How do we determine the attribute scales and questions that we should ask of subjects when evaluating spatial audio quality?

Jan Berg

School of Music, Luleå University of Technology, P O Box 744, SE-94128 Piteå, Sweden

ABSTRACT

As a result of the research on spatial audio quality, different ways of generating evaluation scales have been utilised. The experience gained shows that attribute scales have successfully been derived through different methods. Comparisons of the results show that similar attributes were found across experiments. From this, the implication is that the choice of language development method is less critical for the resulting attributes. For future tests it is suggested that strategies for how to deal with the complexity of the auditory scene and how to circumvent construct masking should be developed and that the questions to the subjects should reflect these considerations.

1. INTRODUCTION

A variety of audio systems that rely on their capability to render different forms of spatial information exist, e.g. cinema sound, home cinema systems, automotive systems, and handheld terminals for enhanced spatial effects. The spatial quality of such and similar systems has been evaluated in several experiments, e.g. [1][2][3][4][5][6]. An important part of quality assessment is the development of evaluation scales, or the quantification of the qualities of spatial audio.

Spatial quality can be evaluated by either considering it as a holistic concept, or as an entity that is decomposable into its constituent dimensions. Both of these aspects can be combined. The dimensions in the quoted work are captured by utilisation of the spatial attributes of the sounds under test. The spatial attributes are expressed through verbal descriptors or more generic, a language that can be used in the evaluation scales. Consequently, the language development process, or the scale generation, is important, as the resulting scales should, as closely as possible, reflect the audible sensations perceived by listeners.

The most obvious problem then becomes to select which attribute scales are relevant for a specific test. Some fundamental properties of such scales are that they should capture the perceptual characteristics of the stimuli; the scales should also be clear and unambiguous to allow for a common understanding across subjects and the scales should differentiate between stimuli. In addition, if no overlap of the scales is desired, the scales should also be orthogonal.

As a result of the research on spatial audio, different ways of generating the scales have been studied. Therefore, experience of scale development has now been gained.

In this paper, some of the factors influencing the decision on which scale generation approach to be utilised are discussed. The objective is to present some suggestions for points to be observed in future experiments relating to attribute scaling of spatial audio quality.

2. ATTRIBUTE GENERATION

There are basically two ways to generate attribute scales; (1) scales originating from the experimenter's definitions, provided constructs, and (2) scales originating from subjects' constructions, elicited constructs. These constructs may be expressed explicitly by words, phrases or drawings accounting

for the perceived sensations, or implicitly by numbers indicating inter-stimulus similarities or differences [1]. A mix between provided and elicited constructs occurs in some experiments.

2.1. Scales originating from provided constructs

Attribute scales based on constructs that are provided and defined by the experimenter for the test at hand can be used when the attributes have shown to be functional within the given context in previous tests. Such scales can also be applied when there is reason to believe that they will be applicable in the current context, e.g. attributes from concert hall acoustics used for evaluation of reproduced sound. Attribute scales regarded as stable and reliable today may once have been elicited from subjects. One example is some of the attributes used by Toole in 1985 [2] that were defined by Gabrielsson in 1979 [3], who in his turn used partly terms from Kötter's work in 1968 [7] and partly subjects' free descriptions. By means of factor analysis, Gabrielsson reduced a large list of words into a relative few attribute scales.

The risk of using experimenter-defined attribute scales that have not been validated by others is that subjects may not be able to relate to these. The advantage is that a well-defined scale will enable the subjects to focus on the sensations that the experimenter would like to evaluate.

2.2. Elicited constructs

A number of methods for construct elicitation and attribute scale generation are available. In recent years, audio researchers have utilised procedures not previously occurring in the context of spatial audio.

Generally, the methods can be divided into three groups: (a) those that through consensus aim to arrive at a common set of attributes for grading by all panel members, (b) those that are based on free categorisation or individualised scales, and (c) those which use some form of multidimensional analysis based on non-semantic similarity/difference relationships between stimuli. When individual language methods were employed, different approaches to find attributes that are common for a majority (ideally all) of the subjects were utilised in some experiments.

