On Some Biases Encountered in Modern Listening Tests

Slawomir Zielinski

Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, UK

ABSTRACT

Hedonic judgments are prone to many non-acoustical biases. Since audio quality evaluation involves, to some degree, hedonic judgments, the scores obtained from typical audio quality listening tests can be biased if the non-acoustical factors are not properly controlled. In contrast to hedonic judgments, sound character judgments are less prone to the non-acoustical biases, and, if appropriate acoustical anchors are used, can provide highly reliable and repeatable data. However, when listeners are asked to judge multidimensional attributes, like the overall spatial audio fidelity, the resultant data may exhibit a large variation and a multimodal distribution of scores. A possible solution to this problem will be discussed.

1. INTRODUCTION

Listening test methodologies have been greatly improved over the last two decades. Nevertheless, there are still some aspects of modern listening tests that need to be enhanced. In the opinion of this author, one of the main factors introducing error to the results of the modern listening test is attributed to biases involved in hedonic judgments. It will be argued in this paper that hedonic judgments should be avoided in listening tests in order to achieve high reliability and repeatability of conducted experiments.

Toole suggested that audio quality consists of only two major dimensions: pleasantness and fidelity [1]. This implies that a process of sound quality evaluation involves a combination of hedonic judgments (related to pleasantness) and sensory judgments (related to fidelity). The importance of the distinction between hedonic judgements and sensory judgments in listening tests has been emphasised in a number of papers in the audio engineering literature [2], [3]. The aim of sensory judgments is to evaluate sound character, for example loudness, pitch, timbre, angle of sound incidence, sound width, spatial envelopment etc. In hedonic judgments, by contrast, participants are asked about their likes, dislikes, preferences and desires. In Section 2 of this paper it will be shown that hedonic judgments are prone to many non-acoustic biases like situational context, expectations and mood, just to mention a few. Since audio quality evaluation involves, to some degree, hedonic judgments, it may be argued that the scores obtained from typical audio quality listening tests can be biased if the non-acoustical factors are not properly controlled.

In contrast to hedonic judgments, sensory judgments are less prone to non-acoustical biases, and, if appropriate acoustical anchors are used, can provide highly reliable and repeatable results. However, when listeners are asked to judge multidimensional attributes, like the overall spatial audio fidelity, the resultant data may exhibit a large variation and a multimodal distribution of scores (see Section 4). This phenomenon is largely caused by the fact that listeners use different decision criteria when “weighting” the low-level audio attributes.

The paper is organized as follows. The next section demonstrates typical biases involved in hedonic judgments, primarily based on evidence found in food sciences literature. Sections 3 and 4 will describe possible implications of these biases in experiments concerned with overall audio quality and spatial audio quality respectively. Proposed solutions aiming to improve current methodologies of audio quality evaluation are summarized in Sections 6 and 7.

2. BIASES IN HEDONIC JUDGEMENTS

According to Koster [4], it is difficult to design an experiment involving hedonic judgments that leads to conclusive and meaningful results. He identifies a number of problems with hedonic judgments that will be briefly discussed below.

2.1. Between-Subject Inconsistency

It is often implicitly assumed that data acquired in subjective tests have a unimodal distribution and therefore it is legitimate to average the results across the listeners in order to find the average score representing the "mean opinion" of all participants. However, experiments involving hedonic judgments often yield results that exhibit a bimodal or even multimodal distribution caused by the fact that participants are not homogenous in terms of their affective judgments (what one person likes may be disliked by another person). If the experimental data show a bimodal or multimodal distribution, it may not be legitimate to average the data across all participants but it may be necessary to employ some form of segmentation of subjects in order to identify groups of participants with similar scores.

