What are the requirements of a listening panel for evaluating spatial audio quality?

Nick Zacharov[1], and Gaëtan Lorho[2]

[1]*Nokia Corporation, Technology Platforms, Visiokatu 1, FI-33721 Tampere, Finland*

[2]*Nokia Corporation, Technology Platforms, P.O. Box 407, FIN-00045 NOKIA GROUP, Finland*

Correspondence should be addressed to Nick Zacharov (`nick.zacharov@nokia.com`)

ABSTRACT

This paper addresses the topic of subject and listening panel performance in descriptive analysis applications. A summary of the status in the field of audio is provided regarding the selection, development, training and monitoring of subjects and listening panel. Subject categorisation and associated terminologies are discussed, placing emphasis on the difference between affective and descriptive tasks. A number of statistical analysis methods are presented that can be applied to the evaluation of subject and a panel's performance in descriptive analysis tasks which are supported by some illustrative examples. Additionally, the application of these analysis methods to spatial sound evaluation tasks is discussed.

1. INTRODUCTION

As audio systems and spatial sound systems continue to develop there is an increased need and interest to find appropriate means to evaluate their perceptual characteristics. Whilst global evaluation of audio quality such as the so-called mean opinion score, listening quality or basic audio quality have been popular in the field of audio these are found increasingly to be insufficient to describe the detailed perceptual nature of such complex systems and stimuli. As a result, researchers are turning to descriptive analysis techniques to provide a more detailed and objective perceptual evaluation. Over the last few years a number of researchers in the field of audio have applied a range of descriptive analysis methods to several different product categories including mobile phones, loudspeaker and headphone spatial sound reproduction systems and so forth.

Sensory evaluation is dependent on the performance of the subjects performing the evaluations. For the sensory profiles to be meaningful, subjects, forming a panel, should be as objective as possible and in agreement on the usage of a consensus descriptive vocabulary. It is this matter of subject reliability in descriptive analysis tasks that forms the core of this study.

The development of an appropriate set of attributes for the description and discrimination of a given set of stimuli, such as those found in spatial sound reproduction, is a challenging problem. Methods based on consensus descriptive analysis are known to require a lot of time and effort. Panel training might range from 10 hours for a simple stimulus set (e.g. Stone & Sidel [75]) to 120 hours for more complex stimuli (e.g. Meilgaard et al. [54]). As an example in the field of audio, the descriptive analysis experiment reported by Zacharov & Koivuniemi [82, 40, 81] employed a panel of 12 assessors who developed a set of 12 attributes to describe the characteristics of different spatial sound reproduction systems for a range of program recordings in 30 hours. The reliability of a descriptive analysis (DA) panel is therefore an aspect that should deserve significant attention, considering the efforts it requires. Panel performance monitoring can also be a very useful tool when used to identify and attempt to fix potential problems with attributes or assessors during the vocabulary development process. It should also be mentioned that communication of performance to assessors by the panel leader is an important aspect of the panel training. This feedback to panellists can have a positive effect on training and motivation and thus on the resulting quality of the panel (Meilgaard et al. [54], Findlay et al. [23]).

The following section will present a short set of terms of reference to clarify the context and application domain of the performance measures to be discussed in Section 4. Section 3 will provide a short overview of descriptive analysis in order to clarify the different types of performance evaluation stages. Section 4 presents a number of assessor evaluation methods with examples, leading to a short discussion and conclusion of the paper.

2. TERMS OF REFERENCE

2.1. Affective versus descriptive

The matter of whether a task is affective/hedonic or descriptive is an important differentiation that has significant impact. Perhaps the most significant influence is associated with the type of subject to be employed in a test. Scriven [72] discusses this topic in relation to different type of panels and starts with a description of the two stages of human response to stimuli as follows:

1. The primary response is to recognise and measure the stimulus (Descriptive)
2. The secondary response is to form a judgment about what is perceived, e.g. liking or acceptability judgements (Affective/Hedonic)

Scriven's model is illustrated in relation to panels in Table 1, which is found to be similar to the filter model concept introduced by Pedersen and Fog [64] and refined by Bech and Zacharov [5].

Typically, naïve subjects are only aware of the second response. Thus it is viewed that obtaining affective measures such as liking, preference, acceptance is best achieved with naïve subjects. Such subjects are not conscious of the primary response, which is typically developed through training and awareness. It should also be noted that according to the assessor definitions provided by ISO 8586-2 [34], subjects typically develop from being naïve to being initiated assessors, and so on, in a continuous manner (see Figure 1).

In the context of sensory evaluation, the experimenter is typically only interested in performing objective assessment of attributes developed through a vocabulary development process. As stated by Blauert and Jetosch [13] "to obtain a sensory profile of a product such that the results show reasonable objectivity, trained and experienced panels of expert are, in fact, indispensable". Thus only the use of trained assessors will be considered in this paper

2.2. Individual versus consensus

In the domain of descriptive analysis a number of methods exist aimed at the same purpose, i.e. the development of descriptive vocabularies for the objective assessment of product characteristics. In the category of verbal and direct elicitation methods [5], two major routes exist for establishing the vocabulary, namely, consensus and individual vocabulary development procedures.

Consensus vocabulary development procedures include methods such as Quantitative Descriptive Analysis (QDA) [75] or Flavor Profile [54] for example. The main idea of these methods is to employ a panel of assessors to develop a set of common perceptual attributes to describe the sensory properties of the stimuli under investigation. Standard methodologies have been developed for this type of consensus vocabulary development process [32] and this method has proved to be successful in the food industry.

Individual vocabulary development procedures include methods such as free choice profiling (FCP) [42, 79, 78, 50], repertory grid technique (RGT) [39, 7] and flash profile [74, 20] for example. In these methods, each assessor develops his or her own set of attributes, which removes the need for construct alignment between the assessors.

The example methods discussed in this paper and presented in Section 4.3 are predominantly applicable to consensus vocabularies. However some of these methods can also be considered for performance assessment of the case of individual vocabulary methods, e.g. for a given attribute and assessor, or for a given individual sensory profile. Methods have also been developed to derive an 'average' sensory profile from the individual datasets, and the resulting data structure can be exploited to some extent to assess the performance of the panel.

2.3. Subject categorisation

Whilst the terms untrained, naïve, experienced and expert are often employed in the audio literature to

Product measurement or trained panel	**Consumer panel**
Selected or trained subjects	Naïve subjects
Subjects selected to measure characteristics of products	Subjects selected to represent wider consumer population (target market)
Measure primary response to product as an indirect measure of ingredients or processes to extrapolate to whether differences between products are real or not Performed in sensory laboratory	Measure primary response to product as an indirect measure of ingredients or processes and measure secondary responses to extrapolate to what the total population might like Ecologically motivated test environment (real usage)
8 well trained assessors	50 - 100 assessors

Table 1: Comparison of trained versus consumer panels, adapted from Scriven [72].

describe the nature of a subjects performance, a review of a number of the standards and recommendations (e.g. [17, 35, 37, 38, 36]) will yield inconsistent definitions of these terms. This topic is discussed further by Bech and Zacharov in [5]. However, a well formulated summary of assessor categorisation has been developed in the field of agricultural food products and reported in several ISO standards discussed in the following section.

2.3.1. ISO assessor categorisation

The ISO has standarised a set of terms which are employed in the agricultural food industry to describe different kinds of assessors also commonly referred to in audio as subjects. This terminology is defined predominantly for application in descriptive analysis tasks, where objective assessment of attributes is the primary purpose.

Two key standards on this topic exist including ISO standard 8586-1 [31] and 8586-2 [34]. The former focus upon the selection, training and monitoring of selected assessors whilst the latter consider these aspects for experts. In order to clarify these meanings and terms, please refer to Table 2. Figure 1 also provides some information regarding how subjects can progress from being untrained assessors through to being expert assessors.

It is considered by the authors that this terminology is very clearly defined and can be applied to any field of sensory perception/evaluation irrespective of the specific nature of that field. It is suggested that this unambiguous terminology be adopted in the field of audio in order to clarify communication regarding assessor categorisation and their associated performance.

As a summary of the focus of this paper, it can be stated that this work relates to direct, verbal, descriptive, (mostly) consensus vocabulary application using quantitative scales with selected and trained assessors.

