows mean values and 95% CI. Data obtained from Hardisty and formerly published by Zieliński et al. in [3]. Arrows indicate the same conditions.

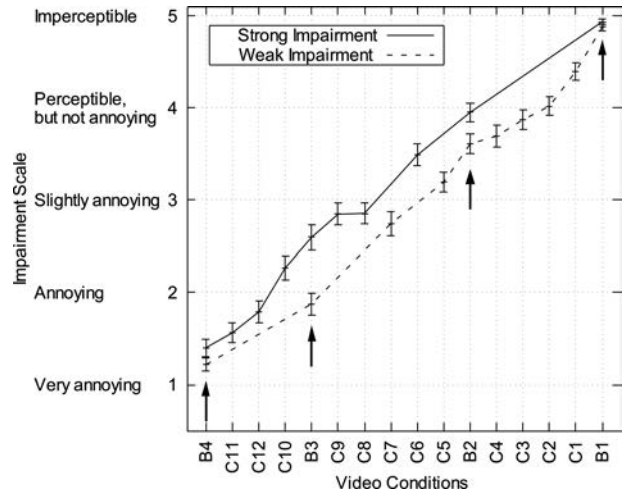

Fig. 8. A shift in video quality impairment scores likely caused by the stimulus spacing bias. Results of the international study averaged across four laboratories. Graph shows mean values and 95% CI obtained for DSIS II method. Data extracted from ITU-R Doc. 11E/34-E (pages 47 and 48) [43]. Arrows indicate the same video anchor conditions.

Table 2. Magnitude of the bias observed in Fig. 7.

Condition	A	B	C	D
Difference	11%	22%	16%	8%
Cohen's *d*	0.81	1.73	1.46	0.77

regardless of the actual quality differences between the stimuli under assessment. It is likely to occur when the distribution of stimuli under assessment is particularly uneven, for example when a pool of evaluated items contains predominantly high quality recordings (negative skew of distribution) or when it contains predominantly low quality recordings (positive skew of distribution). Listeners tend to expand the differences within densely populated stimuli and compress the differences within sparsely distributed stimuli. In contrast to the previously discussed biases, it neither affects the range of the resultant scores nor the way the scores "float" on the scale but it does affect the relative distances between the adjacent quality scores. Consequently, it distorts equal-interval property of a rating scale causing its nonlinear warping. The effect of the stimulus spacing bias on human judgments was demonstrated by Mellers and Birnbaum [16]. Its graphical model can be found in [1] and [2].

An example of the stimulus spacing bias in MUSHRA tests is presented in Fig. 7 and Table 2. It contains the results of the sound quality assessment of the low-pass filtered audio recordings obtained in two separate experiments, one with a positively skewed distribution of the stimuli (Test 1) and another one with a negatively skewed distribution of the stimuli under assessment (Test 2) [3]. In each experiment a different group of 15 listeners was employed.

The horizontal axis in Fig. 7 represents objective quality scores for the anchor conditions derived from the expert system known as Quality Adviser [42]. The objective scores from the expert system were chosen to form a baseline data for the plotted results. Although the Quality Adviser was developed using a large dataset of results coming from experiments with varied conditions, potentially neutralizing systematic errors, it is not claimed that the scores on the *x*-axis are bias-free. They were chosen here to serve as a frame of reference. As it can be seen in the figure, the stimulus spacing bias caused the discrepancy in results between the tests, with the maximum effect observed in the middle of the scale and diminishing effects towards the ends of the scale. For the worst and the best recordings in terms of audio quality, referred to in the figure as Anchor and Hidden Reference respectively, the bias had no effect. The plotted results exemplify an important property of the stimulus spacing bias: it predominantly affects the scores in the middle part of their distribution, causing a nonlinear warping of the assessment scale.

More research would be needed to verify whether the scale warping property of the stimulus spacing bias is universal or specifically related to some quality assessment methods. Corriveau et al. reached the conclusion that some assessment methods might be more prone to the stimulus spacing bias than others [13]. In their experiment they compared the three methods known as Double Stimulus Continuous Quality Scale (DSCQS), Double Stimulus Impairment Scale–variant II (DSIS II), and Comparison method. They examined the resilience of each method to the skew of the distribution of the stimuli under assessment. Out of the three methods the DSIS II technique was the most susceptible to the bias, which is illustrated in Fig. 8 and summarized in Table 3. Although the example presented in the figure is concerned with an assessment of picture quality, it was included in the paper due to its uniqueness, as it comes from a large-scale international study designed to investigate the contextual bias effects. The figure presents

Table 3. Magnitude of the bias observed in Fig. 8.

Condition	B4	B3	B2	B1
Difference	5%	18%	9%	1%*
Cohen's *d*	0.29	0.74	0.64	0.18*

Statistically not significant ($p > 0.05$)

two sets of results for the anchor conditions representing various levels of picture quality (B1, B2, B3, and B4). The solid line represents the results obtained for the stimuli set with prevalence of strongly impaired excerpts whereas the dashed line signifies the scores obtained for the stimuli set with dominance of weakly impaired items. The results for the strong and weak impairments were obtained from the two groups of viewers consisting of 54 and 56 participants respectively. Similarly to the previously presented example, the stimulus spacing bias caused an uneven shift in the data, with the most pronounced effect observed in the middle of the scale (anchors B2 and B3) and with only small effects at its ends. The difference between the conditions for B1 anchor was not statistically significant but the effect size was calculated and included in Table 3 for completeness.

Corriveau et al. found out that for the Comparison method the magnitude of the bias was less pronounced than for the DSIS II technique, while for DSCQS method no bias effect was observed (differences between the scores obtained for the anchors' conditions were not statistically different at $p = 0.05$ level). It is impossible to infer from their research how resilient to the stimulus spacing bias (and to what degree) are the methods used in speech and audio quality assessment, such as the ACR or the technique recommended in ITU-R BS. 1116. This issue constitutes an open research topic.

In order to reduce the stimulus spacing, bias stimuli under assessment should ideally be uniformly distributed. This requirement is often impractical since an experimenter may have little control over the distribution of the conditions under assessment. For example, in benchmarking tests a group of codecs under evaluation may exhibit only high quality levels, thus distorting the uniformity of distribution. Nevertheless, using a set of uniformly distributed anchor conditions may to some extent balance out an uneven distribution of stimuli under assessment. In addition, these anchors may be used as diagnostic conditions to check for a presence of the stimulus spacing bias between the listening tests. For example, in speech quality tests it is already a common practice to use a set of seven anchors (direct speech condition and six modulated noise reference unit conditions [MNRU]) [44].

Another hypothetical solution to reduce the stimulus spacing bias would involve using a medium quality anchor assigned directly to a mid-point of the scale. In principle, the above approach could be extended to any number of direct anchors associated with equidistantly distributed points along the rating scale. The proposed methods of direct anchoring, to the knowledge of the author, were never systematically tested or reported in the audio engineering literature.

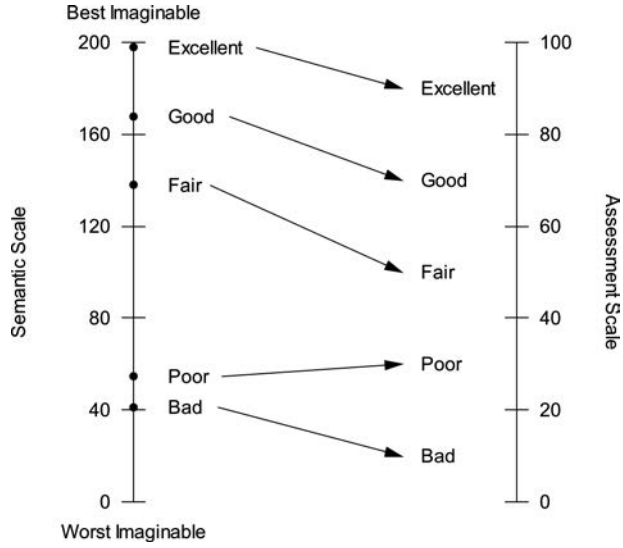

Fig. 9. A graphical model of bias due to perceptually nonlinear distribution of labels. Data on the left-hand side of the figure plotted based on the results obtained by Watson (taken from Table 15 on page 131 in [45]).