Examples of audio evaluation versions of methods in the aforementioned groups are: (a) Descriptive Analysis [4][6][8]; (b) Repertory Grid Technique [1] and Individual Vocabulary Profiling [9] (inspired by Free-Choice Profiling and Flash Profile techniques); Perceptual Structure Analysis [10] (based on Knowledge Space Theory and Formal Concept Analysis); The graphical assessment language GAL [11]; (c) Multidimensional Scaling (MDS) [12], Individual Differences Scaling (INDSCAL) [13].

In addition to these, there are elicitation methods not yet tested in audio evaluation, like Hierarchical Dichotomisation and Free Elicitation. These methods were compared to Repertory Grid Technique by Steenkamp & van Trijp [14] in a market evaluation context.

3. INFLUENCING FACTORS

When deciding on which approach to be followed, the experimental context comprises a number of factors having a possible influence on which scales should be used. In this section some of these factors are discussed.

As in all research, the most important question to ask oneself before pursuing an experiment is: Which question or problem should the experiment provide data for solving? What is the objective of the test?

A selection of possible objectives of an experiment might be to find: the perceptual structure of the stimulus set expressed as its attributes; one or a few attributes accounting for a desired feature; the stimulus that evokes the greatest preference; how a particular attribute is affected by the stimuli; the differences between groups of listeners; a linear combination of attributes contributing to preference; etc.

When the aim and the objective are established, these will govern other factors like the following:

What types of listeners will be participating in the test? Will they be naïve, experienced or trained? A naïve listener may require another test procedure compared to an experienced. A listener categorised as experienced may possess some degree of uncontrolled experience compared to a listener having received a known training.

The stimuli may range from being perceptually simple to complex, which will affect the subjects' prospect to focus on the desired characteristics of the auditory scene.

The number of stimuli will affect the size of the evaluation experiment. This depends on whether all

combinations of every stimulus are to be evaluated by each subject.

In different phases of product development, a fast evaluation may be required in order for development to proceed. This may limit the time available for evaluation.

Other factors are bias issues and generalisability of the results.

All these factors will have an influence on which scales should be utilised in the evaluation and how the scales should be generated.

4. OBSERVATIONS AND SUGGESTIONS

The work within the field has contributed to more understanding of how scales can be generated and interpreted. In this chapter, observations on the work are reported and some suggestions for future tests are given.

4.1. Stimulus complexity

The selection of stimuli is governed by a number of factors, e.g. the objective of the test. If a home cinema system is to be evaluated, stimuli that are normally occurring in such a system can be expected in the test. In that case, the stimuli will no doubt be complex to their nature, e.g. film sound tracks with dialogue, music and effects. The human ability to discriminate between different events in an auditory scene, as described by Bregman [15], also opens the possibility for subjects to focus on different parts, or streams of the scene. In an evaluation situation, one subject's attention may be directed towards one subset of the auditory scene, whereas another subject may focus on another subset, although the stimuli are identical. This implies that the judgements of those two subjects cannot automatically be compared as they might refer to different percepts.

In the case of multiple streams, a suggested approach for acquiring the relevant constructs is to clarify and/or define the scene's auditory streams, e.g. single musical instruments or groups of instruments as employed in [16] and described in [17]. An alternative solution is to involve the subject in the scene definition process. Hence, the elicited constructs can be linked to the appropriate auditory stream. Elicitation questions that support this approach may facilitate a more effective use of the resulting attributes. This also helps to formulate more precise questions to be asked to the listeners.

4.2. Listener experience

An experienced listener is likely to have reflected more on what he/she perceives and is thereby assumed to have a greater ability to both detect and express audible differences of the stimuli. For naïve listeners, an elicitation method that does not assume that they directly can name the attributes of the stimuli can be more appropriate. In descriptive analysis, the normal procedure is to have a group of experienced listeners. A claimed feature of the Repertory Grid Technique is the ability to reveal knowledge that the elicitee (listener) possesses, but is unable to express directly. This indicates that less experienced listeners may benefit from this method. An elicitation relying on a more indirect method can be one measure to compensate for a listener's lack of experience.

An example of varying verbal skills was observed in [18], when some subjects during an elicitation experiment used gestures instead of words. The ability to express what one perceives can be considered as sort of listener expertise as well.