2.2. Within-Subject Inconsistency

In subjective tests it is often assumed that participants act as "meters" that are relatively stable in their calibration over the duration of the experiment and hence yield relatively repeatable and consistent "measurement results". However, there is some evidence that when a subject is asked to evaluate a stimulus using hedonic judgments, his/her scores can vary substantially over time. For example, emotive attitudes of participants towards different objects under evaluation can be changed by a new fashion, peer pressure, by watching a TV advertisement etc. Koster claims that people change in time, especially when hedonic scales are involved. He says [4]:

> **"In fact, it can easily be shown that changes in preference and choice do take place and even to a degree that casts serious doubts on the predictive validity of hedonic and consumer studies that rely on single measurement sessions."**

According to Koster, up to 50% of participants can change their mind over the period of the experiment itself. Consequently, it may be questionable whether it is legitimate to extrapolate the results from a single test involving affective judgments into a long-term future or even whether it is possible to draw any meaningful conclusions from such tests.

2.3. Discrepancy between "Words and Actions"

It is widely known that there is some discrepancy between what people claim they like or prefer compared to what their actual behaviour reveals about their likes and preferences. Hence, experiments based on hedonic judgments may not lead to reliable observations. Studies involving observation of people's behaviour, rather than hedonic-oriented experiments, may render a more accurate picture about what people like or prefer.

2.4. Context Dependency

Food scientists observed that hedonic judgments may change depending on the situational context. For example, some food or beverage products are more liked in a restaurant than in a home setting. In other words, the same product may fit one situation and not another. This may imply that for a given sound stimulus, its quality may be evaluated differently depending on the situational context. For example, sound quality of a recording with certain type of spatial distortions may be unacceptable for listening at home but can be tolerable when listened to in a car.

2.5. Expectation Dependency

Another important factor that affects the hedonic judgments is expectation. For a given object under evaluation, participants give different scores depending whether the object meets their expectations or not (they will like the objects which meet their expectations and dislike any object which departs from their internal standard of expectation). Moreover, subjects can be biased by their expectation due to factors such as visual appearance, price and branding.

2.6. Hedonic Judgments and Mood

It is important to distinguish between an emotion and mood, the former being a specific reaction to a stimulus, whereas the latter is a general "background" feeling. Both emotions and mood can have some effect on hedonic judgments. For example, Vastfjall and Kleiner [5] investigated the situation in which mood may affect the results of the audio quality evaluation. In their experiments, participants were asked to evaluate audio quality using an annoyance scale, which can be considered as a special case of a hedonic scale. According to their results the subjects' mood biased the results of evaluation of audio quality by as much as 40%.

3. IMPLICATIONS FOR BASIC AUDIO QUALITY EVALUATION

The previous section showed that hedonic judgments are inevitably prone to a number of possible biases that cannot be neglected during the experimental design. This gives rise to a question about whether the same biases are involved in audio quality evaluation methods and if so, to what extent. In order to answer this question it is necessary to establish whether audio quality evaluation involves sensory judgments or hedonic judgments or both. If audio quality involves, at least to a degree, hedonic judgments, it would mean that all the biases discussed above may potentially affect the results of the audio quality tests.

Fig. 1 shows a hypothetical hierarchy of selected attributes that could be used for evaluation of audio quality. On top of this hierarchy there are purely hedonic attributes related to likes, dislikes and preferences. At the bottom of the hierarchy there are a number of low-level attributes (preferably unidimensional) that can be used to evaluate solely sound character. These emotion-free, low-level attributes are only related to sensory judgments (they do not involve any hedonic judgments). The layer above the low-level attributes consists of high-level multidimensional attributes, for instance timbral fidelity and spatial fidelity. Basic Audio Quality is placed between the top layer containing hedonic attributes (likes, dislikes, preferences) and the layer of high-level attributes (timbral and spatial fidelities). According to this model, basic audio quality comprises a combination of both sensory judgments as well as hedonic judgments.