5 BIAS DUE TO SCALE NONLINEARITY

Studies by Watson [45] and by Jones and McManus [55] on semantic scaling of the verbal terms used in quality assessment concluded that the presence of the labels along the rating scales might introduce substantial non-linear warping of the assessment scale (see the caption of Fig. 12 in our previous paper [1] for more references). Although this conclusion is well grounded on the outcomes of their experiments, it is likely to be incorrect. In our previous paper we already highlighted some contradictory findings indicating that the above departure from linearity, if any, was less than could be inferred from the semantic scaling experiments [1]. The examples illustrated below in this section imply that the magnitude of the bias due to the labels is smaller than commonly asserted. In Sec. 7.3 we also provide the arguments indicating that the original studies on semantic scaling of the verbal terms used in quality assessment might have been flawed and need to be revisited. Moreover, we introduce a hypothesis that the major factor causing a nonlinear warping of data could be due to the stimulus spacing bias, not due to the presence of the labels.

As we demonstrated in the previous section, the stimulus spacing bias could violate a linear property of an assessment scale. Another potential cause of a departure of metric properties of an assessment scale from linearity is typically attributed to uneven semantic differences between quality descriptors attached to scales, such as the terms "excellent," "good," "fair," etc. This effect is demonstrated in Fig. 9. The data points on the left-hand side of the figure, together with the associated verbal labels, were positioned using the empirical results of semantic scaling of a group of adjectives, which was part of the study conducted by Watson [45]. According to her results, British English speakers regard the terms "poor" and "bad" as similar, but the terms "poor"

and "fair" as distinctively different, which is reflected in the nonequidistant distribution of the quality terms on left-hand side of the figure. The right-hand side of the figure shows a standard uniform distribution of the labels along the assessment scale used in listening tests. As it is shown in the figure, the nonequidistant label points from the perceptual (semantic) scale are mapped onto the equidistant points on the assessment scale, causing a nonlinear warping effect. Some distances between adjacent labels are compressed while others get expanded. More examples of such nonlinear mapping, including several languages, can be found in [1] (see Fig. 12).

Metric properties of the standard rating scales used in audio quality assessment were investigated by Zieliński et al. [4]. In their case study, focused on the MUSHRA test, they asked three independent groups of listeners, consisting of 15, 15, and 13 listeners respectively, to assess audio quality of the same set of audio stimuli using three types of scales respectively. In the first case, they used the standard quality scale with the equidistantly spaced quality terms such as "excellent," "good," "fair," etc. In the second case they exploited the standard impairment scale with the equidistantly spaced impairment descriptors such as "imperceptible," "perceptible, but not annoying," etc. In the third case the researchers used a scale without any verbal descriptors. The last case was assumed by the authors to represent a bias-free scenario. Their results are summarized in Fig. 10.

The top graph (a) of Fig. 10 shows the scatter plot of the scores obtained using the scale with the standard quality terms plotted against the baseline scores obtained with the label-free scale. The diagonal line represents a linear relationship between the scales ($y = x$). Under the bias-free condition, all the measurement points should be positioned along the reference line. According to the graph, there is no departure from linearity in the lower part of the quality scale. However, the presence of the quality descriptors caused a nonlinear warping of the scale in its upper part, resulting in a slight underestimation of the scores. The maximum departure from linearity ($y = x$) is equal to only 4% of the range of the rating scale ($d = 0.51$).

The effect of the presence of the equidistantly spaced labels on the impairment scale is depicted in the middle graph (b). The results indicate almost perfect collinear relationship between the scores obtained using the labeled impairment scale and those achieved with the label-free scale, with a slight negative offset. The maximum value of this offset is equal to 6% of the range of the rating scale ($d = 0.67$). The exception is the score obtained for the hidden reference, which in both cases was placed at the top end of both scales. According to the results presented in graphs (a) and (b), the departure from linearity, if any, was much smaller than one would expect from the distribution of data points on the left-hand side scale of previously discussed Fig. 9.

Fig. 10(c) was included for completeness. It represents the scores obtained using the labeled quality scale plotted against the labeled impairment scale. Considering different semantic meaning of the labels on both scales, the obtained results could be regarded as intriguing. There is almost a

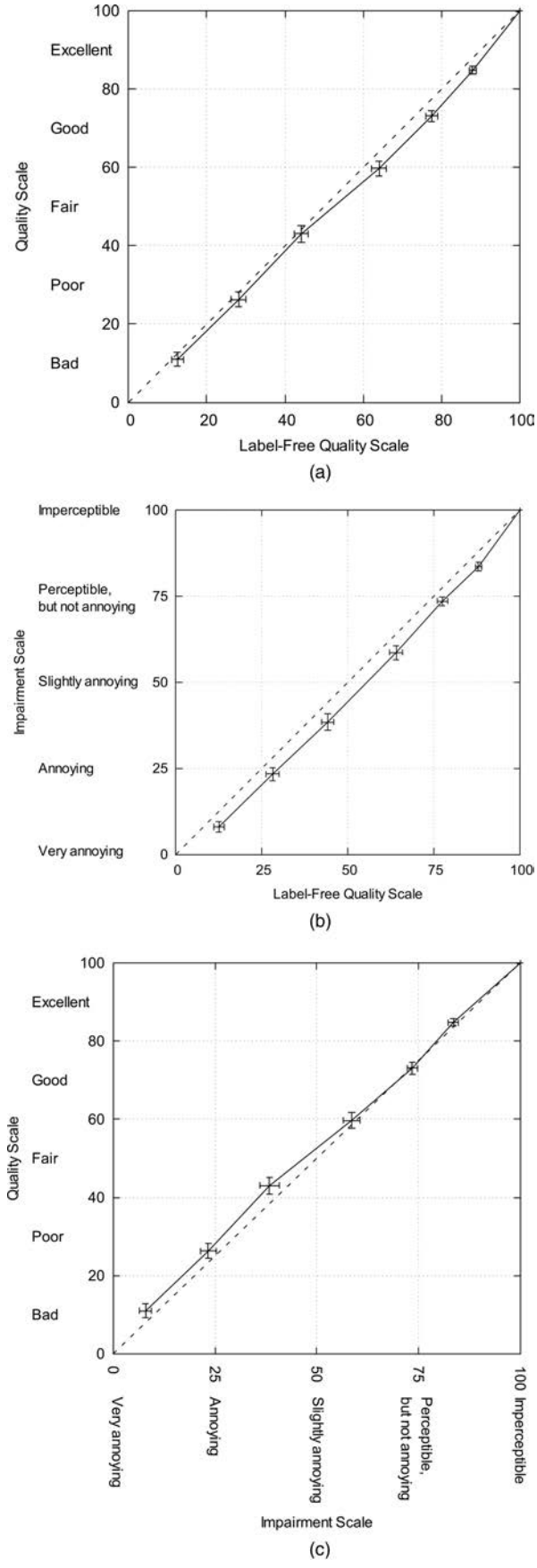

Fig. 10. Comparison of the MUSHRA test scores obtained using three types of the scales (see text for explanation). Graph shows mean values and 95% CI. Based on the data obtained by Brooks and formerly published by Zieliński et al. in [4].

perfect linear match between the scores with only a slight departure from the reference line at the lower part of the scales. The maximum magnitude of this departure is equal to only 5% of the range of the rating scale with a Cohen's d value of 0.34.

A potential bias caused by the presence of verbal labels along the scale may be reduced either by using a graphical scale with only numbers and two verbal descriptors at its ends or by removing all the labels from the scale and indicating to assessors its polarization. An example of the study where the label-free quality scale was used can be found in [46].

Note that the magnitude of the bias effects presented in graphs (a), (b), and (c) in Fig. 10 is very small. The maximum experimental size of the bias, observed in Fig. 10(b), is equal to only 6% of the range of the scale ($d = 0.67$). This magnitude is less than the effect size of the other biases illustrated in this paper. This observation implies that the standard quality labels attached to rating scales are less "problematic" in terms of biasing the data than commonly assumed. This point will be discussed further in the remainder of the paper (Secs. 7.1 and 7.3).