3. DESCRIPTIVE ANALYSIS OVERVIEW

Often when starting a descriptive analysis task from scratch, the vocabulary and attributes may not yet have been developed. In such cases a particular process is required to develop the vocabulary and attributes. However, once these exist, then they can be applied to the objective evaluation of products/stimuli. It is this latter application that we focus on here, as it is in the rating phase that the performance of the assessor is of greatest interest.

Assuming that the vocabulary already exists, a typical descriptive analysis experiment will follow the following 4 stages:

- Pre-selection
- Training
- Product evaluation
- Post-selection

The exact location of the product evaluation phase may also occur following post-selection. This case may occur when a permanent panel is established

Assessor category	**Definition**
Assessor	Any person taking part in a sensory test
Naïve assessor	A person who does not meet any particular criterion
Initiated assessor	A person who has already participated in a sensory test
Selected assessor	Assessor chosen for his/her ability to carry out a sensory test
Expert	In the general sense, a person who through knowledge or experience has competence to give an opinion in the field about which he/she is consulted. (Please note that the term *expert* does not provide any indication regarding the qualification or suitability of the individual to perform listening tests.)
Expert assessor	Selected assessor with a high degree of sensory sensitivity and experience in sensory methodology, who is able to make consistent and repeatable sensory assessments of various products
Specialised expert assessor	Expert assessor who has additional experience as a specialist in the product and/or process and/or marketing, and who is able to perform sensory analysis of the product and evaluate or predict effects of variations relating to raw materials, recipes, processing, storage, ageing, and so on.

Table 2: Summary of assessor categories employed in sensory analysis, as defined in ISO standard 8586-2 [34], applied to the food industry and recommended for adoption in the field of audio.

and the training - product evaluation - post-selection is a continuous process. Alternatively this may occur when post-selection is performed following a specific experiment to evaluate the assessor's performance as opposed to a product evaluation experiment.

Pre- and post- selection and training will be discussed in further detail in the following sections.

3.1. Pre- and post- selection

When considering panels of assessors, a number of different methods and selection criteria can be applied to evaluate and assess the suitability of the subjects. Broadly speaking these fall into two categories, namely pre- and post-selection procedures.

Pre-selection procedures comprise of those where subjects are evaluated for their *potential* of being a good listener, suitable for participating in (certain types of) listening tests or to become a members of a listening panel. Some examples of pre-selection procedures are described in [52, 29].

Post-selection considers the performance of listeners following a listening test to evaluate the *goodness* of their performance in this task. This goodness can be assessed in a large number of ways which will be discussed later in Section 4.

Whilst pre-selection provides a raw screening of subjects potential, post-selection provide concrete information regarding a subject's performance. It can be

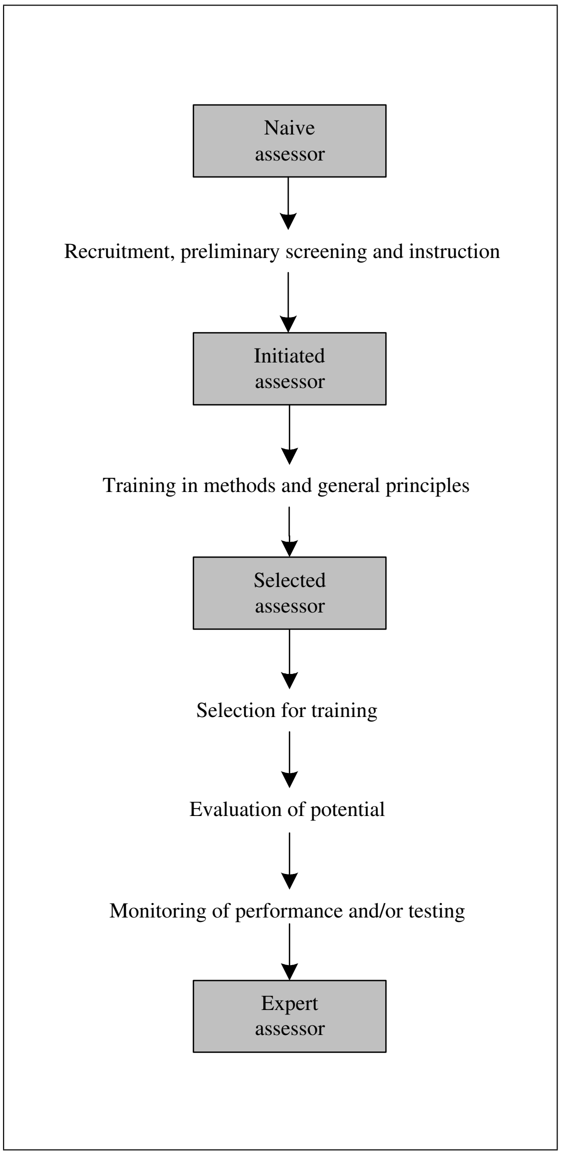

Naive assessor
Recruitment, preliminary screening and instruction
Initiated assessor
Training in methods and general principles
Selected assessor
Selection for training
Evaluation of potential
Monitoring of performance and/or testing
Expert assessor

Fig. 1: The process of sensory assessor development according to ISO 8586-2 [34].

considered that the process of post screening is quite expensive, as subjects are analysed following an experiments and if their performance is very poor their data may not be meaningful/valid. In such cases the subject may require further training or in the worst case not continue to perform listening tests with the panel. Whilst this is indeed a time consuming and expensive process, it is the only true way to establish the performance level of a subject and the panel as a whole. This is considered to be a good investment, as noisy or erroneous data is of little value to the experimenter. In order to manage the costs of post-selection at the early stages of establishing a panel, specific experiments can be defined that test the performance of the subjects. Such experiments may not yield meaningful experimental data beyond the evaluation of the subjects performance. This is safe, as it will ensure that no scientific or product related decisions will be made based on these finds, but the panel performance will be refined for its future use.

3.2. Training

Training is a key way to improve the performance of subjects and increase their expertise. Through listening to relevant sound stimuli that exemplify a certain set of auditory characteristics so subjects can explore and become more sensitive to the nature of these characteristics and their evaluation. Often subjects sensitivity to such characteristics will increase with exposure and hopefully also their ability to reliably rate their perceptions (see Bech [3]). However, it should be noted that not all subjects have the capacity to improve and furthermore not all subjects can become expert in all areas of perception.

Training programs take several different forms and are typically targeted to certain applications in audio. These are typically divided into two types:

- Training by listening only,
- Adaptive training with feedback.

A large number of training schemes exist a few of which are listed here as examples. However, presently, no extensive training schemes yet exist for training of spatial evaluation tasks.

Some examples of listening only training schemes can be found from [57] and [2]. A larger number of adaptive training schemes exist, including timbre solfege programme [46, 56] and timbral ear training (TET) software developed by Quesnel [67, 65, 66]. More specific training schemes for particular application or attributes may be found from the following sources [18, 62, 63, 73, 61, 60, 55].

The benefits of training subjects for descriptive analysis or sensory profiling is discussed in a number of cited papers. Labbe et al. [41] provide a short and concise summary of the benefits of training. It is stated that training

- allowed the assessors to become familiar with the vocabulary[1] and to use it reliably,
- completely modified the mean profiles and made them reliable,
- improved discrimination ability of the panel,
- simultaneously reduced the individual session effect and improved individual discrimination ability,
- improved the consensus within the panel.

With respect to this study, training can be view as a means to improve the performance of assessors and it a vital part of that process.

4. DESCRIPTIVE PANEL PERFORMANCE

In this section, an overview of the ideas and terminologies commonly employed for consensus panel assessment are presented. The performance criteria considered in this paper focus on the restricted case of a single sensory panel. Inter-panel performance assessment in which several panels are compared, or cases where external information such as preference data or instrumental measurements is present, require slightly different approaches, see e.g. McEwan et al. [53] but this is not covered in the present paper. A more general discussion of potential criteria for measuring panel performance and theoretical problems with some of them can be found in Wolters & Allchurch [80]. It is also important to remember that a DA panel should be seen as a human instrument from which we expect objective and reliable measurements. Also, a panel of assessors is employed rather than a single assessor to take into account the fact that human subjects are not equally sensitive to sensory stimuli, might also vary in their ability to discriminate different perceptual aspects of the stimuli and can be subject to judgment bias. In comparison to the profile produced by a single expert, the sensory profile obtained by a panel of assessors is considered to represent a more general and more stable description of the stimuli.