There are a number of factors supporting the correctness of this hierarchical model, indicating that the evaluation of audio quality involves a combination of sensory and hedonic judgments. The early definition of audio quality could be paraphrased as a measure of “goodness of fit” of a perceived sound relative to one’s expectation or as a measure of “satisfaction”. For example, in 1989 Letowski defined sound quality as **“that assessment of auditory image in terms of which the listener can express satisfaction or dissatisfaction with that image”** [2]. A few years later, Blauert defined audio quality as **“adequacy of a sound in the context of a specific technical goal and/or task”** [6]. According to a more recent proposal by Jekosch in 2004, sound quality was defined as the **“result of an assessment of the perceived auditory nature of a sound with respect to its desired nature”** [7]. On the basis of these definitions it can be concluded that evaluation of sound quality does not involve only sensory judgments of sound character but also, to some extent, hedonic judgments of “satisfaction” or of sound “adequacy”. Without involving some form of hedonic judgment in audio quality evaluation, neither “adequacy of a sound” nor its “desired nature” can be assessed. Hence, it seems to be legitimate to assert that the evaluation of audio quality does not involve only one process (cognitive only), but two processes - cognitive and affective, as it was also pointed out by Vastfjall and Kleiner [5]. Thus, having established that some form of hedonic judgment is involved in

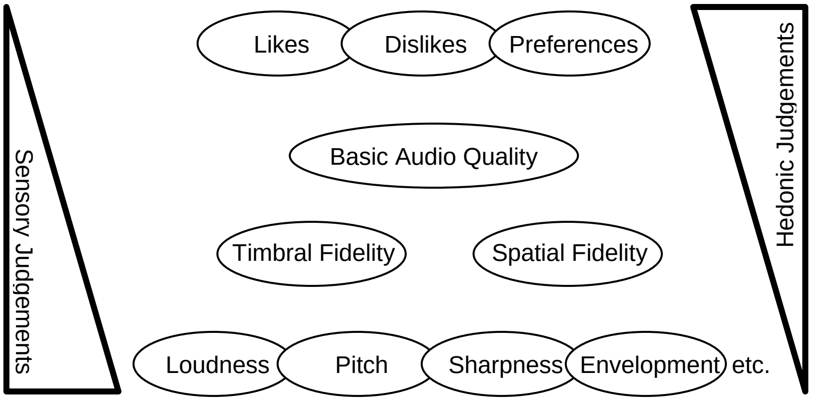

Fig. 1. Hypothetical hierarchy of audio quality attributes

evaluation of audio quality, it is possible to conclude that all the biases involved in hedonic judgments discussed above may also influence the results of basic audio quality evaluation.

One may argue that the two currently most popular methods for evaluation of audio quality [8], [9] are free from the aforementioned biases, as they use an emotion-free definition of audio quality which is substantially different from the definitions quoted above. According to both standards, the basic audio quality is defined as a **single, global attribute used to judge any and all detected differences between the reference and the object**. This definition does not make any references to the "satisfaction", "adequacy" or "desired nature" of a sound but to the perceptual "difference" between the audio reference and the object under evaluation. Since the perceptual "difference" can be considered as an emotion-free attribute, one could conclude that in these two standardised methods there is no place for any hedonic judgments. However, a close examination of the grading scales used in these standard techniques reveals that this conclusion is flawed. According to the ITU-R BS. 1116 recommendation, a 5-point impairment scale should be used in listening tests involving small audio quality impairments [8]. It can be seen in Fig. 2 that the two ends of the scale do not contain bipolar labels, as the top end of the scale is concerned with imperceptibility of impairments whereas the middle and bottom parts of the scale are used to represent different levels of annoyance. In other words, this scale can be described as a "hybrid", combining two different perceptual constructs at two ends of the scale; perceivability at the top and annoyance at the bottom. Since the "annoyance" construct is directly related to disliking, it can be inferred that the middle and bottom part of the scale will involve a substantial proportion of hedonic judgments. Hence, all the biases discussed in the previous section can potentially affect the results obtained using the ITU-R BS. 1116 recommended method.