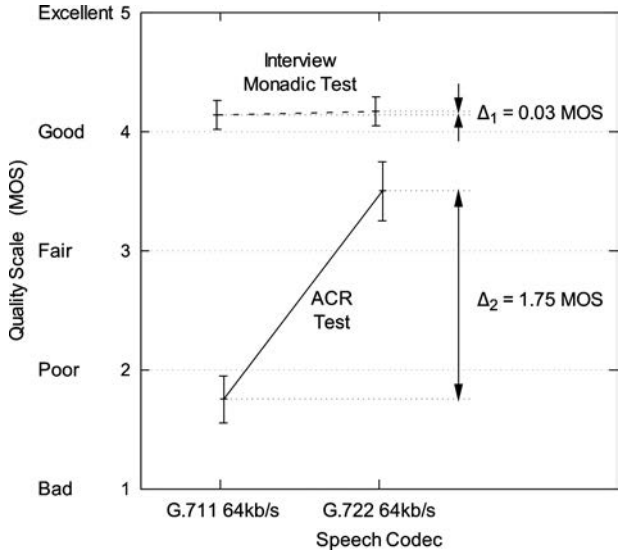

Fig. 11. Results of speech quality assessment based on the data published by Daengsi et al. (Interview Monadic Test [47]) and by Raake et al. (ACR Test [49]). Graph shows mean values and 95% CI.

6 CONTRACTION BIAS

The contraction bias causes shrinking of the distribution of the scores. Consequently, instead of spanning the whole range of a scale, data is shrunk and projected onto its part. The likelihood of the contraction bias occurrence is, loosely speaking, inversely proportional to the number of stimuli under assessment. Therefore, the monadic method of sound quality assessment is said to be the most conducive for the occurrence of the contraction bias [1]. The monadic method is a subjective quality assessment technique in which each participant assesses one and only one stimulus, without any comparison with other stimuli.

Under its extreme manifestation, the contraction bias may introduce so strong a data shrinking effect, that it could mask any genuine experimental factors. This in turn may lead experimenters to getting null results. For this reason it is very difficult to find in the literature any graphical examples of the contraction bias since the researchers are reluctant to publish reports with null results. Nevertheless, Fig. 11 presents the results that could be explained using the contraction bias model. The MOS values plotted at the top of the figure, labeled as Interview Monadic Test, were extracted from Table 2 of the paper by Daengsi et al. [47]. The error bars were calculated by this author based on the standard deviation and the number of votes provided in the aforementioned table, using the standard equation for 95% confidence intervals. In this example the researchers attempted to quantify a benefit of using a wideband speech codec (G.722) over the conventional narrowband codec (G.711). To this end they undertook a large scale experiment in which 100 listeners assessed the quality of the narrowband codec and another group of 101 listeners rated the quality of the wideband codec. Each listener assessed one and only one codec (monadic test) using an interview test. The obtained MOS results for the narrowband and wideband codecs were remarkably similar, being equal to 4.14 and 4.17 respectively, with the difference of only 0.03 MOS points, which constitutes 0.75% of the range of the scale, with an effect size d being equal to 0.05. According to the results of the Student's t-test, there was no statistical difference between the mean scores ($p = 0.743$). Nevertheless, the aforementioned effect size was calculated and included here for completeness. Its magnitude should be treated with reservation as it accounts for experimental (statistical) noise, not for a genuine experimental effect. It can be argued that the listeners, having no comparison with other stimuli, rated stimuli as "good," perhaps giving them a "benefit of the doubt." Consequently, the scores were compressed and projected near the "good" category (see the figure).

In their experiment Daengsi et al. tested five hypotheses and despite a large number of assessors none of them proved to be statistically significant [47]. They explained the null results by stating that for Thai speakers, who took part in the experiment, the spectral content above the conventional narrowband was of no benefit (most important speech spectral content bound within the traditional narrow telephone bandwidth) [47]. This explanation can be challenged. Although it might be true that for a tonal language bandwidth enhancement beyond 3.4 kHz might be of little benefit in terms of speech intelligibility, there is a growing body of research demonstrating that bandwidth extension does improve overall speech quality for many languages, including the tonal ones [44] [48]. Therefore, it is more plausible that the null results obtained by Daengsi et al. were due to the contraction bias.

For comparison, the data for G.711 and G.722 codecs obtained using the standard ACR test were also included in Fig. 11. The results were taken from the study undertaken by Raake et al. (data extracted from Fig. 6 in [49],

conditions N, M). The difference between the MOS scores was in this case equal to 1.75 MOS points (44% of the scale), indicating a substantial quality difference between the codecs. Möller et al. calculated the impairment factors for various codecs using the approach based on ITU-T E-model [50]. According to their results, the impairment factor for G.711 codec was higher, thus indicating lower quality, than the one calculated for G.722 codec. The values of the impairment factors for G.711 and G.722 codecs were equal to 34.4 and 12.3 respectively [52]. This outcome is in accordance with the results obtained using the POLQA standard, which indicate that the speech quality of G.722 codec is superior to that of G.711 codec. The objective MOS values for G.711 (A law, 64 kbit/s) and G.722 (64 kb/s) codecs obtained from POLQA method in a superwideband mode with level pre-alignment are equal to 3.34 (min. 2.79, max. 3.80) and 4.34 (min. 4.04, max. 4.46) respectively. The POLQA scores were taken from Table II.2 in ITU-T Rec. P.863.1 [34].

As already mentioned, the difference between the experimental results for the interview monadic test, denoted as Δ_1 in Fig. 11, was equal to only 0.03 MOS points. In contrast, for the ACR method the difference between the scores, represented by Δ_2, amounted to 1.75 MOS points. Since the above two tests were not designed to be directly compared, only an approximate estimation of the contraction bias effect could be attempted. The magnitude of the contraction bias might be estimated as the difference between the values of Δ_2 and Δ_1, normalized to the length of the scale. Using this approach, the magnitude of the bias illustrated in Fig. 11 was estimated to be equal to about 43% of the range of the scale, which constituted the maximum bias effect presented in this paper. Since the standard deviations for the ACR test were unknown to the author, the calculations of Cohen's d value were omitted.

The magnitude of the contraction bias is a function of the number of stimuli: the more stimuli under assessment, the smaller the magnitude of the contraction bias. Since in typical listening tests undertaken according to the standard methods the number of stimuli assessed by each listener is relatively large, the likelihood of the contraction bias is rather small. Nevertheless, the contraction bias is typically observed in the results of the speech quality tests according the ACR method. The scores from such tests, instead of spanning the whole scale (from 1 to 5 MOS points), range from 1.5 to 4.5 MOS points, thus limiting the operational range of the standard MOS scale by 1 MOS point.

The reason for using monadic tests by researchers springs either from the difficulty in the direct comparison of some audio conditions, e.g., comparing quality of automotive and laboratory acoustical environments [5], or from the pursuit to enhance the ecological validity of the experiments [53]. Note that none of the current standards for speech and audio quality assessment recommends using monadic tests and hence chances of getting the null results as in Fig. 11 are small, if the modern guidelines are adhered to. In order to reduce the contraction bias as a minimum precaution each listener should be familiarized with the range of the audio conditions under assessment prior to commencement of a listening test. The familiarization phase allows assessors to learn the range of the stimuli and it helps them to establish a rule by which they make their assessments. In theory, the magnitude of the contraction bias gets reduced with the increased number of stimuli under assessment [2]. Using multiple comparison tests, such as recommended in the MUSHRA method, combined with indirect or direct anchoring techniques may also help to reduce or even to remove the contraction bias [1].

7 DISCUSSION

Due to the reasons already explained in the introduction, the examples presented in this paper should be treated with some reservation as the plotted results could have been affected by a mixture of various factors, not necessarily by a single "isolated" bias.

Table 4 summarizes the biases illustrated in this paper and reiterates the example factors affecting their magnitude. As can be seen in the table, the range equalizing bias depends on the number of stimuli under assessment. According to Poulton [2] an increase in the number of stimuli under assessment may raise a magnitude of the range equalizing bias. An opposite effect may be observed for the contraction bias. In this case an increase in a number of stimuli may decrease the size of the contraction bias. As shown in the table, the centering bias depends on the mid-point between the maximum and the minimum quality stimuli [2]. Although the exact value of the mid-point may be difficult to ascertain, it can be controlled by adjusting the maximum and minimum stimuli (the range of stimuli). For the standard audio quality assessment methods, such as the MUSHRA test [7] or the method based on ITU-R Rec. BS.1116 [10], the maximum stimulus is determined by the hidden reference (assessed at the top of the scale), whereas the minimum stimulus is normally defined by the low-quality anchor. As it is indicated in the table, the stimulus spacing bias is dependent on the skew of stimuli under assessment and the number of stimuli under assessment. The possible causes of the scale non-linearity bias could be attributed to uneven semantic differences between quality descriptors attached to rating scales. In general, more experimental factors may contribute to the biases summarized in Table 4. They could be identified using the systextual experimental design described below in Sec. 7.4.