4.1. Terms of reference

Repeatability The *precision* of a measurement method is defined in e.g. ISO 5725-1 [33] as the closeness of agreement between mutually independent test results obtained under stipulated conditions. In the context of a consensus DA panel, this measure can be understood at the level of a single assessor or a panel level. These two cases highlight a different aspect of the sensory panel by considering either one component or the sum of the components of the sensory measurement tool. Precision is often stratified into the two different concepts of *repeatability* and *reproducibility.* ISO 5497 [30] defines repeatability as the closeness of agreement between mutually independent test results obtained under conditions where mutually independent test results are obtained with the same method on identical test material in the same laboratory by the same operator using the same equipment within short intervals of time. Reproducibility is defined as the closeness of agreement between test results obtained under conditions where test results are obtained with the same method on identical test material in different laboratories with different operators using different equipment. In the context of the performance assessment of a single sensory panel, repeatability is a more appropriate measure. However, another approach could also be adopted which consider assessors as different "laboratories". For example, Rossi [68] defined repeatability and reproducibility respectively as the abilities of an assessor to score the same product consistently for a given attribute and to score the products, on average, similarly to the other panel members. Reproducibility in this case relates more to the concept of accuracy defined next.

Agreement *Accuracy* is defined in e.g. ISO 5725-1 [33] as the closeness of agreement between a test

[1] In this study a glossary of 17 pre-defined attributes were employed.

result and the accepted reference value. Considering that only data from a single sensory panel is available in our case, we can interpret the test result as the set of scores given by one assessor and the accepted reference value as the set of "mean" scores over the panel[2]. In practice, an assessor can be very precise but not accurate, which means that this assessor disagree with the rest of the panel. The term *agreement* (or *disagreement*) is commonly used in sensory science literature to describe accuracy. However, the term *proficiency* is also found to express the accuracy of either a given assessor or the full panel. It should be noted at this point that repeated measurements of the same samples are necessary to measure precision and that disagreement can not be separated out from repeatability if the sensory profiling test does not include repetitions.

Discrimination The concepts of precision and accuracy would be sufficient for performance measurement if a true external reference was available for comparison with the test result but this is rarely the case in sensory science. In our specific case of a single panel, the "internal" accuracy we measure has no absolute meaning and there is always a risk to obtain precise and accurate results from a panel without the required level of discrimination. The concept of discrimination relates to the ability of an assessor or a panel to make a perceptual distinction between several stimuli with a given attribute or a set of attributes. The discrimination of a panel depends on assessor reliability and panel agreement. These performance measures are nested in the sense that no discrimination will be achieved if assessors show a poor reliability, but even when assessors are discriminative on an individual level, discrimination between products might still remain poor at a panel level if a large disagreement exists between the different panellists.

Multivariate sensory information: product discrimination and attribute redundancy.

A clear distinction appeared in the previous paragraphs between the performance of a given assessor and the performance of the panel. Another particularity of sensory profiling relates to the difference between the measurement of performance at an attribute level and at a complete (or partial) sensory profile level. Whereas univariate methods can be used to compare means for each attribute separately; multivariate methods aim at comparing product positions in the sensory space and allow for the assessment of product discrimination in terms of both strength (distance between products in the sensory map) and complexity (dimensionality of the sensory space). Another aspect of multivariate analysis of sensory profiles concerns the measurement of redundancy in attributes. Redundancy is employed here instead of correlation to highlight the importance of meaning rather than statistical relationship, as described in Wolters and Allchurch [80]. Looking at panel performance from an efficiency point of view, we could test the appropriateness of the number of attributes employed for the sensory profile. In pragmatic terms, we would be looking for a sensory profile with an optimal balance between a too large and a too small number of attributes. The former case will imply a larger experimental effort and might result in a redundancy in the sensory profile; the later case could result in low product discrimination or could fail to capture some aspects of the sensory differences between the products. However, this topic is not well documented in the literature, to the knowledge of the authors.

4.2. Attribute scale usage and attribute interpretation

A sensory score is the result of a complex process including physiological, psychological, and grading tasks. The assessor scoring on a linear continuous scale is the last step of this measurement process and can be affected by individual variations relating to scale usage. Such variations can be very large but have usually no relation to product characteristics. They are often considered as nuisance effects

[2] It should be noted that the concept of "true value" versus "expected value" is problematical in sensory science as it depends on the presence and nature of an external reference, e.g. the accuracy of a given panel can only be assessed in the context of a larger experiment including several sensory panels (McEwan et al. [53])

by the experimenter and methods to correct scores from these individual differences have been developed in sensometrics (Næs & Solheim [59], Brockhoff & Skovgaard [15]). Brockhoff [14] described four basic assessor differences that can be encountered on a univariate scale, as illustrated in **Figure 1**:

1. Level difference: Assessors use different parts of the scale, which is reflected by a different mean score over all products for a given attribute.
2. Scaling difference: Assessors use different amount of the scale, which is reflected by a different variance over all products for a given attribute.
3. Variability: The precision of assessors differ, which is reflected by a different variance over the repetitions of a given product and attribute.
4. Disagreement: the non-linear individual variation not attributable to the scaling differences previously described.

A set of formal univariate statistical models was proposed by Brockhoff to derive these four assessor effects, which gives useful information about panellist differences in scale usage for each attribute.

Another aspect of individual variation in attribute scale usage concerns the issue of disagreement on the sensory concept associated to each descriptor. Panel training can solve this problem to some extent but a perfect concept alignment of the assessors is usually not feasible even with extensive training. Examples of attribute misinterpretation can include the possible confusion between two perceptually related attributes or the use of a combination of these attributes to assess a certain perceptual aspect of the stimuli. Multivariate analysis methods have been exploited to address such issues with the idea of a simulated individual vocabulary profile to test the validity of consensus use of the attribute by the panel [22].

4.3. A review of panel performance assessment methods

The topic of sensory panel performance assessment is largely documented in the literature but numerous research studies continue to be published in the field of sensometrics, which illustrate the complexity of this topic. Very different methods have been presented with performance criteria ranging from relatively simple statistics (see Rossi [68] or Næs [59]) to more advanced statistical methods employing univariate models (Brockhoff [14]) or multivariate models (see Sclich et al. [71]). A short literature review is presented in this section to illustrate how different approaches focus on different aspects of the panel performance.

As described in the previous section, sensory data can be regarded from a univariate (uniscalar) or a multivariate (multi-scalar) point of view. Panel assessment on an individual attribute level provides very detailed information about the performance of the panel. For example, Rossi [68] used two measures relating to the repeatability and disagreement criteria defined in the previous section. These measures are computed per product, assessor and attribute (the second criterion is referred to as reproducibility in the original paper) and are based on a statistical model defined by Mandel [49]. This approach results in a large set of descriptive statistics, which Rossi summarised with several graphical techniques. Another statistical method based on classical reliability theory was used by Bi [12] to compute the same criteria. However, these two statistics are only computed per attribute and assessor, which yields two graphical displays only. Similar visual representations of panel performance have been presented for the reliability and discrimination criteria defined in the previous section, e.g. by Næs and Solheim [59] or Lea et al. [44] employing statistical techniques based on ANOVA[3]. Schlich [69] proposed a method for graphical representations of assessor performances based on individual and global ANOVAs for each attribute separately. This method described in more details in the next session covers three of the performance criteria described in the previous section, i.e. repeatability, disagreement and discrimination. Brockhoff [14] presented a univariate assessor performance method taking into account the 4 different basic assessor differences presented in the pre-

[3] Performance measures based upon ANOVA methods have also been applied to hedonic or affective data in the field of audio by Gabriellson [25] and Bech [3, 4]. A good introduction to the application of ANOVA to sensory data can also be found from Lea et al. [43]