Fig. 3 shows a five-interval quality scale included in the so-called MUSHRA recommendation (ITU-R BS. 1534) [9]. It can be seen that the labels used to describe the intervals have a hedonic nature. Hence, the results obtained using the tests involving this scale can also be affected by the biases discussed in the previous section.

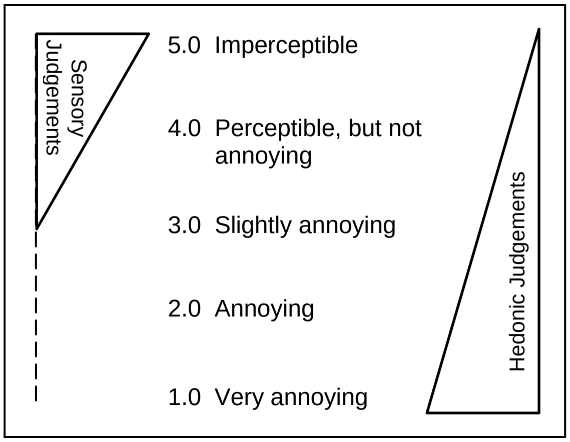

Fig. 2 Impairment scale according to the ITU-R BS.1116 Recommendation [8].

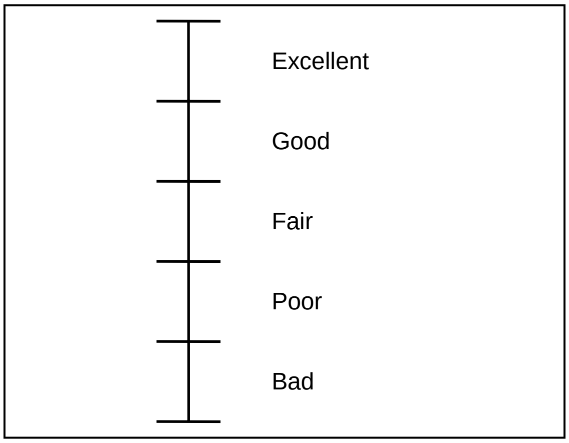

Fig. 3 Continuous quality scale according to the MUSHRA Recommendation [9].

3.1. Data Distribution Distortions

As mentioned in Section 2.1, the main problem related to hedonic judgment is a bimodal or even a multimodal distribution of data. This problem makes it difficult to establish the "mean" score representing a typical answer among the listeners unless a systematic segmentation of participants is undertaken. For example, Fig. 4 shows a result of the listening test where participants were asked to indicate their preference between two processed versions of the same recording. For both processed versions the overall bit-rate was identical but in the first case (recording A) the reduction in bit-rate was achieved predominantly at the expense of the timbral fidelity whereas in the second case (recording B) the reduction in bit-rate was achieved by "sacrificing" the spatial fidelity. The value of "-2" in Fig. 4 represents a "Strongly prefer recording A" category, grade "2" represents "Strongly prefer recording B", whereas "0" corresponds to the category of "Neither prefer A nor B". The mean value of the scores obtained from all participants was equal to 0.03, which could misleadingly indicate that the listeners neither preferred recording A nor B. However, according to Fig. 4 it is clear that the listening panel consisted of two groups of participants (segments) having opposite preferences: one group preferred recording A whereas the second group preferred recording B. More details are provided in Rumsey *et al* [10].

Another example of a "problematic" distribution is presented in Fig. 5 and was obtained in the experiment undertaken by Beresford *et al* [11]. A group of thirty listeners was asked to listen to a novel type of a multichannel audio classical music recording and to express their opinion as to whether they would be willing to purchase it. The listeners used a 9-point Likert scale where "-4" corresponded to the "Strongly disagree" category whereas "4" represented the "Strongly agree" category. Since the willingness to purchase a recording is strongly correlated to how much people like it, one may argue that in this particular case the audio evaluation process involved predominantly hedonic judgments. Hence, according to the discussion included in Section 2.1, the data might exhibit a multimodal distribution. This supposition was confirmed by the obtained results which clearly indicated a multimodal distribution (see Fig. 5). By contrast, for the same recording the distribution of data obtained where the listeners were asked to judge the sound character (sensory judgments) is predominantly unimodal, which supports the hypothesis that sensory judgments are less affected by the between-subject variability than the hedonic judgments.