7.1 Comparison of the Bias Effects

The purpose of this section is to provide an approximate comparison of the effect sizes associated with the biases illustrated in this paper. The data provided below should not be used to make any universal or final inferences, since the effect size of each bias is dependent on the specific experimental design and therefore one cannot generalize the results. Nevertheless, the compared results, shown in Table 5, not only give some initial insight into the "strength" of the discussed biases but may also be used to draw some preliminary conclusions, with the reservations indicated.

Table 4. Summary of the biases illustrated in the paper with selected factors affecting their magnitude.

Bias Type	Selected Factors Affecting the Bias Magnitude
Range Equalizing Bias	Number of stimuli under assessment [2], [17], [18]
Centering Bias	Mid-point between the minimum and maximum stimuli (range of stimuli) [2]
Stimulus Spacing Bias	Skew of a distribution of stimuli under assessment [2], [13], [17]; Number of stimuli under assessment [51]
Scale Non-Linearity	Uneven semantic differences between quality descriptors attached to scales (see Fig. 12 in [1]), [45], [55]
Contraction Bias	Number of stimuli under assessment [2]

Furthermore, the presented comparison may prompt some researchers to undertake a more detailed investigation into the bias effects, e.g., using the systextual design discussed below.

Out of all compared effects, the contraction bias exhibited the maximum magnitude, being equal to approximately 43% of the rating scale (see Table 5). If this size of the effect is validated by future research, the monadic method should be avoided in listening tests. As mentioned before, none of the standard methods for the subjective audio or speech quality assessment recommends the monadic test. Nevertheless, some researchers may still choose to use this method either to enhance the ecological validity of their experiments or due to the circumstances preventing listeners from a direct comparison of audio stimuli. In such cases a caution must be taken when interpreting the results as the data may be substantially "shrunk" due to the contraction bias.

The size of the experimental effect shown in the illustrations exemplifying the centering bias, the range equalizing bias, and the stimulus spacing bias was of a similar magnitude, ranging from 22% to 31%, and therefore these three biases were placed in the second rank in Table 5. No further sub-ranking between them was attempted due to the reasons explained at the outset of this section. Although one cannot exclude the possibility that in general the magnitude of these three biases may exceed 31%, the author is not aware of any examples in the literature demonstrating a greater magnitude of the experimental bias than the one presented in Table 5. In his recent experiment designed to investigate the stimulus spacing bias in the MUSHRA method, Zieliński [35] reported the maximum observed effect to be equal to 27%, which fits within the above range. Therefore, one can tentatively conclude that as long as there are no more extreme conditions compared by listeners than the ones investigated in this paper, the magnitude of the centering, range equalizing, and stimulus spacing biases is likely to be less than that of the contraction bias.

The bias effect due to the scale non-linearity was tentatively placed under the third rank in Table 5, as it amounted to only 6%. This is a small magnitude compared to the effect sizes exhibited by the other biases illustrated in the paper. If the rank order of the experimental effects presented in the table is confirmed by future research, with the scale non-linearity effect ranked at the bottom, one could conclude that a bias due to the verbal labels attached to rating scales is less "problematic" than the effects originating from the other biases. Hence, it can be tentatively concluded that such factors as the range, the distribution skew or the number of stimuli under assessment may have far greater impact on the experimental results than the presence of the quality labels (or lack of thereof). Note that the above conclusion was drawn from the experiments employing the standard descriptive labels, as defined by the current recommendations [7] [10] [11]. Therefore, it cannot be generalized to listening tests in which non-standard labels are used.

7.2 Relative Properties of Scores

The presented examples, such as those illustrating self-calibrating and "floating" properties of the assessment scales, provide evidence that the scores obtained using the standard audio and speech assessment methods are inherently relative, not absolute. This conclusion is in line with the observations made by psychologists [2] [17]. The above assertion is also in line with the observation made by other researchers working in a similar discipline [57]. This conclusion is also consistent with the recently issued ITU-T Recommendation P.800.2 [32] regarding the interpretation of MOS results. The recommendation states that the MOS values depend on experimental factors including the context of the stimuli. It also explains that the word "absolute" used in the name of the Absolute Category Rating (ACR) method does not mean that the resultant scores are absolute but that the listeners' judgments are made in isolation, without relative comparisons. Nevertheless, the term "absolute" used in the ACR method could still be misinterpreted.

Table 5. Comparison of the observed effects presented in this paper. The presented rank order of the biases is for illustrative purposes with limited external validity.

Rank Order	Type of Bias	Illustration	Max. Difference Observed	Max. *d* value
1	Contraction Bias	Fig. 11	43%	–
2	Range Equalizing Bias	Fig. 2	31%	2.00
	Centering Bias	Fig. 5	26%	1.21
	Stimulus Spacing Bias	Fig. 7	22%	1.73
3	Scale Non-Linearity	Fig. 10(b)	6%	0.67

Since most of the algorithms used for the objective measurement of speech or audio quality were calibrated to the data originated in the listening tests, they might have "inherited" the properties seen in the listening tests, such as those illustrated in this paper. Consequently, the results of quality measurements obtained using such algorithms should be treated as relative, not absolute. It follows that the standard objective algorithms can be used for comparative measurements but they might not be legitimately used for absolute benchmarking. For example, it might not be a legitimate practice to assign an objectively measured MOS value to a given codec and treat it as an absolute measurement result.

7.3 Labels and Non-Linearity Issues

The purpose of this section is to show that the original experiments concerned with the semantic scaling of the verbal descriptors used in the standard quality assessment methods might have been flawed and need to be revisited. Moreover, in this section we introduce a new hypothesis that the major factor causing a non-linear warping of data could be due to the stimulus spacing bias, not due to the presence of the labels.

The results of the semantic scaling of the verbal descriptors used in the standard quality assessment methods published by Watson and by other researchers indicate that the verbal labels might cause substantial departure from linear metric properties of grading scales [45] [54] [55]. However, the results presented in Fig. 10 show much smaller, if any, departure from linearity and thus contradict the conclusions derived from the semantic scaling experiments. The observation that the effect of the labels on the nonlinearity of the scale is much smaller than initially assumed is in line with the reports from the international multi-laboratory experiments. Despite a great likelihood of nonlinear bias in such tests, for example due to translation issues, the data obtained for the same stimuli from various laboratories are typically collinearly related [57] [58]. Interestingly, in one of the pioneering studies from the field of speech quality assessment, some researchers used the label "unsatisfactory" instead of the standard term "bad." Although the term "unsatisfactory" is not on the same semantic continuum as the standard quality terms ("bad," "poor," "fair," etc.), the obtained scores were also collinearly related to that obtained in other laboratories [59].

These results may question the correctness of the scaling experiments performed by Watson and by other researchers. In fact, there are reasons to believe that their experiments were subject to the previously discussed stimulus spacing bias. For example, an unusually large expansion of the distance between "poor" and "good" labels in the Watson's data could have been caused by presence of nine intermediate labels in the scaling experiment, compared to an average of two extra labels between the other pairs (see Table 15 and Figure 20 in [45]). Hence, it would be advisable to re-examine their results by undertaking further scaling experiments using only the standard descriptors, without any intermediate items such as "satisfactory," "good enough," "acceptable," "adequate," "sufficient," "passable," etc.

The presented examples cast some doubt on usefulness of the verbal descriptors placed along the rating scales. As already demonstrated, some biases can shift scores by one or even by two categories, say from "poor" to "good." This may imply that listeners tend to ignore the labels and perhaps use them only as an indicator of a polarization of a scale. This supposition is supported by the results presented in Fig. 10(c). Considering the large semantic differences of the labels on both scales, one would expect a greater discrepancy between the data. However, the scores exhibited almost a perfect match, suggesting that the listeners might have ignored the meaning of the labels. The unusually close linear match between the scores obtained from distinctively different labeled scales was also observed in the more recent experiments reported by Raake et al. [62], Tominaga et al. [63], and Kawano et al. [64]. These examples, taken from different disciplines, indicate that the verbal quality descriptors and the impairment descriptors might be used interchangeably or could be entirely removed from the scales without fear of jeopardizing the metric properties of a rating scale. These modifications to the standard scales are formally permitted in the recently standardized recommendation for the subjective assessment of video quality, audio quality, and audiovisual quality of Internet video (ITU-T Rec. P.913 [65]). One of the most straightforward ways of reducing a potential bias caused by the semantically uneven distribution of quality labels is to use a graphic scale only with numbers (without semantic quality descriptors). An example of the study in which a label-free scale was used can be found in [46].

