Evaluation of Perceived Spatial Audio Quality

Jan BERG[1]

School of Music, Luleå University of Technology
Piteå, Sweden

ABSTRACT

The increased use of audio applications capable of conveying enhanced spatial quality puts focus on how such a quality should be evaluated. Different approaches to evaluation of perceived quality are briefly discussed and a new technique is introduced. In a series of experiment, attributes were elicited from subjects, tested and subsequently used for derivation of evaluation scales that were feasible for subjective evaluation of the spatial quality of certain multichannel stimuli. The findings of these experiments led to the development of a novel method for evaluation of spatial audio in surround sound systems. Parts of the method were subsequently implemented in the OPAQUE software prototype designed to facilitate the elicitation process. The prototype was successfully tested in a pilot experiment. The experiments show that attribute scales derived from subjects' personal constructs are functional for evaluation of perceived spatial audio quality. Finally, conclusions on the importance of spatial quality evaluation of new applications are made.

Keywords: Spatial, Audio, Quality, Attribute, Evaluation, Software.

1. INTRODUCTION

Many of the available audio platforms now enable the distribution of multichannel (comprising more than two channels) sound. The increased use of such sound systems has created a vast number of possibilities for producers, engineers, editors and consumers to create and/or alter the sound image finally reproduced at the consumer's end of the chain. This sound image is able to give the listener an improved feeling of presence and more directional cues. One of the predominant features of a multi-channel sound system is the spatial impression created by the system, i.e. how the system deals with the three-dimensional character of the sound sources and their environment. This is referred to as the spatial quality of the system.

The spatial quality may be influenced by the different processes present in the signal path from recording via post-production and distribution to reproduction. In order to evaluate how these processes affect the spatial quality, methods for capturing the different aspects of the quality have to be developed.

This paper begins with by a short review on different approaches to evaluation of perceived spatial quality. The main part reports on the results of a series of experiments, where attributes of spatial quality were elicited, tested and subsequently used for derivation of evaluation scales. A software prototype, designed to facilitate the elicitation process, was developed and tested in a pilot experiment. Finally, conclusions on the importance of spatial quality evaluation of new audio applications are made.

2. APPROACHES TO EVALUATION

Various approaches have been utilised in order to assess different aspects of a sound system's performance. These could roughly be divided into two categories of approaches: 'objective' and 'subjective'. The objective approach refers to the use of parameters measurable by means of some (electrical) instrument, whereas the subjective approach encompasses methods where human subjects are used for detecting and quantifying some characteristics of interest. The subjective approach is essential as it seeks to capture how humans perceive the sound. This is also significant when an objective instrument for measuring spatial quality is to be constructed, as the instrument has to be correlated to human perception to ensure the instrument's validity.

In order to assess the perceived spatial quality of a sound system, it is important to know the components of this conception. The problem can be formulated as a need to find the perceived components of spatial sound and to scale them. Since human perception is the scope of different sciences, e.g. psychophysics, research methods from these must be considered. It is well known from psychology that certain perceptual variables or components cannot be observed directly [1], which has resulted in techniques for extracting underlying components or latent variables. One way of finding these components is to observe a number of perceptual variables and test for similarities and differences between them. If a subset of the variables could be expressed as a common factor or an attribute, it would be an indication of some common dimensionality within the subset.

Some evaluation methods rely on the assessment of the quality on a holistic basis, without the intention of decomposing it into constituent dimensions. In these methods, the participating

[1] The author was previously affiliated also with Sonic Studio, Interactive Institute, Sweden.

subjects judge the total audio quality by assigning grades on a pre-defined scale. Two methods in common use for quality test of codecs used for bit-rate reduction, e.g. MP3, are the ITU-R BS. 1116-1 [2] and BS. 1534-1 [3]. In these cases, the subjects have to weigh all aspects of the different qualities, e.g. spatial and spectral, into a single number. A disadvantage of such a method is that two sound stimuli may sound very different, but be given similar grades, due to the internal weighing by the subjects.

The approach utilising decomposition of the total audio quality or subsets thereof into attributes scales was employed by other authors, e.g. Gabrielsson [4], Toole [5], Rumsey [6], Bech [7] and Koivuniemi & Zacharov [8]. These and others are reviewed by the author in [9].

A problem common to all methods making use of attribute scales is the selection and definition of the attributes to be included. To serve their purpose, the attributes should be unambiguous, i.e. perceived similarly across the subject group. If verbal descriptors are used to define the attributes, these should have the same meaning to all the subjects. If this is not the case, the grades given on a certain scale does not refer to the same sensation, which subsequently causes problems when analysing the results. One implication of a heterogeneous interpretation of the scales may be a high noise level in the data. Therefore, measures have to be taken to ensure the relevancy of the attribute set.

The attribute set can be generated mainly in two ways, defined by the researcher, provided attributes, or obtained from subjects, elicited attributes. Using provided attributes, the risk is that the researcher's preconceptions influence the selection and definition of the attributes in a way that certain characteristics of the stimuli remains undetectable.

3. ATTRIBUTE DEFINITION

In an attempt to overcome some of the problems with attribute definitions, an approach involving the elicitation of verbal descriptors of spatial audio quality was developed and tested. The elicitation method employed was the Repertory Grid Technique (RGT), originally resulting from work by Kelly [10], and later adapted by Berg and Rumsey [11] to fit audio quality evaluation purposes.