Another example supporting the hypothesis that hedonic judgments typically yield a "problematic" distribution of data is given by Beidl and Stucklschwaiger [12] who observed a bimodal distribution of scores resulting from paired comparison of auditory stimuli. In their experiments, participants were asked to listen to pairs of car audio noises and express their preference using a hedonic scale. According to the obtained results, one group of participants preferred quieter noises whereas another group preferred louder noises claiming that "the higher the speed, the more powerful, the more sporty, the more dynamic and, therefore, better".

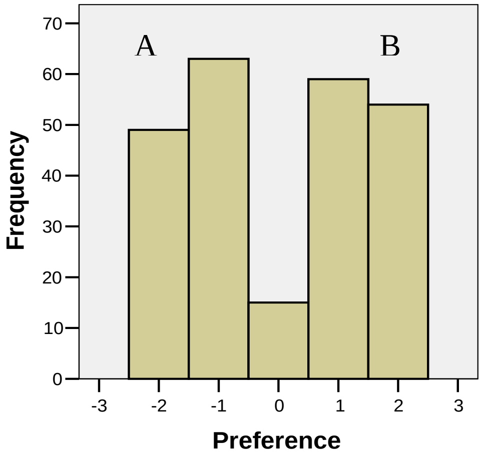

Fig. 4 Example of a bimodal distribution

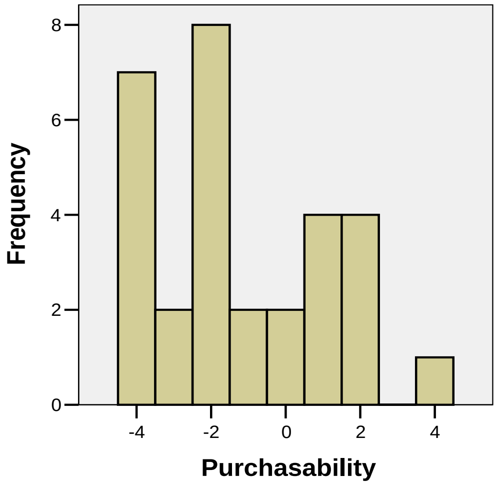

Fig. 5 Example of a multimodal distribution (taken from [11])

3.2. Bias Due to Situational Context

As it was mentioned in Section 2.5, the situational context may have a significant effect on the results of hedonic judgments. Although some authors seem to support this hypothesis (e.g. Toole [13]), there is no direct data available in its support. On the contrary, there is some evidence contradicting this hypothesis. For example, Gros and Chateau [14] showed that the context of environmental or visual cues has a very weak influence on audio quality, although it might be difficult to generalise their finding as their experiment was concerned with speech signals.

3.3. Bias Due to Expectation

In 1994 Toole and Olive demonstrated that both experienced and inexperienced listeners were prone to non-acoustical factors such as visual cues and "product identity" when they were they were rating their preferences of loudspeakers (hedonic judgments) [15]. Their paper is well known in the audio engineering community and often quoted as an example of how important it is to undertake blind listening tests to reduce any non-acoustic biases.

Another example, perhaps less known, of how expectation can affect evaluation of audio quality is given by Bentler *et al*. [16]. In their experiment a group of listeners were asked to assess the audio quality of the same type of a hearing aid, labelled either as "digital" or "conventional". They found that the effect of labelling accounted for between 2% and 32% of the variance in individual outcome measures. Like Toole and Olive, Bentel *at al* also emphasised the importance of undertaking blind listening tests in order to minimise any non-acoustic biases.

A more recent example of how expectation of listeners can affect the results of audio quality evaluation is given by Vastfjall [17]. In his experiment the expectation of the participants was controlled by asking them to read different consumer reports. In the listening test participants were using a hedonic scale. It was found that people who had low expectations on average rated the sounds as less annoying than people who had high expectations. The magnitude of the effect was of the order of 20%.