Considering the preliminary outcomes of the comparison of the effect sizes between the biases illustrated in Table 5, with the reservations indicated in Sec. 7.1, the magnitude of the effect due to the labels is smaller than the one caused by the other biases. We already illustrated in Fig. 7 and Fig. 8 that the stimulus spacing bias can introduce a substantial non-linear warping of the data. Hence, based on the data gathered so far, we propose a hypothesis that the major factor introducing a non-linear warping of experimental data (violation of linear metric properties of an assessment method) could be due to the stimulus spacing bias, not due to the presence of the labels along rating scales, as commonly assumed. In order to check the validity of this hypothesis more research would be required.

7.4 Systextual Design

Some researchers prompted by this discussion may want to include in their listening tests extra experimental factors allowing them to check to what extent their results are dependent on the biases illustrated in this paper. To this end a so called systextual design could be employed. This approach, originally introduced by Birnbaum [66], involves a systematic change of the experimental design itself, including an experimental context determined by the stimuli.

In Table 6 we propose an example of such design. The first three factors included in the table are related to the stimuli themselves. The systextually designed experiment, based on these three factors, would involve manipulating

Table 6. Example factors proposed to be included in the systextual design of listening tests.

Experimental Factors	Biases Likely to be Affected				
	Contraction Bias	Centering Bias	Range Equalizing Bias	Stimulus Spacing Bias	Scale Non-Linearity Bias
Number of Stimuli	x		x	x	
Range of Stimuli		x			
Skew of Stimuli Distribution				x	
Calibration of a Rating Scale	x	x	x		x
Type of Listeners	x	x	x	x	

the number, the range, and the distribution skew of stimuli under assessment. As shown in the table, changing the number of stimuli is likely to affect the magnitude of the contraction bias, of the range equalizing bias, and of the stimulus spacing bias; whereas manipulating the range of the stimuli is likely to influence the magnitude of the centering bias. Modifying the skew of the stimuli distribution is likely to affect the magnitude of the stimulus spacing bias.

Considering the literature on the topic of audio and speech quality assessment methodologies over the past 15 years, the first three factors listed in Table 6 seem to be almost entirely overlooked. Most of the research on the topic was focused on such factors as the design of the assessment scales, semantic issues, the methods of stimuli presentation, the assessment environment or the types of assessors, while the number, the range or the distribution skew of the stimuli under assessment were almost entirely ignored in the experimental design. For example, Raake at al. recently undertook an experiment aiming to compare the results obtained using the ACR and the MUSHRA methods [62]. In their experiment they kept the number of stimuli constant. Based on their empirical data they established a mapping function allowing one to convert the data between the methods. However, considering Parducci range-frequency theory [17] [18] it is likely that their mapping function would not be universal since the biases observed in both methods depend on the number of stimuli under assessment. Consequently, it is recommended that at least the first three factors from Table 6 are included in future experiments devoted to the methodology of a subjective quality assessment.

The way an experimenter chooses to calibrate a rating scale may have a direct bearing on the magnitude of the centering bias ("floating scale"), the range equalizing bias ("rubber ruler"), the contraction bias, or may violate liner properties of a rating scale (scale non-linearity bias), as indicated in Table 6. In the simplest case of the systextual design, modifying a rating scale calibration could involve using a scale interchangeably with or without the verbal quality labels; however more options should be considered. For example, some researchers may wish to quantify the effect of calibrating a rating scale using the direct or indirect auditory anchoring techniques, which are said to have a potential of reducing or stabilizing some of the biases illustrated in the paper (see [1] for more details). Some experimenters may also try to modify a distribution of the labels along the scale, their wording, or a distribution of numbers with associated tick marks along a rating scale.

The last factor proposed to be included in the systextual design is the type of listeners. The discussion presented in this section is predominantly limited to two types of listeners according to their level of experience; however, more generic criteria of listeners' segmentation could be explored as part of the systextual design. For example, listeners could be categorized according to their habitual, cultural, linguistic [71] or demographic background. While there are many reports in the literature comparing the expert and naïve listeners in terms of their performance in the listening tests (e.g., [40] [67]), the author is not aware of any study investigating to what degree various groups of listeners are liable to introduce the bias effects illustrated in this paper.

According to Bech [67], experienced listeners are more sensitive to quality distortions and more discriminating. Hence, it is hypothesized here that they may introduce a smaller magnitude of a contraction bias compared to naïve listeners. According to the results obtained by Beresford et al. [68] and by Schinkel-Bielefeld et al. [69], the scores obtained from naïve listeners in the MUHSRA test could be consistently higher than those acquired from experts. These results indicate that these two groups of listeners may vary in terms of their susceptibility to the centering bias (vertical offset of scores). In his recent paper Zieliński [35] tentatively concluded that the magnitude of the stimulus spacing bias in the MUSHRA test introduced by naïve listeners might exceed the magnitude introduced by the experienced assessors. Thus, based on the preliminary observations, it is hypothesized here that experienced listeners are likely to introduce a smaller magnitude of the biases illustrated in this paper, compared to non-experts. If this hypothesis is confirmed by future research, this would constitute another strong argument for using expert assessors in listening tests. It has to be emphasized here that although the current audio engineering standards for the subjective quality listening tests explicitly recommend using experienced listeners, in other research and development sectors, most notably in telecommunications, using naïve listeners is still the norm. It is not the intention of the author to imply the superiority of experienced raters over naïve ones. There are still some merits in using naïve assessors in listening tests, which are not mentioned in this paper. Nevertheless, as indicated above, more research is needed to compare experienced and naïve listeners in terms of their susceptibility to the bias effects.

There is cost involved in using a systextual design, as introducing any extra factor to listening tests may prolong their duration or may require employing more listeners. However, the advantages of systextually designed experiments may outweigh their increased cost. The systextual design may help experimenters to identify biases distorting their data, quantify their magnitude, and has a potential of enhancing validity of listening tests.

7.5 Implications for Development and Use of Objective Models

The objective methods for audio and speech quality assessment are becoming increasingly popular [30] [31] [56]. They are easy to operate and allow researchers to obtain results quickly and in a repeatable way. However, the objective models provide only estimates (predictors) of the average scores acquired from individual voters scoring in listening tests [70]. Since objective models are calibrated to data obtained from subjective quality assessment, their accuracy and validity would not exceed that of properly executed and analyzed listening tests. Moreover, it is of imperative importance that every potential bias during listening tests, whose results are subsequently used to calibrate objective models, is reduced (see the sections on direct and indirect anchoring in our previous paper [1]). Otherwise, there is a risk of a bias propagation from subjective tests to objective models.

Systextual design could be incorporated in the calibration and validation of models intended for objective assessment of audio quality. This would provide information on how robust models are to the bias effects. Currently, the description of the existing models, provided by their developers, is typically limited to their statistical performance and a scope of applicability (see [30] and [31] as examples). By reading their extended specifications, including the results of systextually designed calibration and validation tests, a prospective user could gauge to what degree given models are susceptible to variation in the quality range of stimuli, their number or their distribution skew. Not only this would enhance external validity of objective models but would allow prospective users to apply them in a more conscious and careful way, being aware of their limitations.

As mentioned above, there is cost involved in systextual design. However, a large number of listening tests, typically undertaken during development of objective models, could be traded for fewer but systextually designed ones.

The onus of a bias compensation, described in detail in the remainder of this section, is currently placed on users of objective models (see ITU-T Rec. P.1401 [70] or ITU-T P.863 [31]). It is suggested, therefore, that a new generation of models should include some form of artificial intelligence accounting for bias effects potentially present in user data or perpetuated in objective models (or both). For example, the new models could be equipped with an option of self-calibration to the user data.

It needs to be stressed here that a current generation of algorithms for objective assessment of audio quality is still not capable of modeling the bias effects illustrated in this paper. They use fixed perceptual models without any mechanism of self-adjusting to the range, the number or the distribution skew of stimuli under assessment. Hence, if the measurements obtained using objective models are compared to the results of subjective tests with a different number of stimuli, a different quality range or a different stimuli skew, compared to those used during the model calibration, a mismatch between the objective measurements and subjectively obtained scores is likely to occur; a phenomenon that is well known to the experts developing the objective models [70] (and perhaps surprising to the uninitiated users of objective models). As a result of the above mismatch a graph of subjectively obtained scores plotted as a function of the objective scores may exhibit a combination of the following phenomena:

- Offset effect,
- Gradient effect,
- Non-linear warping effect.

Examples of such graphs can be found in ITU-T Rec. P.1401 [70].