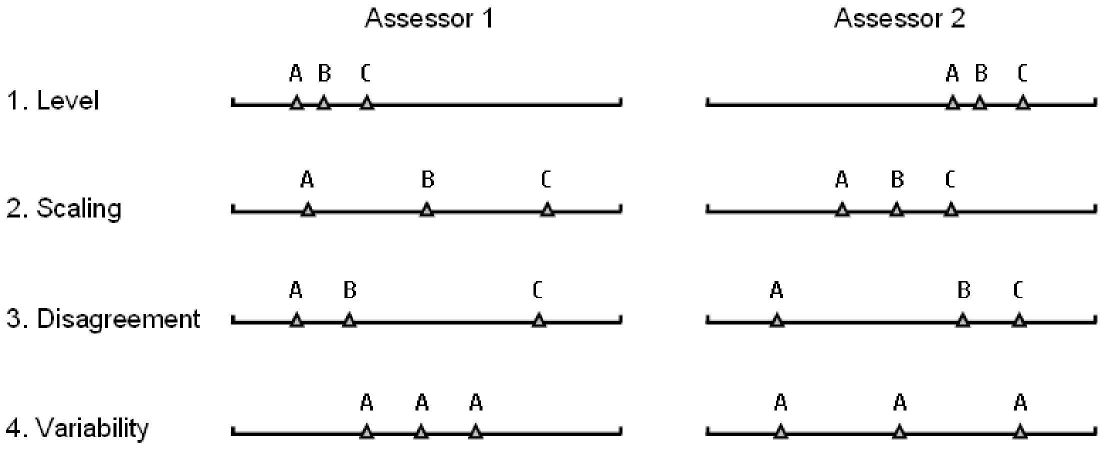


Fig. 2: The four basic assessor differences for a single sensory attribute (from Brockhoff [14]). Letters A, B and C represent the grading of three different products.

vious section. A set of formal statistical models based on ANOVA was employed and significance tests were proposed for the following assessor effects: differences in variability (see 4 in Figure 2), presence of disagreement (see 3 in Figure 2), differences in scaling (see 2 in Figure 2) and differences in sensitivity (defined as a signal-to-noise ratio). Brockhoff's approach is one of the most thorough univariate tools for assessor performance monitoring and can be run with an available SAS macro called PANMODEL which is of interest to evaluate in future. Finally, clustering methods or factorial methods have also been used for panel performance assessment of each attribute separately, e.g. Dijksterhuis [21] or Couronne [19] presented PCA-based methods to assess disagreement between panellists.

The analysis of scores at a sensory profile level offers a different perspective on the assessor performance. When multiple attributes are considered at once, a matrix of samples by attributes has to be handled for each assessor. Ledauphin et al. [45] presented a simple approach to measure agreement among assessors, which includes the computation of a weighted average for the panel, an index of agreement for each assessor and a test of statistical significance for this index (an example of application of this method is presented in the next section). Multivariate data analysis techniques can be used to identify underlying differences between products and to interpret these differences in terms of attributes. The performance of the panel can also be assessed from this type analysis. Factorial methods are commonly used for this purpose, e.g. techniques based on principal component analysis (PCA) (Husson et al. [28]) or canonical variate analysis (CVA) [71], which can incorporate visual representations of individual assessor variability. The use of partial least squares regression models was also reported by Thybo and Martens [77] who measured signal to noise ratios (ratio of systematic between-object variation and residual noise) for attributes, assessors and products to highlight problems of assessor disagreement and attribute discrimination. Finally, another group of multivariate data analysis methods can be considered to test assessor agreement on the interpretation of attributes. Multivariate data analysis methods such as Generalized Procrustes Analsyis (GPA, Gower [26]) or STATIS (Schlich [70]) take this assumption into account by allowing for rotations between individual assessor configurations.The principle and application of GPA is described with an example in the next section.

4.4. Examples of panel performance assessment methods

In this section, a practical example of sensory profile is considered to illustrate different approaches to panel performance assessment. The sensory dataset analysed in this paper originates from a study by Folkenberg et al. [24] in which cocoa-milk products were evaluated in three replicates by a panel of assessors using a set of consensus sensory attributes. The products were evaluated on a 15 cm unstructured scale with two anchor points placed 1 cm from each scale end marked as low and high. The original dataset consisted of 7 assessors, 14 products, 15 attributes and 3 replicates. However, one extra assessor was added to this dataset to illustrate the effect of noisy data on the panel performance criteria. The scores of this assessor comprised of random numbers selected on a magnitude scale corresponding to the original data scale. The sensory dataset including replicates was analysed with the univariate analysis method proposed by Schlich [69]. Then, the data was averaged over replicates to form a 3-way dataset (8 assessors x 14 products x 15 attributes)[4] and two multivariate methods focusing on a different aspect of the sensory dataset were applied. The technique of Ledauphin et al. [45] was considered to measure agreement among assessors, and the GPA method [26] was applied to assess consensus in attribute usage. All computations for this study were performed with the Matlab software.

4.4.1. Graphical Representation of Assessor Performances

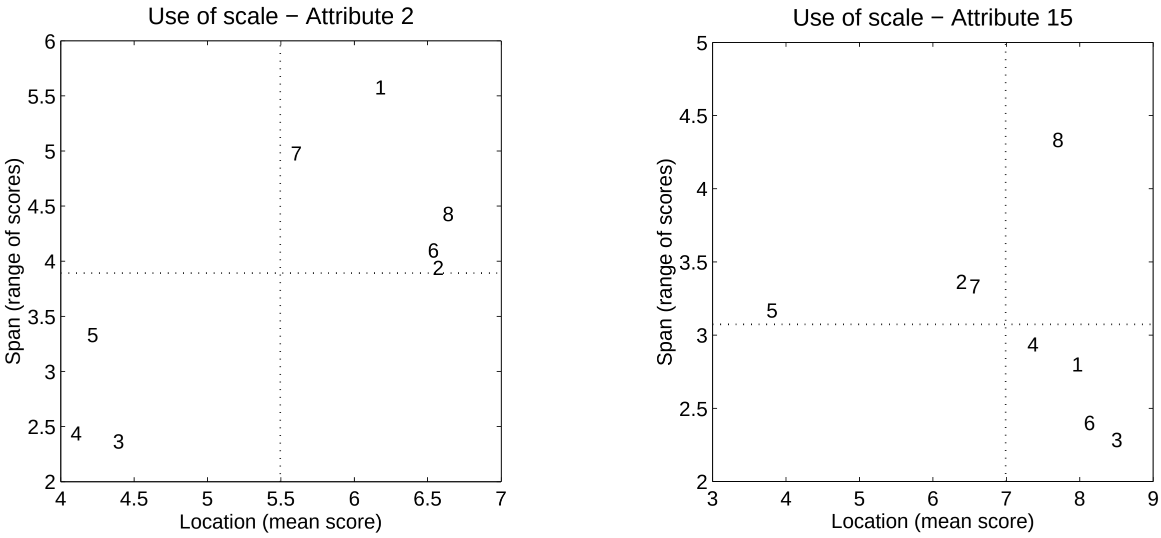
The Graphical Representation of Assessor Performances (GRAPES) [69] method proposed by Schlich was developed to address the issue of scale usage, repeatability over sessions, discrimination and disagreement for each attribute separately. The six statistics proposed in this method are derived from two ANOVA models for a given attribute. The first one (model 1) is applied on the repeated scores of one assessor and the second one (model 2) is applied on the repeated scores of all assessors. Span (range of scores) and location (mean score) reflect the use of the scale by the different assessors. The unreliability statistic relates to the product-by-session interaction of model 1 and the drift-mood statistic is the average variation among sessions. These two measures address the issue of assessor precision in scoring. Finally, Discrimination (F-test of product difference in model 1) and disagreement (contribution of an assessor to the product-by-assessor interaction in model 2) measure the extent to which each assessor discriminates the products and agrees with the rest of the panel.

A subset of the GRAPES algorithm output is presented below with a slightly different graphical representation from the original paper. Figure 3(a) and 3(b) illustrate scatter plots of scale usage for two attributes. The distribution of mean scores in Figure 3(a)highlights a level difference between two groups of assessors. The lower location and span of assessors #3, #4 and #5 indicate that they employed a lower part of the scale in comparison to the other assessors. It appears from Figure 3(b) that assessors #5 and #8 use the scale differently from the other assessors.

In Figure 4 bar graphs of the three key performance criteria of GRAPES are presented for four different attributes. The unreliability measure is presented on the left side of the graphs with a panel mean (horizontal dashed line). Assessor discrimination measures are shown in the centre plot with a panel mean (horizontal dashed line) and a 5% significant level for this F-ratio statistic (horizontal dotted line). The right side of the graphs shows the disagreement measure with a panel mean corresponding to the total product-by-assessor interaction in model 2 and a 5% significant level for this F-ratio statistic.