3.4. Bias of Unconscious Choice

As it was discussed in Section 2.3 above, there is some psychological evidence that what people actually do or choose could be different from what they indicate in subjective tests. No evidence of this bias in audio quality evaluation tests has been found so far. However, if this bias proved to play some role in audio quality tests, experiments involving hedonic judgments would have to be re-designed in order to minimise it. For example, instead of asking people direct questions like "Which sound do you prefer, A or B?", we could simply ask people to listen to either of these sound for a relatively long period of time and check which they really preferred to listen to. It is believed that listeners' behaviour would reveal a more accurate picture of their preferences than a simple preference tests. If this approach is successful, this could form a new listening test paradigm.

3.5. Bias Due to Mood

As mentioned in Section 2.6, Vastfjall and Kleiner [5] investigated the effect of mood on audio quality evaluation. In their experiments participants were asked to evaluate audio quality using the annoyance scale, which could be considered as a special case of a hedonic scale. According to their results, mood biased the evaluation of audio quality by as much as 40%. Moreover, in a more recent experiment Vastfjall observed that listeners who had a positive attitude judged the pleasantness of sound about 20% higher than people who had a negative attitude. In addition, it was found that those listeners who were annoyed evaluated sounds about 30% higher on the annoyance scale compared to the listeners in neutral mood [17]. These examples illustrate how hedonic judgments of audio quality are "prone" to non-acoustic factors.

4. IMPLICATIONS FOR SPATIAL AUDIO EVALUATION

It is believed that all the biases affecting hedonic judgments can manifest themselves in experiments involving evaluation of spatial audio quality. In fact, Fig. 4 and Fig. 5 discussed above were obtained from experiments concerned with spatial audio. It is expected that evaluation of sound character of spatial audio (e.g. envelopment, angle of incidence of audio sources, source width, and frontal spatial fidelity) will be much less biased than affective evaluation of spatial audio quality (e.g. preference, liking, desires). The experimental results obtained by Rumsey [18] seem to support this assertion. In his study he achieved conclusive results for front image quality, however the results obtained for spatial impression and for listener preference (the latter is a standard example of a hedonic judgment) were not so definite due to distortions in data distribution.

As it was asserted above, it is expected that for sensory judgments of spatial audio the biases discussed will be less prominent than in the case of hedonic judgments. However, this does not guarantee that the distribution of data will be unimodal. On the contrary, the data may still exhibit, to a degree, a bimodal or even a multimodal distribution. It must be remembered that some of the attributes used in spatial audio experiments are multidimensional in nature and hence are difficult to evaluate. Consequently, the listeners have to undertake some internal "weighting" in order to produce the "overall" score for a number of multidimensional attributes. Depending on the importance listeners attach to the individual sub-attributes, the resultant scores can differ substantially between the participants, potentially leading to multimodal distribution. For example, when listeners are asked to evaluate the overall spatial fidelity of a large set of spatially distorted audio recordings, some people may prioritise accuracy of the frontal image over the importance of envelopment, whereas some other listeners can attach the greatest "weight" to the envelopment. This, obviously, will give rise to the inter-subject discrepancy in the data. For this reason some spatial audio attributes, like overall spatial fidelity, are not easy to evaluate as it is not always straightforward to decide how to "prioritise" different sub-attributes of the overall spatial fidelity during the evaluation process.