The offset effect is an indication that the scores from a subjective test are consistently higher (overestimated) or lower (underestimated) compared to those obtained from an objective model. The offset effect could be attributed to the centering bias, with its magnitude being a function of a mid-point of stimuli under assessment (a range of stimuli) [2]. The offset effect could also be linked to other factors such as overcritical or too lenient listeners, or even listeners' cultural, linguistic or demographic background [71]. Without applying systextual design, described in Sec. 7.4 above, it is impossible to ascertain which factor plays here a more dominant role.

Under a bias-free condition the scores obtained from a listening test, plotted against the objective scores, should be scattered along a diagonal line ($y = x$) at an angle of 45 degrees. A gradient effect manifests itself by a plot where scores are scattered along a line having a steeper or shallower slope compared to the target one. This effect could be caused by a different span of scores along x and y axis, and therefore it might be attributed to the range equalizing bias or to the contraction bias. Hence, among other factors, the number of stimuli under assessment or a type of anchoring technique used might be responsible for a gradient effect.

The subjectively obtained scores plotted against the scores calculated by the objective models may also exhibit some non-linear patterns. Such warping effects could be attributed to the stimulus spacing bias or to the scale non-linearity bias. Consequently, its origin could be linked to a skew of stimuli distribution. It could also be caused by a rating scale design, e.g., by uneven semantic distribution of labels.

The first two effects could be compensated by normalization of the subjectively obtained data using a linear regression, whereas the third effect (non-linear warping) could be compensated using a third-order polynomial mapping, constrained to be monotonic. Examples of mapping

functions recommended for a bias compensation can be found in ITU-T P.862 [30], ITU-T P.862.1 [72], ITU-T P.563 [73], and ITU-T P.863 [31]. The above-described normalization procedure may not only help to compensate for biases specific to listening tests but, as it is explicitly acknowledged in the POLQA standard, it "may also compensate for some of the systemic prediction errors caused by the objective measurement method" [31].

As mentioned above, the responsibility of compensation for bias effects is currently placed on the users of objective models. We do hope, however, that future models would prove to be more intelligent in this respect, accounting for the common biases and assisting their users in the procedure of bias compensation.

7.6 Other Considerations

The range equalizing bias discussed in Sec. 2 was presented as an undesired factor adversely affecting the resultant scores. However, some researchers deliberately introduce this effect by post-processing the data, using it to their advantage (see [56] for example). They normalize the scores in such a way that the maximum score from each listener is scaled to the top value of the scale, the minimum score is assigned the bottom value of the scale, and the remaining scores are scaled accordingly using a linear transformation. This procedure equalizes the data from each listener so that their scores always span the whole range of a scale. It has a potential of increasing the sensitivity of a test by removing inter-listener variability and by "stretching" the distribution of scores.

There is a growing body of literature promoting an alternative approach to sound quality assessment using indirect assessment methods such as paired comparison or ranking techniques (see [78] for example). Their proponents advocate the methods based on the mathematical assumptions, asserting that they are free of the typical biases encountered in the direct assessment techniques. However, more research would be needed to validate their assertions.

8 CONCLUSIONS AND RECOMMENDATIONS

Contrary to a popular opinion, the presented examples demonstrate that the scores obtained from the standard quality assessment methods are inherently relative. There are three implications of the above assertion. First, one should exercise caution making absolute inferences based on the resultant scores. For example, researchers should avoid making conclusions about the quality of tested systems by associating numerical scores with verbal descriptors placed at fixed positions along the scale. Second, broadcasters should be careful in selecting codecs based on absolute thresholds applied to the listening tests results. Third, the researchers should avoid terminology explicitly or implicitly implying that the standard assessment methods yield absolute scores.

According to the presented examples and the recent literature, the primary factor potentially introducing a departure from linearity of an assessment scale is related to an uneven distribution of stimuli under assessment, not the verbal labels or translation issues, as typically assumed. The verbal descriptors cause small departure of assessment scales from linearity. The outcomes of the classical studies of semantic scaling of the quality descriptors should be re-examined in view of the current literature as they could have been affected by the stimulus spacing bias.

Although some of the presented examples question the usefulness of the verbal descriptors used along the scales, more research would be needed to determine whether they are obsolete. There is still some merit in using them as they indicate the polarization of the scale and also allow researchers "backward comparability" with previously undertaken experiments.

The presented examples show the importance of using direct or indirect anchoring techniques as diagnostic or error-reducing tools. Since the nature of the standard anchors is no longer representative of the modern audio systems under assessment, further improvements to the MUHSRA standard are required, beyond what has been achieved in its recent revision [12]. In particular, universal anchors, which could be applied as diagnostic and error-reducing tools in joint speech and audio listening tests, still need to be defined.

As we illustrated in the paper, the latest objective methods for speech quality assessment produce their outputs using various scales in order to accommodate for the range equalizing effect. In view of the technological convergence across the disciplines it would be advisable to aim at developing methods using a single full-bandwidth scale both for speech and audio applications. Some researchers already attempted this challenge by choosing a single-scale approach [60] [61].

A systematic study to characterize the standard methods used in speech and audio quality assessment in terms of their susceptibility to the biases described in this paper is still needed. Since a magnitude of the biases in quantifying judgments depends strongly on the number of stimuli under assessment, their range, and their distribution skew, it is recommended that these factors are included in future experiments in the area of methodologies of speech and audio quality assessment. In addition, further studies on biases in quality assessment should include such factors as the type of listeners and a method of calibration of a rating scale. Such approach to experimental design may also improve usability and enhance external validity of models intended for objective assessment of sound quality.

The bias examples discussed in this paper were limited to the area of audio and speech quality assessment. During the recent years it is possible to observe a shift of interest among some researchers from the traditional quality assessment to the evaluation of "quality of experience" (QoE); as the latter approach is claimed to be more ecologically valid [74] [75] [76] [77]. The aim of the former (traditional) approach is to assess the degree to which a system under appraisal meets expectations of human evaluators based on subjectively perceived audio stimuli, with deliberately reduced or controlled influence from non-auditory factors. In contrast, the scope of QoE is much broader,

as it extends to extra-auditory factors including, for example, interaction features, usage and service features, or even socio-economic features [77]. While some consensus among scientists has been reached with regard to the taxonomy of this new field of research (see Möller and Raake [77]), it is still unclear what types of biases are specific to the evaluation of QoE. Therefore, an important question that needs to be posed in conclusion to this paper, requiring further investigation, is whether the five types of biases discussed are also pertinent to the assessment of QoE.

We hope that the article will promote further research in the area of audio and speech quality assessment and will encourage the researchers to take all possible precautions to minimize any potential experimental bias in audio and speech quality listening tests.

9 ACKNOWLEDGMENT

The author would like to thank Prof. F. Rumsey and Prof. S. Bech for their invaluable advice, guidance, and tuitions on the methodologies of listening tests during his work at Surrey University (2000–2009). The author would like to express his gratitude to Dr. Sean E. Olive for granting permission to include a copy of the figure from his *JAES* paper (Fig. 6). Special thanks are due to Dr. Jordan Cheer, Dr. J. Skoglund, Phil Hardisty for sharing their experimental data, and to the anonymous reviewers for their helpful comments on the previous version of the manuscript. The figures were plotted using *gnuplot*. This work was supported by Białystok University of Technology (Project No. S/WI/1/2013).

10 REFERENCES

[1] S. Zieliński, 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.).

[2] E. C. Poulton, *Bias in Quantifying Judgments* (Lawrence Erlbaum, London, 1989).

[3] S. Zieliński, P. Hardisty, C. Hummersone, and F. Rumsey, "Potential Biases in MUSHRA Listening Tests," presented at the *123rd Convention of the Audio Engineering Society* (2007 Oct.), convention paper 7179.

[4] S. Zieliński, P. Brooks, and F. Rumsey, "On the Use of Graphic Scales in Modern Listening Tests," presented at the *123rd Convention of the Audio Engineering Society* (2007 Oct.), convention paper 7176.

[5] K. Beresford, N. Ford, F. Rumsey, and S. Zieliński, "Contextual Effects on Sound Quality Judgements: Listening Room and Automotive Environments," presented at the *120th Convention of the Audio Engineering Society* (2006 May), convention paper 6648.

[6] Y. Jiao, S. Zieliński, and F. Rumsey, "Towards Consistent Assessment of Audio Quality of Systems with Different Available Bandwidth," ITU-T Workshop on "From Speech to Audio: Bandwidth Extension, Binaural Perception," Lannion, France, 10–12 September, 2008.