A clear pattern is visible for the assessor #8 in this series of graphs. The high unreliability, low discrimination and high disagreement observed in all graphs for this assessor show that the method detected the random nature of this assessor data. An overall difference between average discrimination of the different attributes can also be noted. Attribute #2 shows the largest discrimination in these four graphs. It is interesting to note that the unreliability and disagreement measures are respectively the lowest and the highest for this subset of attributes. The attribute #4 in Figure 4(b) shows the lowest mean discrimination with a non-significant discrimination for two assessors (#1 and #4). This attribute is also the only one with a non-significant panel disagreement. On an individual level now, Figure 4(a) illustrates that the assessors #1 and #7 are more discriminative than the assessor #4, but the right plot also

[4] A new set of random number was generated for the assessor #8 for this averaged dataset.



(a) Attribute 2

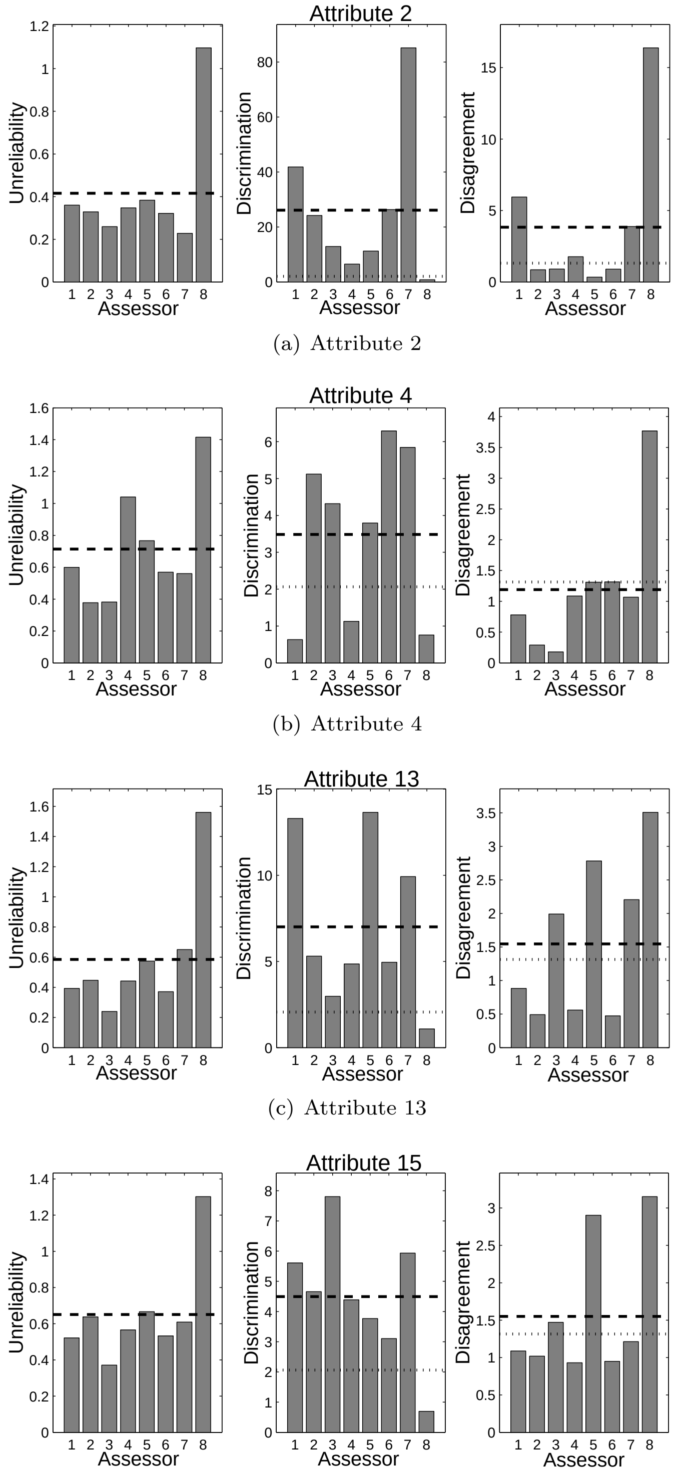
(b) Attribute 15

Fig. 3: Scatter plots of differences in scale usage for a given attribute, as found in GRAPES [69].

shows in this case a significant panel disagreement, to which these two assessors contribute largely. A different pattern can be seen in Figure 4(c) with two different types of assessor behaviour. Assessors #1, #5 and #7 all show a high discrimination, but only assessors #5 and #7 contribute to the panel disagreement.

4.4.2. A multivariate measure of assessor agreement [45]

The method proposed by Ledauphin et al. [45] to assess the agreement among assessors in a consensus vocabulary profile consists of four steps. The 3-way dataset (assessors x products x attributes) is pre-processed first by centering and scaling. A similarity matrix between the different assessors is calculated next, from which a weighting coefficient is derived for each assessor. A weighted average configuration can then be obtained and finally an index of agreement between each individual configuration and this weighted average configuration is computed. An additional step includes the testing of significance for these performance indices based on permutation tests. This method was applied to the sensory profile of each session separately and to the profile averaged over the three sessions. The result of this analysis is presented in Table 1 for separate sessions and in Table 2 for the average profile. Weighting factors used for the calculation of the average configuration and an individual index of agreement are shown for each assessor. A global index corresponding to an average over assessors is also included. Assessor #8 appears clearly as an outlier in this analysis with a performance index close to zero in all sensory profiles. It can also be noted than the weight applied to this assessor for the weighted average configuration is around 30% lower than the weight of the other assessors. A progression in the average performance index is visible over the three sessions (Table 3), which indicates that an improvement in agreement was achieved during the sensory profiling experiment. It is however interesting to note that all assessors do not behave similarly. Assessor #2 shows the best performance index in all cases but did not improve during the experiment. On another hand, the assessor #3 shows the best progression over sessions (Table 3) but still shows the lower performance index in the average sensory



(a) Attribute 2

(b) Attribute 4

(c) Attribute 13

(d) Attribute 15

Fig. 4: Bar graphs of individual panel performance for unreliability, discrimination and disagreement measures, based on GRAPES [69].

	Session 1		*Session 2*		*Session 3*	
Assessor	**Weight**	**Performance Index**	**Weight**	**Performance Index**	**Weight**	**Performance Index**
1	0.1283	0.742	0.1275	0.740	0.1286	0.790
2	0.1356	0.865	0.1342	0.852	0.1317	0.841
3	0.1234	0.661	0.1294	0.772	0.1296	0.806
4	0.1272	0.724	0.1282	0.753	0.1284	0.787
5	0.1337	0.833	0.1268	0.729	0.1301	0.815
6	0.1314	0.795	0.1328	0.829	0.1323	0.851
7	0.1311	0.790	0.1334	0.839	0.1309	0.829
8	0.0894	0.093	0.0877	0.081	0.0884	0.130
mean		0.773		0.788		0.817

Table 3: Performance index of assessor agreement based on Ledauphin et al. [45]. Sensory profile of three different sessions.

Assessor	**Weight**	**Performance Index**
1	0.1293	0.844
2	0.1344	**0.926**
3	0.1289	0.838
4	0.1307	0.866
5	0.1319	0.886
6	0.1328	0.900
7	0.1325	0.896
8	0.0796	**0.047**
mean		0.879

Table 4: Performance index of assessor agreement based on Ledauphin et al. [45]. Sensory profile averaged over sessions.

profile (Table 4).