Rumsey *et al* [19] investigated the difference between naïve and experienced listeners who assessed band-limiting and down-mixing of multichannel audio. They observed that experienced listeners considered the frontal spatial fidelity as important, whereas the naïve listeners did not. In addition, for the naïve listeners, the surround spatial fidelity was a more important factor than for the experienced listeners. An obvious observation here is that these two subject groups put different weights on the scales used. Hence, a thorough scale generation achieved through elicitation from one group of listeners might pose a problem if those scales were applied to a group with another level of experience.

4.3. Construct masking

Another possible problem is that a stimulus can have multiple attributes. If one attribute is perceptually dominant over another, the less dominant attribute risks being masked and therefore undetected by the subjects.

One measure to counteract this is to exhaust all perceived attributes of the stimulus, or combination of stimuli, under consideration by the subject, before the stimuli are changed. A way to elicit as many attributes as possible is to perform a number of repetitions and/or refinements of the elicitation question. The assumption here is that the most perceivable characteristics of the stimulus will be

revealed first, and subsequently the less dominant in an onion-peeling fashion. Not until all constructs seem to be elicited, a new stimulus/stimuli is presented to the subject.

4.4. Relative or absolute elicitation?

An elicitation process can employ an absolute approach, where one stimulus at a time is presented to the subject, or a relative, where two or more stimuli are available for comparison.

According to Nunnally and Bernstein [19], human judgements of a stimulus show better consistency when they are compared with other stimuli, rather than with a fixed 'internal' scale. This implies that relative judgements are easier performed than absolute. This was also noted by Mason *et al* [21], who used both an absolute and a relative technique during elicitation; the absolute elicitation technique was less sensitive and consistent than the relative was.

These observations seem to support the use of comparisons in the elicitation process. However, there may be cases where an attribute is present in two stimuli at a perceptually equal magnitude. Hence, the inter-stimulus difference on that attribute is undetectable. This situation will prevent the attribute from being elicited through comparison, although the attribute may be an important characteristic for the stimuli.

If the objective of the evaluation is to find differences between known stimuli, the relative approach seems more convenient. The absolute approach, however, gives information about the subject's internal reference.

4.5. Duration of language development sessions

How much time can be allotted for the development of a language that is functional for attribute scales? It seems reasonable to assume that the more thorough the process is, the longer the duration, How does the experimenter know when the language has the desired precision? Are there certain methods that are faster than others? These questions have not systematically been studied yet.

Lorho [9] notes that an individual language development process is less time-consuming compared to three other experiments using the consensus language approach, in his case 3 hours compared to 20, 30 and 60 hours respectively. In an experiment based on the Repertory Grid Technique with triadic elicitation, the process inclusive grading took 55 minutes in average across subjects [22]. These results are not directly comparable due to different experiment conditions, but raise the question whether there is time to be saved with a 'correct' choice of language development method.

4.6. Level of abstraction

In a non-audio context, Steenkamp & van Trijp [14] observed that Free Elicitation yielded a higher number of constructs and that these were more abstract than those resulting from RGT and Hierarchical Dichotomisation. If this is valid for audio as well has to the knowledge of this author not been established.

4.7. Comparison of resulting scales

Experiments utilising different methods show similarities on many of the resulting spatial attributes, e.g. the work by Toole [2], Koivuniemi & Zacharov [4], Berg & Rumsey [1] and for some attributes also by Ford [11] and Lorho [6]. Attributes relating to width/broadness, locatedness/sense of direction, image focus, and source distance were occurring in most of the cited experiments.

This similarity across experiments suggests that the different language elicitation methods all have the potential of revealing the relevant attributes.

5. SUMMARY

The experience gained through spatial audio quality evaluation shows that attribute scales have successfully been derived through different methods. Comparisons show that similar attributes were found across experiments. For future experiments the following is suggested:

- Strategies for how to deal with the complexity of the auditory scene should be developed, i.e. the questions to the listener should reflect the scene.
- Ways of circumvent construct masking should be tried, by e.g. repetition and/or refinement of the questions.
- The choice of the elicitation method should be based on the expected experience of the listeners.
- The duration of language development process should be investigated. Researchers are encouraged to record and publish the time spent on different parts of the process.

It is also noted that the language development method seems less critical for the resulting attributes, as similar attributes result from different methods.

6. REFERENCES

[1] Berg, J. & Rumsey, F. (2006): Identification of quality attributes of spatial audio by repertory grid technique. *J. Audio Eng. Soc.* Accepted for publication.