[7] ITU-R Rec. BS.1534-2, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," International Telecommunications Union, Geneva, Switzerland (2014).

[8] S. Bech and N. Zacharov, *Perceptual Audio Evaluation. Theory, Method and Application* (Wiley, Chichester, UK, 2006).

[9] A. Raake, *Speech Quality of VoIP: Assessment and Prediction* (Wiley, Chichester, UK, 2006).

[10] ITU-R Rec. BS.1116-3, "Methods for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunications Union, Geneva, Switzerland (2015).

[11] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," International Telecommunications Union, Geneva, Switzerland (1996).

[12] J. Liebetrau, F. Nagel, K. Watanable, C. Colomes, P. Crum, T. Sporer, and A. Mason, "Revision of Rec. ITU-R BS.1534," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9172.

[13] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach, "All Subjective Scales Are Not Created Equal: The Effects of Context on Different Scales," *Signal Process.*, vol. 77, pp. 1–9 (1999). DOI: 10.1016/S0165-1684(99)00018-3.

[14] L. E. Marks, T. G. Shepard, K. Burger, and E. M. Chakwin, "Flavor-Intensity Perception: Effects of Stimulus Context," *Physiology and Behaviour*, vol. 105, pp. 443–450 (2012 Jan.). DOI: 10.1016/j.physbeh.2011.08.039.

[15] H. N. J. Schifferstein, "Contextual Effects in Difference Judgments," *Perception and Psychophysics*, vol. 57, pp. 56–70 (1995). DOI: 10.3758/BF03211850.

[16] B. A. Mellers and M. H. Birnbaum, "Loci of Contextual Effects in Judgment," *J. Exp. Psychol.: Human Perception and Performance*, vol. 8, pp. 582–601 (1982). DOI: 10.1037/0096-1523.8.4.582.

[17] A. Parducci, "Contextual Effects: A Range-Frequency Analysis," in *Handbook of Perception*, vol. 2, E. C. Carteret and M. P. Friedman, Eds. (Academic Press, London, 1974).

[18] A. Parducci, "Category Ratings: Still Mode Contextual Effects!," in *Social Attitudes and Psychophysical Measurement*, B. Wagner, Ed. (Lawrence Erlbaum, Hillsdale, NJ, 1982).

[19] N. Schinkel-Bielefeld, "Quantifying Sequential Context Effects in Subjective Quality Evaluation," Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 269–274, 18–20 Sept. (2014). DOI: 10.1109/QoMEX.2014.6982330.

[20] W. Cools, J. Hofmans, and P. Theuns, "Context in Category Scales: Is 'Fully Agree' Equal to Twice Agree?" *European Rev. Applied Psych.*, vol. 56, pp. 223–229 (2006 December). DOI: 10.1016/j.erap.2005.09.007.

[21] L. Gros, S. Chateau, and S. Busson, "Effects of Context on the Subjective Assessment of Time-Varying Speech Quality: Listening/Conversation, Laboratory/Real Environment," *Acustica/Acta Acustica*, vol. 90, pp. 1037–1051 (2004).

[22] D. C. Howell, *Statistical Methods for Psychology* (Wadsworth, UK, 2002).

[23] H. T. Lawless and H. Heymann, *Sensory Evaluation of Food. Principles and Practices* (Kluwer-Plenum, London, 1998).

[24] S. Zieliński, F. Rumsey, R. Kassier, and S. Bech, "Comparison of Basic Audio Quality and Timbral and Spatial Fidelity Changes Caused by Limitation of Bandwidth and by Down-Mix Algorithms in 5.1 Surround Audio Systems," *J. Audio Eng. Soc.*, vol. 53, pp. 174–192 (2005 Mar.).

[25] S. Zieliński, F. Rumsey, and S. Bech, "Effects of Down-Mix Algorithms on Quality of Surround Sound," *J. Audio Eng. Soc.*, vol. 51, pp. 780–798 (2003 Sep.).

[26] J. Cheer, "The Investigation of Potential Biases in Speech Quality Assessment," Technical Project (BMus Dissertation), Institute of Sound Recording, University of Surrey (2008).

[27] ITU-T Technical Paper, "GSTP-GVBR Performance of ITU-T G.718," International Telecommunications Union, Geneva, Switzerland (2010).

[28] M. Xie, D. Lindbergh, and P. Chu, "ITU-T G.722.1 Annex C: A New Low-Complexity 14 kHz Audio Coding Standard," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5 (2006). DOI: 10.1109/ICASSP.2006.1661240.

[29] J. Pomy, "The Next-Generation Mobile Voice Quality Testing Standard," ITU Workshop, Moscow, 27–29 April (2011).

[30] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," International Telecommunications Union, Geneva, Switzerland (2001).

[31] ITU-T Rec. P.863, "Perceptual Objective Listening Quality Assessment," International Telecommunications Union, Geneva, Switzerland (2014).

[32] ITU-T Rec. P.800.2, "Mean Opinion Score Interpretation and Reporting," International Telecommunications Union, Geneva, Switzerland (2013).

[33] ITU-T Rec. P.862.2, "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," International Telecommunications Union, Geneva, Switzerland (2005).

[34] ITU-T Rec. P.863.1, "Application Guide for Recommendation ITU-T P.863," International Telecommunications Union, Geneva, Switzerland (2013).

[35] S. Zieliński, "Is a Multi-Slider Interface Layout Responsible for a Stimulus Spacing Bias in the MUSHRA Test?" *Archives of Acoustics*, vol. 40, no 4, pp. 585–594 (2015 Dec.). DOI: 10.1515/aoa-2015-0058.

[36] J.P. Guilford, *Psychometric Methods* (McGraw-Hill, London, 1954).

[37] S. George, S. Zieliński, F. Rumsey, P. Jackson, R. Conetta, M. Dewhirst, D. Meares, and S. Bech, "Development and Validation of an Unintrusive Model for Predicting the Sensation of Envelopment Arising from Surround Sound Recordings," *J. Audio Eng. Soc.*, vol. 58, pp. 1013–1031 (2010 Dec.).

[38] J. Paulus, Ch. Uhle, and J. Herre, "Perceived Level of Late Reverberation in Speech and Music," presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8331.

[39] J. Skoglund, "Listening Tests of Opus at Google" http://www.opus-codec.org/comparison/GoogleTest2.pdf (Fall 2011) – accessed in November 2014.

[40] S. E. Olive, "Differences in Performance and Preference of Trained versus Untrained Listeners in Loudspeaker Tests: A Case Study," *J. Audio Eng. Soc.*, vol. 51, pp. 806–825 (2003 Jun.).

[41] S. Lee, Y-T. Lee, J. Seo, M-S. Baek, Ch-H. Lim, and H. Park, "An Audio Quality Evaluation of Commercial Digital Radio Systems," *IEEE Transactions on Broadcasting*, vol. 57, pp. 629–636 (2011 Sep.). DOI: 10.1109/TBC.2011.2152910.

[42] S. Zieliński, F. Rumsey, and R. Kassier, "Development and Initial Validation of a Multichannel Audio Quality Expert System," *J. Audio Eng. Soc.*, vol. 53, pp. 4–21 (2005 Jan./Feb.).

[43] ITU-R Document 11E/34-E, "Canada, France, Germany, Switzerland. Investigation of Contextual Effects," International Telecommunications Union, Geneva, Switzerland (1997).

[44] 3GPP Document TR 26.952, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Codec for Enhanced Voice Services (EVS); Performance Characterization," Release 12 (2015).

[45] A. Watson, "Assessing the Quality of Audio and Video Components in Desktop Multimedia Conferencing," *Ph. D. thesis*, Department of Computer Science, University College London (1999).

[46] R. Conetta, T. Brookes, F. Rumsey, S. Zieliński, M. Dewhirst, P. Jackson, S. Bech, D. Meares, and S. George, "Spatial Audio Quality Perception (Part 1): Impact of Commonly Encountered Processes," *J. Audio Eng. Soc.*, vol. 62, pp. 831–846 (2014 Dec.). DOI: 10.17743/jaes.2014.0048.

[47] T. Daengsi, Ch. Wutiwiwatchai, A. Preechayasomboon, and S. Sukparungsee, "Speech Quality Assessment of VoIP: G.711 VS G.722 Based on Interview Tests with Thai Users," *I.J. Information Tech. Computer Sci.*, pp. 19–25 (2012). DOI: 10.5815/ijitcs.2012.02.03.