4.4.3. Generalized Procrustes Analysis

Procrustes Analysis originated as a method for matching two configurations (e.g. a matrix of n samples by m attributes for two different assessors). The idea was generalised to a set of configurations with possibly different numbers of columns (e.g. attributes) in the 1970s (see Gower [26]). In the context of sensory profiling data, the method can be described as an iterative transformation of individual configurations by translation, rotation, reflections and isotropic scaling, with the aim of minimising the Procrustes distance between each configuration and a common configuration, the later being the mean of the transformed configurations. More details about the mathematics associated with Procrustes analysis can be found in [27] and a good review of the technique and its application to sensory profiling can be found in [1] or [22]. The transformations performed in GPA are aimed at correcting a number of assessor effects. The translation removes differences in level (effect 1 in Figure 2 for each assessor of each assessor and the isotropic scaling relates to the second assessor difference in Figure 2, but is applied to all the attributes of an assessor simultaneously. The rotation/reflection step is more specific to attribute usage and aims at aligning individual attribute descriptions. The main output of the GPA algorithm is a group average configuration, i.e. a matrix of average scores on a common set of “underlying” attributes. Information about samples, attributes and panellists is also a useful outcome of the GPA procedure. This method is commonly used in the field of sensory science for the analysis of individual vocabulary profiling data (e.g. Free Choice Profiling [79], Repertory Grid [76] and Flash profile [20]) but it has also been applied to consensus vocabulary profiles in e.g. [22] and [16]. Each assessor develops his or her own set of attributes in these methods, which removes the need for construct alignment between the assessors.

In the current study, GPA was applied to the average sensory profile to assess the consensus in attribute interpretation between the assessors. Other aspects relating to scaling factors or measures of noise in individual configurations e.g. GPA residual variance, can also be addressed with GPA, but this was not considered here. In the following step of the data analysis, a PCA was performed on the resulting group average data. Only the two first principal component were considered in this study, which explained 85.4% of the variation in the data. The correlation loading plot approach [51] is employed here to illustrate relationships between the original individual attributes and the principal components. Variables with large correlation loadings in this type of plots are considered more important to explain the differences observed between the systems. In the present case, a separate correlation loading plot is presented for each attribute to evaluate differences between assessors, though the underlying principal components are in fact identical. Circles indicating respectively 50% and 100% explained variance are included to help visual interpretation and correlation loading vectors of the average configuration are also added to illustrate the consensus direction.

Figure 5 illustrates the result of this analysis for the same subset of attributes considered with the GRAPES method. The assessor #8 appears as an outlier here again, as can be seen from the correlation loading vectors consistently in disagreement with the rest of the panel. Attribute #2 shows the largest average correlation loading vector. This attribute is therefore the most important to explain the differences observed between the products along these two dimensions or, in other words, the panel is very discriminative for this attribute. A large agreement between assessors is also visible for this attribute except in the case of assessor #4. Attribute 15 shows the lowest average correlation loading vector and the largest disagreement too. The very low correlation loading vector of the assessors #1 and #5 for the attribute #4 also indicates very poor product discrimination. Considering all the correlation circles of attributes, it appears that assessor #2 is in good agreement with the GPA average configuration, except for two attributes not presented here.

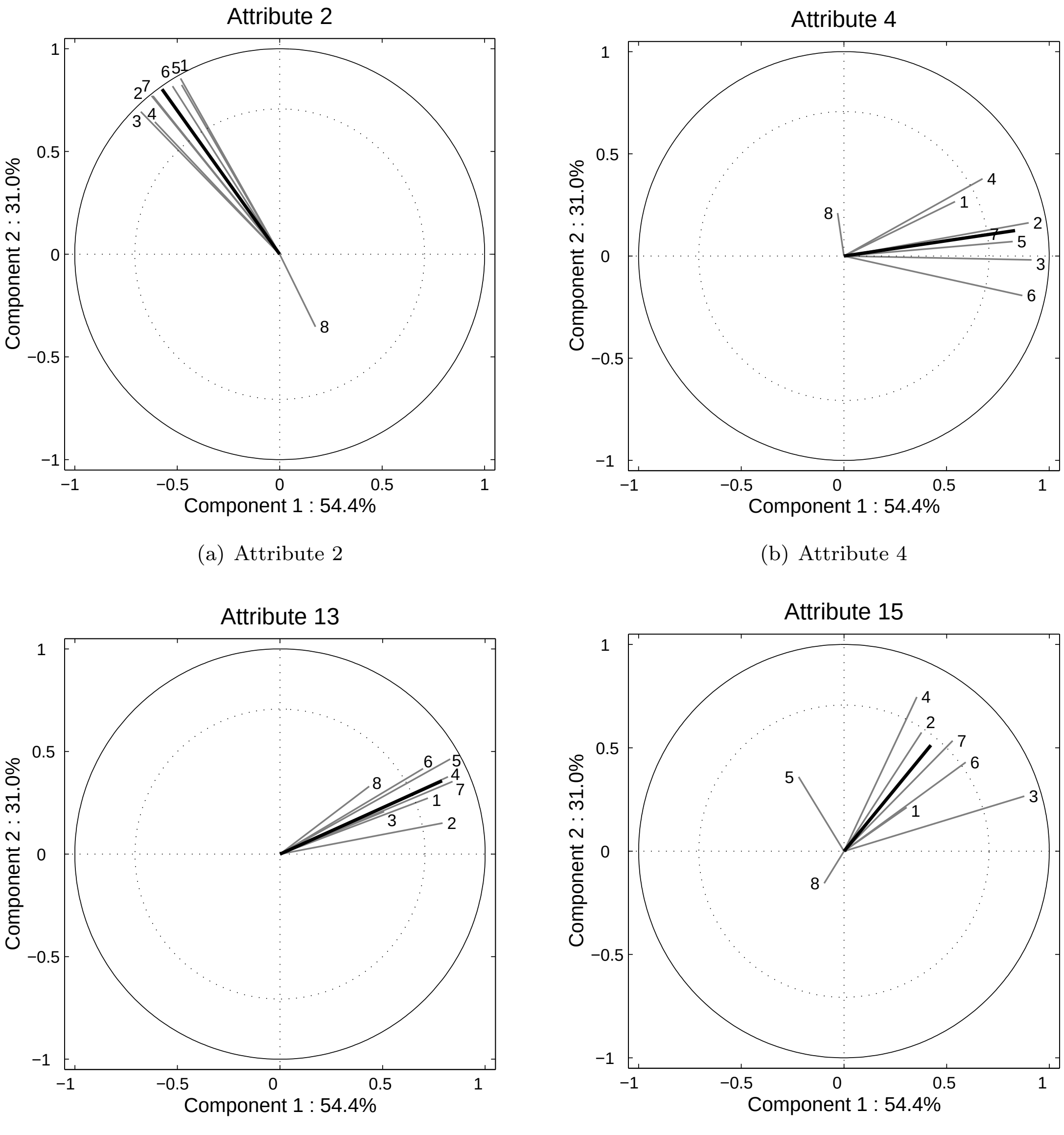
4.5. Brief comparison of the three methods

The three analysis techniques presented in this section give an idea of the different aspects of panel performance that can be addressed. The univariate method developed by Schlich offers a very detailed view of the performance of each assessor on an attribute basis taking into account replicate information, whereas the multivariate method proposed by Ledauphin et al. gives more compact representation of panel performance. The GPA method appears as a good compromise offering very detailed attribute-based information as illustrated in this study, but also more global performance criteria for each assessor (not covered in this paper). Comparing the results of the three methods now, it appears that all methods highlighted the poor performance of the artificial assessor #8. Assessor #2 was identified as the best assessor in terms of agreement performance index of Ledauphin et al., and this result was confirmed in qualitative terms by the GPA technique. The discrimination measure also compared well between the GRAPES and GPA methods at a panel level. However, some discrepancies were found between these two approaches at a assessor level, which might be explained by the very different statistical models underlying the performance criteria definitions of these two techniques.

4.6. Application to spatial sound perception

The above methods for evaluating the performance of assessors were targeted towards direct attribute rating methods using quantitative line scale. The methods are not developed for specific applications but may be applied to any domain of sensory evaluation. As a result their direct application to spatial sound can occur assuming that an existing and suitable vocabulary exists that can be taught to the subjects.

Several different studies have focused upon the development of vocabularies for the evaluation of spatial sound characteristics. For example Zacharov and



(a) Attribute 2

(b) Attribute 4

(c) Attribute 3

(d) Attribute 15

Fig. 5: PCA analysis of the GPA average configuration using the correlation loading plot method to illustrate relationships between the attributes and the principal components for each assessor. A separate plot is shown for four different attributes.