[2] Toole, F. (1985): Subjective measurements of loudspeaker sound quality and listener performance. *J. Audio Engineering Society.* 33, pp 2-32.

[3] Gabrielsson, A. (1979): Dimension analyses of perceived sound quality of sound-reproducing systems. *Scandinavian Journal of Psychology* **20**, pp 159-169.

[4] Koivuniemi, K. & Zacharov, N. (2001): Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training. Presented at *AES 111th Convention, New York*. Preprint 5424. Audio Engineering Society.

[5] Bech, S., Ellermeier, W., Ghani, J., Gulbol, M.-A., and Martin, G. (2005): A Listening Test System for Automotive Audio - Part 2: Initial Verification. *AES 118th Convention, Barcelona*. Preprint 6359. Audio Engineering Society.

[6] Martin, G. & Bech, S. (2005): Attribute Identification and Quantification in Automotive Audio - Part 1: Introduction to the Descriptive Analysis Technique. Presented at *AES 118th Convention, Barcelona*. Preprint 6360. Audio Engineering Society.

[7] Kötter, E. (1968): *Der Influss übertragungstechnischer Faktoren auf das Musikhören*. Arno Folk Verlag, Köln.

[8] Lorho, G. (2005): Evaluation of Spatial Enhancement Systems for Stereo Headphone Reproduction by Preference and Attribute Rating. Presented at *AES 118th Convention, Barcelona*. Preprint 6514. Audio Engineering Society.

[9] Lorho, G. (2005): Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction. Presented at *AES 118th Convention, Barcelona*. Preprint 6629. Audio Engineering Society.

[10] Choisel, S. & Wickelmaier, F. (2005): Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound. Presented at *AES 118th Convention, Barcelona*. Preprint 6369. Audio Engineering Society.

[11] Ford, N., Nind, T. and Rumsey F. (2002): Evaluating Influences of a Central Automotive Loudspeaker on Perceived Spatial Attributes using a Graphical Assessment Language. Presented *at AES 113th Convention, Los Angeles*. Preprint 5707. Audio Engineering Society.

[12] Mattila, V.-V. (2002): Ideal Point Modelling of Speech Quality in Mobile Communications Based on Multidimensional Scaling (MDS). Presented at *AES 112th Convention, Munich*. Preprint 5546. Audio Engineering Society.

[13] Martens, W. & Zacharov, N. (2000): Multi-dimensional Perceptual Unfolding of Spatially Processed Speech I: Deriving Stimulus Space Using INDSCAL Presented at *AES 109th Convention, Los Angeles*. Preprint 5224. Audio Engineering Society.

[14] Steenkamp, J.-B. E. M. & van Trijp, H. C. M. (1997): Attribute Elicitation in Marketing Research: A Comparison of Three Procedures. *Marketing Letters* 8:2, pp 153–165, Kluwer Academic Publishers, Netherlands.

[15] Bregman, A. S. (1990): *Auditory scene analysis*. MIT Press, Cambridge, Mass.

[16] Berg, J. & Rumsey, F. (2002): Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques. Presented at *AES 112th Convention,Munich*. Preprint 5593.

[17] Rumsey, F. (2002): Spatial quality evaluation for reproduced sound: terminology, meaning and a scene-based paradigm. *J. Audio Eng. Soc.* **50** pp. 651-666.

[18] Berg, J. & Rumsey, F. (2003): Systematic evaluation of perceived spatial quality. In proceedings of *AES 24th International Conference on Multichannel audio*. Audio Engineering Society.

[19] Rumsey, F., Zieliński, S., Kassier R., and Bech, S. (2005): Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences. *J. Acoust. Soc. Amer.*, **117**, 6, pp. 3832-3840.

[20] Nunnally, J. C. & Bernstein, I. H. (1994): *Psychometric theory*. McGraw-Hill.

[21] Mason, R., Ford, N., Rumsey, F. and de Bruyn, B. (2000): Verbal and non-verbal elicitation techniques in the subjective assessment of spatial sound reproduction. Presented at *AES 109th Convention, Los Angeles*. Preprint 5225. Audio Engineering Society.

[22] Berg, J. (2005): OPAQUE – A Tool for the Elicitation and Grading of Audio Quality Attributes. Presented at *AES 118th Convention, Barcelona*. Preprint 6480. Audio Engineering Society.