5. IMPLICATIONS FOR DEVELOPMENT OF AUDIO QUALITY EXPERT SYSTEMS

Current methods for objective audio quality prediction, e.g. [19], are only concerned with the physical characteristics of audio recordings under test. In other words, they operate solely in the physical domain and do not take into consideration any non-acoustical information. However, according to Blauert and Jeckosh [6] **"sound-quality is not an inherent property of the product, but rather something which develops when listeners are auditorily exposed to the product and judge it with respect to their desires and/or expectations in a given situational context."**

In other words, sound quality is not only "rendered" by a listener depending on the auditory stimuli but also based on how well or badly a particular stimulus meets a listener's expectations for a given situational and emotional context. Consequently, an "artificial intelligence" approach is not sufficient for successful development of objective methods for audio quality evaluation as one also needs an "affective intelligence" in order to be able to predict audio quality accurately, which was pointed out by Vastfjall and Kleiner [5]. Another possibility is to "give up" with objective audio quality predictors, and concentrate on the development of objective audio fidelity predictors (the latter would be concerned with trueness of stimuli with respect to a reference). For example, new algorithms could be developed to predict timbral fidelity, spatial audio fidelity, surround spatial fidelity, localisation fidelity etc. As these predictors do not depend on non-acoustical information, there would be a higher chance of success in the development of the algorithms for automatic prediction of these attributes based solely on acoustical inputs.

6. PROPOSED SOLUTIONS

It was shown that hedonic judgments may introduce more bias to the results of audio quality listening tests than sensory judgments. Consequently, hedonic judgments should be avoided in audio listening tests if possible. For instance, the participants could be asked to evaluate sound character or audio fidelity (trueness with respect to a reference) rather than how much they like, dislike, prefer or desire certain audio stimuli. Hedonic scales should be avoided wherever possible. For example, the hybrid (sensory/hedonic) scale recommended by the ITU-R BS. 1116 standard should be used with caution or could be replaced by a "pure" sensory scale describing the magnitude of perceptual differences between the reference and the object under evaluation. Carefully calibrated anchors could improve the repeatability of the listening tests even further as they reduce centring and contraction biases [20].

If hedonic judgments cannot be avoided, special care should be taken during the experimental design to control any non-acoustic factors that could potentially bias the results. For example, a situational context could be controlled by explaining the context of the experiment in the instructions for the listeners (the listeners could be even provided with a story setting the "scene" for the audio evaluation). Extra data should be acquired from the listeners prior to the listening test, which could be used after the experiment to undertake the segmentation of participants if the data exhibits a multimodal distribution. For example, the listeners can be asked about their listening habits and preferences, personal and professional background, type of equipment they use, etc. These data could be mapped onto clusters of scores during the subject segmentation procedure and could help to explain why people vary in their hedonic judgments.

In the case of spatial audio fidelity, a pilot test can be undertaken prior to a proper listening test in order to check whether the data exhibits any distribution distortions. If this is the case, using low-level attributes instead of high-level ones might help (see Fig. 1). For example, if listeners are asked to evaluate the overall spatial fidelity and if the resultant data exhibit multimodal distribution, it might be necessary to ask participants to evaluate a number of low-level attributes instead.

Since, as pointed above, sound quality is not an inherent property of the product or of the sound itself but depends on subject's expectation, mood, situational context, just to mention a few factors, an "artificial intelligence" approach is not sufficient for successful development of objective methods for audio quality evaluation as one also needs an "affective intelligence" in order to be able to predict audio quality accurately. Alternatively, the reliability of objective methods could be increased by predicting only scores obtained from sensory judgments, like timbral fidelity scores, spatial fidelity scores or envelopment scores.

7. SUMMARY

This paper describes a number of potential biases related to hedonic judgments that can affect both basic audio quality as well as spatial audio quality evaluation. It was shown that experiments involving hedonic judgments are particularly prone to many non-acoustic biases like mood, expectation, inter-personal differences, situational context and the problem of unconscious choices. The data obtained from experiments involving hedonic judgments are usually distorted by inter-subject differences and exhibits a bimodal or multimodal distribution. By contrast, the experiments involving sensory judgments are less affected by these biases and the data obtained in such experiments normally exhibits unimodal distribution, unless the listeners have to evaluate sound character using high-level multidimensional attributes involving some "trading" between changes in low-level characteristics of sound. A number of solutions were identified that could help to reduce the biases described in the paper.