[48] Z. Cai, N. Kitawaki, T. Yamada, and S. Makino, "Comparison of MOS Evaluation Characteristics for Chinese, Japanese, and English in IP Telephony," *4th International Universal Communication Symposium*, Bejing (2010). DOI: 10.1109/IUCS.2010.5666762.

[49] A. Raake, M. Wältermann, and S. Spors, "Which Wideband Speech Codec? Quality Impact due to Room-Acoustics at Send Side and Presentation Method," presented at the *127th Convention of the Audio Engineering Society* (2009 Oct.), convention paper 7823.

[50] ITU-T Rec. G.107, "The E-Model: A Computational Model for Use in Transmission Planning," International Telecommunications Union, Geneva, Switzerland (1998–2015).

[51] A. Parducci and D. H. Wedell, "The Category Effect With Rating Scales: Number of Categories, Number

of Stimuli, and Method of Presentation," *J. Experimental Psychology: Human Perception and Performance*, vol. 12, pp. 496–516 (1986). DOI: 10.1037//0096-1523.12.4.496.

[52] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, "Impairment Factor Framework for Wide-Band Speech Codecs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1969–1976 (2006). DOI: 10.1109/TASL.2006.883262.

[53] N. Staelens, S. Moens, W. Van den Broeck, I. Mariën, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing Quality of Experience of IPTV and Video on Demand Services in Real-Life Environments," *IEEE Trans.on Broadcasting*, vol. 56, pp. 458–466 (2010). DOI: 10.1109/TBC.2010.2067710.

[54] ITU-R Rep. BT.1082-1, "Studies toward the Unification of Picture Assessment Methodology," International Telecommunications Union, Geneva, Switzerland (1990).

[55] B. L. Jones and P. R. McManus, "Graphic Scaling of Qualitative Terms," *SMPTE J.*, pp. 1166–1171 (1986). DOI: 10.5594/J04083.

[56] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI) Version 2," *J. Audio Eng. Soc.*, vol. 62, pp. 99–117 (2014 Mar.). DOI: 10.17743/jaes.2014.0006.

[57] M. H. Pinson, L. Janowski, R. Pepion, P. Le Callet, M. Barkowsky, and W. Ingram, "The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study," *IEEE J. Selected Topics in Signal Procesing*, vol. 6, pp. 640–651 (2012 Oct.). DOI: 10.1109/JSTSP.2012.2215306.

[58] M. Barkowsky, J. Li, T. Han, S. Youn, J. Ok, Ch. Lee, Ch. Hedberg, I. V. Ananth, K. Wang, K. Brunnström, and P. Le Callet, "Towards Standardized 3DTV QoE Assessment: Cross-Lab Study on Display Technology and Viewing Environment Parameters," *Proc. SPIE 8648, Stereoscopic Displays and Applications XXIV, 864809* (March 12, 2013.). DOI: 10.1117/12.2004050.

[59] D. Goodman and R. Nash, "Subjective Quality of the Same Speech Transmission Conditions in Seven Different Countries," *IEEE Transactions on Communications*, vol. 30, pp. 642–654 (1982 Apr.). DOI: 10.1109/ICASSP.1982.1171565.

[60] S. Möller, A. Raake, M. Wältermann, and N. Côté, "Towards a Universal Scale for Perceptual Value," Second International Workshop on Quality of Multimedia Experience (QoMEX), pp. 142–146, 21–23 June (2010). DOI: 10.1109/QOMEX.2010.5516216.

[61] A. Ramo and H. Toukomaa, "Voice Quality Characterization of IETF Opus Codec," Interspeech, Florence, Italy (2011).

[62] A. Raake, M. Wältermann, U. Wüstenhagen, and B. Feiten, "How to Talk about Speech and Audio Quality with Speech and Audio People," *J. Audio Eng. Soc.*, vol. 60, pp. 147–155 (2012 Mar.).

[63] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance Comparisons of Subjective Quality Assessment Methods for Mobile Video," Second International Workshop on Quality of Multimedia Experience (QoMEX), pp. 82–87, 21-23 June (2010). DOI: 10.1109/QOMEX.2010.5517948.

[64] T. Kawano, K. Yamagishi, and T. Hayashi, "Performance Comparison of Subjective Assessment Methods for 3D Video Quality," Fourth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 218–223, 5–7 July (2012). DOI: 10.1109/QoMEX.2012.6263833.

[65] ITU-T Rec. P.913, "Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in any Environment," International Telecommunications Union, Geneva, Switzerland (2014).

[66] M. H. Birnbaum, "Controversies in Psychological Measurements," in *Social Attitudes and Psychophysical Measurement*, B. Wegener, Ed. (Lawrence Erlbaum, Hillsdale, NJ, 1982).

[67] S. Bech, "Selection and Training of Subjects for Listening Tests on Sound Reproducing Equipment," *J. Audio Eng. Soc.*, vol. 40, pp. 590–610 (1992 Jul./Aug.).

[68] K. Beresford, N. Ford, F. Rumsey, and S. Zieliński, "Contextual Effects on Sound Quality Judgements: Part II – Multi-Stimulus vs. Single Stimulus Method," presented at the *121st Convention of the Audio Engineering Society* (2006 Oct.), convention paper 6913.

[69] N. Schinkel-Bielefeld, N. Lotze, and F. Nagel, "Audio Quality Evaluation by Experienced and Inexperienced Listeners," *Proceedings of Meetings on Acoustics*, vol. 19, ICA, Montreal, Canada, 2–7 June (2013). DOI: 10.1121/1.4805210.

[70] ITU-T Rec. P.1401, "Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models," International Telecommunications Union, Geneva, Switzerland (2012).

[71] D. U. Ebem, J. G. Beerends, J. Van Vugt, C. Schmidmer, R. E. Kooij, and J. O. Uguru, "The Impact of Tone Language and Non-Native Language Listening on Measuring Speech Quality" *J. Audio Eng. Soc.*, vol. 59, pp. 647–655 (2011 Sep.).

[72] ITU-T Rec. P.862.1, "Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO," International Telecommunications Union, Geneva, Switzerland (2003).

[73] ITU-T Rec. P.563, "Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications," International Telecommunications Union, Geneva, Switzerland (2004).

[74] F. Kuipers, R. Kooij, D. De Vleeschauwer, and K. Brunnström, "Techniques for Measuring Quality of Experience," Wired/Wireless Internet Communications, Lecture Notes in Computer Science, vol. 6074, pp. 216–227 (2010). DOI: 10.1007/978-3-642-13315-2_18.

[75] M. Schoeffler, B. Edler, and J. Herre, "How Much Does Audio Quality Influence Ratings of Overall Listening Experience?" *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research*, Marseilles, France, 15–18 October (2013).

[76] J.-N. Antons, S. Arndt, K. De Moor, and S. Zander, "Impact of Perceived Quality and other Influencing Factors on Emotional Video Experience," Seventh International Workshop on Quality of Multimedia Experience (QoMEX), pp. 1–6, 26–29 May (2015). DOI: 10.1109/QoMEX.2015.7148124.

[77] S. Möller and A. Raake (Eds.), *Quality of Experience* (Springer, Switzerland, 2014).

[78] F. Wickelmaier, N. Umbach, K. Sergin, and S. Choisel, "Scaling Sound Quality Using Models for Paired-Comparison and Ranking Data," Paper presented at the DAGA 2012 Congress, 19–22 March, Darmstadt, Germany.

THE AUTHOR

Sławomir Zieliński

Sławomir Zieliński received M.Sc. and Ph.D. degrees in telecommunications from Gdańsk University of Technology, Poland. After graduation (1992) he worked as a lecturer at the same University for eight years. In 2000 Dr. Zieliński joined the University of Surrey, UK, where he initially worked as a postdoctoral research fellow and then as a lecturer at the Department of Music and Sound Recording. Since 2009 he has been working as a teacher at the Technical Schools in Suwałki (*Zespół Szkół Technicznych w Suwałkach*). Recently, he has been employed as a lecturer (*adiunkt*) at Białystok University of Technology, Poland.

For the past 20 years Dr. Zieliński has taught classes in broad range of topics including electroacoustics, audio signal processing, sound synthesis, studio recording techniques, and more recently information and communications technologies. He co-supervised six Ph.D. students. He is the author or co-author of more than 70 scientific papers in the area of audio engineering. His current research interests include spatial audio, psychoacoustics, and audio quality assessment methodologies.