Koivuniemi [82, 40, 81] applied consensus language methods to develop attributes associated with loudspeaker based spatial sound reproduction systems. Berg and Rumsey [7, 8, 10, 9, 11, 6] performed similar studies based on individual elicitation using the repertory grid method (RGT). Lorho [47, 48] has applied both individual and consensus language approaches to the development of attributes for the evaluation of headphone spatial sound characteristics. These studies maybe used as a basis for training and subsequent evaluation of assessor performance.

5. CONCLUSION

Based upon this study of listening panel development, training, calibration and assessment methods some guidance can be provided regarding how to select panels for the evaluation of spatial sound quality by sensory evaluation methods.

It is clear from the literature that in order to obtain objective data from sensory evaluations panel that selected and trained subjects are essential. To select such a panel, pre-screening of the subjects is found to be a good practice in order to ensure you are sampling the correct population and that the subjects are potentially suitable (physiologically and demographically). Post-selection methods are then essential to assess the performance of the subjects. Three methods have been discussed in this paper all of which provide good means of evaluating the detail performance of subjects and panels. These methods are typically applicable to any field using consensus sensory vocabularies. The three method provide a slightly different perspective on each assessors performance, but generally provide a very similar picture. The performance index provides a single figure performance for subjects across all attributes. Whilst this is convenient and simple it does not provide the level of detailed analysis provided by the other two methods. GRAPES provides a very detailed analysis of the subjects performance for each attribute and also proposed acceptance criteria. In this manner subject performance or acceptance criteria can be defined and monitored in detail. The GPA based methods provide a more graphical interpretation of subject performance. This method allows easy identification of inter-subject agreement for each attribute and can be used to monitor subject performance by also attribute usage.

Whilst all the discussed method are found to be very suitable for the purpose of subject performance evaluation, the direct application to spatial sound, with associated acceptance criteria is still to be defined.

Lastly, the performance evaluation methods proposed by Brockhoff [14] are considered to be of great interest and should also be evaluated in comparison to the discussed methods.

6. ACKNOWLEDGEMENTS

Part of the work reported in this paper was carried out during a visit of the second author to the Department of Dairy and Food Sciences, The Royal Veterinary and Agricultural University (KVL) in 2003/2004, which was sponsored by a scholarship from NorFA and from the LMC/MRI program (KVL). The authors would like to thank Ditte Folkenberg, Wender Bredie and Magni Martens for allowing the use of their cocoa sensory profiling data [24] for this study.

7. REFERENCES

[1] Arnold, G. M., and Williams, A. A. *The use of Generalised Procrustes Analysis techniques in sensory analysis.* Elsevier Applied Science, 1986, pp. 233–253.

[2] Audio Engineering Society. Perceptual audio coders: What to listen for. Compact Disc, 2001.

[3] Bech, S. Selection and training of subjects for listening tests on sound-reproducing equipment. *Journal of the Audio Engineering Society 40*, 7/8 (1992), 590–610.

[4] Bech, S. Training of subjects for auditory experiments. *Acta Acustica 1* (1993), 89–99.

[5] Bech, S., and Zacharov, N. *Perceptual Audio Evaluation - Theory, Method and Application.* John Wiley & Sons, Chichester, England, 2006.

[6] Berg, J. *Systematic Evaluation of Perceived Spatial Quality in Surround Sound Systems.* PhD thesis, School of Music, Luleå University of Technology, May 2002.

[7] Berg, J., and Rumsey, F. Identification of perceived spatial attributes of recordings by repertory grid technique and other methods. In *Proceedings of the Audio Engineering Society 106th International Convention* (1999), Audio Engineering Society.

[8] Berg, J., and Rumsey, F. Spatial attribute identification and scaling by repertory grid technique and other methods. In *Proceeding of the Audio Engineering Society 16*$^{\text{th}}$ *International Convention* (1999), Audio Engineering Society.

[9] Berg, J., and Rumsey, F. Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction. In *Proceedings of the Audio Engineering Society 109*$^{\text{th}}$ *International Convention* (2000), Audio Engineering Society.

[10] Berg, J., and Rumsey, F. In search of the spatial dimensions of reproduced sound: Verbal protocol analysis and cluster analysis of scaled verbal descriptors. In *Proceedings of the Audio Engineering Society 108*$^{\text{th}}$ *International Convention* (2000), Audio Engineering Society.

[11] Berg, J., and Rumsey, F. Verification and correlation of attributes used for describing the spatial quality of reproduced sound. In *Proceeding of the Audio Engineering Society 19*$^{\text{th}}$ *International Conference* (2001), Audio Engineering Society.

[12] Bi, J. Agreement and reliability assessments for performance of sensory descriptive panel. *Journal of Sensory Studies 18* (1998), 61–76.

[13] Blauert, J., and Jekosch, U. Concepts behind sound quality: Some basic considerations. In *Proceedings of the the 32*nd *International Congress and Exposition of Noise Control Engineering* (Seogwipo, Korea, 2003), no. N466, Inter-Noise.

[14] Brockhoff, P. M. Statistical testing of individual differences in sensory profiling. *Food Quality and Preference 14*, 5–6 (2003), 425–434.

[15] Brockhoff, P. M., and Skovgaard, I. M. Modelling individual differences between assessors in sensory evaluations. *Food Quality and Preference 5* (1994), 215–224.

[16] Byrne, D. V., O'Sullivan, M. G., Dijksterhuis, D. B., Bredie, W. L. P., and Martens, M. Sensory panel consistency during development of a vocabulary for warmed-over flavor. *Food Quality and Preference 12* (2001), 171–187.

[17] CCITT. *Handbook of telephonometry.* International Telecommunications Union, 1992.

[18] Corey, J. An ear training system for identifying parameters of artificial reverberation in multichannel audio. In *Proceedings of the Audio Engineering Society 101*$^{\text{st}}$ *International Convention* (2004), Audio Engineering Society.

[19] Couronne, T. A study of assessors' performance using graphical methods. *Food Quality and Preference 8* (1997), 359–365.

[20] Delarue, J., and Sieffermann, J. M. Sensory mapping using flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products. *Food Quality and Preference 15*, 4 (June 2004), 383–392.

[21] Dijksterhuis, G. Assessing panel consonance. *Food Quality and Preference 6* (1995), 7–14.

[22] Dijksterhuis, G. *Multivariate analysis of data in sensory science.* Vol. 16 of *Data handling in science and technology* [58], 1996, ch. 7: Procrustes analysis in sensory research, pp. 185–217.

[23] Findlay, C. J., Castura, J. C., Schlich, P., and Lesschaeve, I. Use of feedback calibration to reduce the training time for wine panels. *Food Quality and Preference 17* (2006), 266–276.

[24] Folkenberg, D., Bredie, W., and Martens, M. What is mouthfeel? sensory-rheological relationships in instant hot cocoa drinks. *Journal of Sensory Studies 14* (1999), 181–195.

[25] Gabrielsson, A. Statistical treatment of data for listening tests on sound reproduction systems. Tech. Rep. TA 92, Department of Technical Audiology, Karolinska Inst., Sweden, 1979.

[26] Gower, J. C. Generalized procrustes analysis. *Psychometrika 40* (1975), 33–50.

[27] Gower, J. C., and Dijksterhuis, G. B. *Procrustes Problems.* Oxford University Press, 2004.

[28] Husson, F., Lê, S., and Pagès, J. Confidence ellipse for the sensory profiles obtained by principal component analysis. *Food Quality and Preference 16* (2005), 245–250.

[29] Isherwood, D., Lorho, G., Mattila, V.-V., and Zacharov, N. Augmentation, application and verification of the generalized listener selection procedure. In *Proceedings of the 115th Convention of the Audio Engineering Society* (New York, USA, 2003).

[30] ISO. *5497. Sensory analysis - Methodology - Guidelines for the preparation of samples for which direct sensory analysis is not feasible.* International Organization for Standards, 1982.

[31] ISO. *8586-1. Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 1: Selected assessors.* International Organization for Standards, 1993.

[32] ISO. *11035. Sensory analysis – Identification and selection of descriptors for establishing a sensory profile by a multidimensional approach.* International Organization for Standards, 1994.

[33] ISO. *5725-1. Accuracy (trueness and precision) of measurement methods and results – Part 1: General principles and definitions.* International Organization for Standards, 1994.

[34] ISO. *8586-2. Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 2: Experts.* International Organization for Standards, 1994.