8. ACKNOWLEDGMENT

The author would like to express his gratitude to Francis Rumsey and Kathryn Beresford for their comments on the initial draft of this paper.

9. REFERENCES

[1] F. Toole, "Subjective Measurements of Loudspeaker Sound Quality and Listener Performance," *J. Audio Eng. Soc.*, vol. 33 (1/2), pp. 2-32 (1985).

[2] T. Letowski, "Sound Quality Assessment: Concepts and Criteria," presented at the AES 87th Convention, October 13-21, New York, Paper 2825 (1989).

[3] F. Rumsey, "Spatial quality evaluation for reproduced sound: terminology, meaning and a scene-based paradigm," *J. Audio Eng. Soc.*, vol. 50, pp. 651-66 (2002).

[4] E.P. Koster, "The psychology of food choice: some often encountered fallacies," *Food Quality and Preference*, vol. 14 (5-6), pp. 359-373 (2003).

[5] D. Vastfjall and M. Kleiner, "Emotion in Product Sound Design," Proceedings of Journees Design Sonore, Paris (2002).

[6] J. Blauert and U. Jekosch, "Sound Quality - Evaluation – A Multi-Layered Problem," *Acustica united with Acta Acustica*, vol. 83, pp. 747-753 (1997)

[7] U. Jekosch, "Basic Concepts and Terms of 'Quality', Reconsidered in the Context of Product-Sound Quality," *Acustica united with Acta Acustica*, vol. 90, pp. 999-1006 (2004)

[8] ITU-R Recommendation BS. 1116: "Methods for subjective assessment of small impairments in audio systems including multichannel sound systems," International Telecommunications Union (1994).

[9] ITU-R Recommendation BS. 1534, "Method for the subjective assessment of intermediate audio quality," (MUSHRA), International Telecommunications Union (2001).

[10] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences," *J. Acoust. Soc. Am.*, vol. 117 (6), pp. 3832-3840 (June, 2005).

[11] K. Beresford, F. Rumsey and S. Zielinski, "Listener Opinions of Novel Spatial Audio Scenes," to be presented at the AES 120th Convention, Paris, May 20-23 (2006).

[12] C.V. Beidl and W. Stucklschwaiger, "Application of the AVL-Annoyance Index for Engine Noise Quality," *Acustica united with Acta Acustica*, vol. 83, pp. 789-795 (1997).

[13] F. Toole, "Listening Tests – Turning Opinion into Fact," *J. Audio Eng. Soc.*, vol. 30 (6), pp. 431-445 (1982).

[14] L. Gros and N. Chateau, "The impact of listening and conversational situation on speech perceived quality for time-varying impairments," Measurement of Speech and Audio Quality in Networks. On-line Workshop:

http: // wireless. feld. cvut. cz/ mesaqin2002 /contributions.html (2002).

[15] F. Toole and S. Olive, "Hearing is Believing vs. Believing is Hearing: Blind vs. Sighted Listening Tests, and Other Interesting Things," presented at the 97th AES Convention, Paper 3894, San Francisco, November 10-13 (1994).

[16] R. Bentler, D., Niebuhr, T. Johnson, and G. Flamme, "Impact of digital labeling on outcome measures," *Ear and Hearing*, vol. 24, pp. 215-224 (2003).

[17] D. Vastfjall, "Contextual Influences on Sound Quality Evaluation," *Acustica united with Acta Acustica*, vol. 90, pp. 1029-1036 (2004).

[18] F. Rumsey, "Controlled subjective assessment of two-to-five channel surround sound processing algorithms," *J. Audio Eng. Soc.*, vol. 47, pp. 536-581 (1999).

[19] ITU-R Recommendation BS. 1387, "Method for Objective Measurements of Perceived Audio Quality," International Telecommunications Union (1998)/

[20] E.C. Poulton, "*Bias in quantifying judgements,*" Laweence Erlbaum Associates, London (1989).