[35] ITU-R. *Recommendation BS.1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems.* International Telecommunications Union Radiocommunication Assembly, 1997.

[36] ITU-T. *Recommendation P.800, Methods for subjective determination of transmission quality.* International Telecommunications Union, Telecommunications Standardization Sector, 1996.

[37] ITU-T. *Recommendation P.831, Subjective performance evaluation of network echo cancellers.* International Telecommunications Union, Telecommunications Standardization Sector, 1998.

[38] ITU-T. *Recommendation P.832, Subjective performance evaluation of handsfree terminals.* International Telecommunications Union, Telecommunications Standardization Sector, 2000.

[39] Kelly, G. *The psychology of personal constructs.* Norton, New York, 1955.

[40] Koivuniemi, K., and Zacharov, N. Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training. In *Proceedings of the Audio Engineering Society 111th International Convention* (2001), Audio Engineering Society.

[41] Labbe, D., Rytz, A., and Hugi, A. Training is a critical step to obtain reliable product profiles in a real food industry context. *Food Quality and Preference 15* (2004), 341–348.

[42] Langron, S. P. *Sensory Quality in Foods and Beverages: Definition, Measurement and Control.* Horwood, Chichester, UK, 1983, ch. The application of Procrustes statistics to sensory profiling., pp. 89–95.

[43] Lea, P., Næs, T., and Rødbotten, M. *Analysis of variance for sensory data.* Johm Wiley and Sons, Chichester, UK, 1997.

[44] Lea, P., Rødbotten, M., and Næs, T. Measuring validity in sensory analysis. *Food Quality and Preference 6* (1995), 321–326.

[45] Ledauphin, S., Hanafi, M., and Qannari, E. M. Assessment of the agreement among the subjects in fixed vocabulary profiling. *Food Quality and Preference 17* (2006), 277–280.

[46] Letowski, T. Development of technical skills: Timbre solfeggio. *Journal of the Audio Engineering Society 33* (1985), 240–243.

[47] Lorho, G. Evaluation of spatial enhancement systems for stereo headphone reproduction by

preference and attribute rating. In *Proceedings of the 118th Convention of the Audio Engineering Society* (Barcelona, Spain, 2005).

[48] Lorho, G. Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction. In *Proceedings of the 119th Convention of the Audio Engineering Society* (New York, USA, October 2005).

[49] Mandel, J. The validation of measurement through inter-laboratory studies. *Chemometrics and Intelligent Laboratory Systems 11* (1991), 109–119.

[50] Marchall, R. J., and Kirby, S. P. J. Sensory measurement of food texture by free-choice profiling. *Journal of Sensory Studies 3* (1988), 63–80.

[51] Martens, H., and Martens, M. *Multivariate analysis of quality - An introduction.* John Wiley, 2001.

[52] Mattila, V.-V., and Zacharov, N. Generalized listener selection (GLS) procedure. In *Proceedings of the 110th Convention of the Audio Engineering Society* (Amsterdam, Holland, 2001).

[53] McEwan, J. A., Hunter, E. A., van Gemert, L. J., and Lea, P. Proficiency testing for sensory profile panels: measuring panel performance. *Food Quality and Preference 13* (2002), 181–190.

[54] Meilgaard, M. C., Civille, G. V., and Carr, B. T. *Sensory Evaluation Techniques*, 3^{rd} ed. CRC, 1999.

[55] Merimaa, J., and Hess, W. Training of listeners for evaluation of spatial attributes of sound. In *Proceedings of the Audio Engineering Society 117th International Convention* (2004), Audio Engineering Society.

[56] Miskiewicz, A. Timbre solfege: A course in technical listening for sound engineers. *Journal of the Audio Engineering Society 40* (1992), 621–625.

[57] Moulton, D. Golden ears. `http://www.moultonlabs.com/gold.htm`, 2002.

[58] Næs, T., and Risvik, E. *Multivariate analysis of data in sensory science*, vol. 16 of *Data handling in science and technology.* Elsevier, Amsterdam, The Netherlands, 1996.

[59] Næs, T., and Solheim, E. Detection and interpretation of variation within and between assessors in sensory profiling. *Journal of Sensory Studies 6* (1991), 159–177.

[60] Neher, T. *Towards a Spatial Ear Trainer.* PhD thesis, Institute of Sound Recording, University of Surrey, GB, 2004.

[61] Neher, T., Rumsey, F. J., and Brookes, T. Training of listeners for the evaluation of spatial sound reproduction. In *Proceedings of the Audio Engineering Society 112th International Convention* (2002), Audio Engineering Society.

[62] Olive, S. E. A method of training of listeners and selecting program material for listening tests. In *Proceedings of the Audio Engineering Society 97th International Convention* (1994), Audio Engineering Society.

[63] Olive, S. E. A new listener training software application. In *Proceedings of the Audio Engineering Society 110th International Convention* (2001), Audio Engineering Society.

[64] Pedersen, T. H., and Fog, C. L. Optimisation of perceived product quality. In *Euronoise 98* (1998), vol. II, Euronoise, pp. 633–638.

[65] Quesnel, R. Timbral ear trainer: Adaptive, interactive training of listening skills for evaluation of timbre difference. In *Proceedings of the Audio Engineering Society 98th International Convention* (1996), Audio Engineering Society.

[66] Quesnel, R. *A computer-assisted method for training and researching timbre memory and evaluation skills.* PhD thesis, McGill University, Montreal, Canada, 2002.

[67] Quesnel, R., and Woszczyk, W. R. A computer-aided system for timbral ear training. In *Proceedings of the Audio Engineering Society 96th International Convention* (1994), Audio Engineering Society.

[68] Rossi, F. Assessing sensory panelist performance using repeatability and reproductibility measures. *Food Quality and Preference 12* (2001), 467–479.

[69] Schlich, P. Grapes: A method and SAS program for graphical representation of assessor performance. *Journal of Sensory Science 9* (1994), 157–169.

[70] Schlich, P. *Defining and validating assessor compromises about product distances and attribute correlations.* Vol. 16 of *Data handling in science and technology* [58], 1996, pp. 185–219.

[71] Schlich, P., Pineau, N., Brajon, D., and Quannari, E. M. Multivariate control of panel performances. In *Presented at the 7th Sensometrics Meeting, Davis, CA, USA* (2004).

[72] Scriven, F. Two type of sensory panel or are there more? *Journal of Sensory Studies 20* (2005), 526–538.

[73] Shively, R. E., and House, W. N. Listener training and repeatability for automobiles. In *Proceedings of the Audio Engineering Society 104*th *International Convention* (1998), Audio Engineering Society.

[74] Sieffermann, J. M. Flash profiling. a new method of sensory descriptive analysis. In *AIFST proceedings, 35th Convention* (Sidney, Australia, July 21–24 2002).

[75] Stone, H., and Sidel, J. L. *Sensory evaluation practices*, 2nd ed. Academic Press, 1993.

[76] Thomson, D. M. H., and McEwan, J. A. An application of the repertory grid method to investigate consumer perceptions of foods. *Appetite 10* (1988), 181–193.

[77] Thybo, A. K., and Martens, M. Analysis od sensory assessors in texture profiling of potatoes by multivariate modelling. *Food Quality and Preference 11* (2000), 283–288.

[78] Williams, A. A., and Arnold, G. M. A comparison of the aromas of six coffees characterised by conventional profiling, free-choice profiling and similarity scaling methods. *J. Sci. Food Agric. 36* (1985), 204–214.

[79] Williams, A. A., and Langron, S. P. The use of free-choice profiling for the evaluation of commercial ports. *Journal of food an agriculture 35* (1984), 558–568.

[80] Wolters, C. J., and Allchurch, E. M. Effect of training procedure on the performance of descriptive panels. *Food Quality and Preference*, 5 (1994), 203–214.

[81] Zacharov, N., and Koivuniemi, K. Unravelling the perception of spatial sound reproduction: Analysis & external preference mapping. In *Proceedings of the Audio Engineering Society 111*th *International Convention* (2001), Audio Engineering Society.

[82] Zacharov, N., and Koivuniemi, K. Unravelling the perception of spatial sound reproduction: Techniques and experimental design. In *Proceedings of the Audio Engineering Society 19*th *International Conference on Surround Sound* (2001), Audio Engineering Society.