RGT relies on each subject to use his/her own language and personal constructs. An experiment [11] was conducted, where 18 subjects compared reproductions of various audio processes (5-channel recordings and downmixes thereof) containing speech, music and environmental sound through a surround sound system according to ITU 775 [12]. In the experiment, designed to generate personal constructs that describes spatial audio quality, subjects were asked to verbally indicate similarities and differences between the stimuli during individual interviews. The output from each comparison of stimuli was a pair of opposite terms or phrases, referred to as a bipolar construct. The bipolar constructs were written down.

In order to describe the relation between audio stimuli and the elicited bipolar constructs, each subject graded the stimuli on his/her own constructs. An example is shown in Fig 1. After completion of the grading, the data was entered into one matrix, or as it is referred to in RGT, a grid, per subject. Fig 2

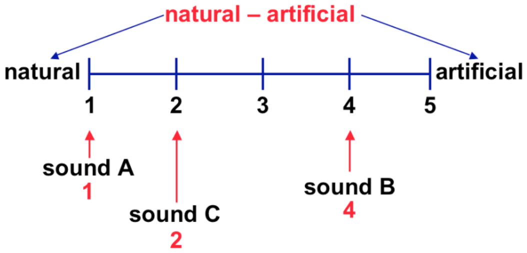
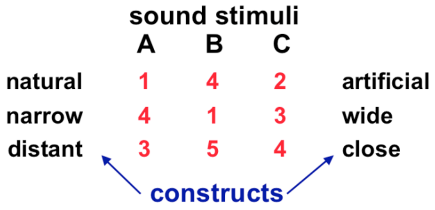


Fig. 1: Three sound stimuli graded on a bipolar construct.



	sound stimuli A	 B	 C	
natural	1	4	2	artificial
narrow	4	1	3	wide
distant	3	5	4	close

constructs

Fig. 2: A grid containing three bipolar constructs and the grades for three stimuli.

The data in the grid was subjected to multivariate analysis by means of principal component analysis and cluster analysis. The purpose was to reduce the data in the grid down to a smaller number of components that account for a considerable part of the variance in the data. Consequently, these components constitute the greater part of the perceived differences of the stimulus set and therefore they represent the attributes of the stimuli. By examining the verbal content of the grid and the interrelation of the constructs, conclusions were drawn upon which attributes were present in the data. The process was also performed across subjects to find inter-subjective similarities. The data reduction process was reported in detail in [13] and resulted in a number of attributes of spatial quality.

4. VERIFICATION OF ATTRIBUTES

In order to verify the validity of the attribute set derived from the elicitation experiment, a new experiment was carried out [14]. A new group of subjects were recruited (n=19). They were listening to the same surround sound system used in the previous experiment. The task was to use the previously generated attributes for scaling of a stimulus set that comprised differences in spatial quality created by downmix algorithms. The assumption was that if the attribute scales were sensitive to differences in the stimulus set, it would be possible to observe significant differences between the stimuli on the scales. However, if the scales were unable to detect differences, the data would mainly contain randomly distributed points, which preclude statistical significance.

The results showed that all attribute scales in the experiment produced significant differences between at least one stimulus and the remaining stimuli. From this it was concluded that attributes resulting from elicitation have a promising potential to be valid.

It was also observed that some attributes were less consistently graded across subject than others. This may serve as an indication of ambiguity within certain attributes. An example of such an attribute is “Naturalness”, from which a higher degree of different interpretations can be expected.

5. VALIDITY OF ATTRIBUTES IN EVALUATION OF 5-CHANNEL MICROPHONE TECHNIQUES

As the previously generated attribute set showed to produce significant results, the validity of the set was tested once more in an experiment reported in [15]. This time, the differences between the stimuli were reduced even further; all reproductions were made through the same type of audio system as in the two previous experiments, but this time no downmixing algorithms were employed. Instead, five different 5-channel microphone techniques were utilised for the recording of two music events in a hall, in total resulting in ten stimuli.

Before the main experiment, a limited elicitation procedure involving the current stimuli took place. The reason for this was to ensure that possible new attributes emerging from the new stimuli set would be captured. In addition, observations from the previous attribute verification experiment were taken into account when deciding which attributes should be included.

The attribute set of this experiment finally encompassed:

- Low frequency content
- Naturalness
- Preference
- Presence
- Ensemble width
- Localisation/locatedness
- Source envelopment
- Source width
- Source distance
- Room envelopment
- Room size
- Room level
- Room width

As can be noted, subjects were instructed to separately judge some characteristics of the source (the instrument playing) as well as of the acoustical environment (the hall).

The results showed statistical significance for all attributes tested. The room attributes were independent of the sound source in most cases. The attribute set seemed to be perceived mainly in three dimensions; width, distance to the source and a sensation of presence in the room/hall.

Examining the attributes of the acoustical environment only, subjects perceived these on two dimensions – one comprised judgements on the size of the room and the level of the reflected sound, whereas the other related to the impression of being present at the venue where the sound is created, i.e. a feeling of presence. These dimensions were also observed in the experiment in sect. 4. The room attributes in the ‘judgement’ dimension showed to still be detectable despite a downmix of the 5-channel recording to mono, which implies that certain properties of the room might be uncorrupted by similar processes.

It was also possible to detect different features of the microphone techniques used, e.g. the omni-directional techniques emphasised width, whereas the coincident technique gave a more focussed position of the sound source.

In summary, the experiment showed that the scales resulting from elicitation based on the stimuli under test produce significant results. Conclusions on the dimensionality of the attributes were also made. More results are available in [15].

6. ELICITATION SOFTWARE

The results from the experiments in sect. 3 through 5 shows a functional approach to the evaluation of spatial quality. The key feature of this approach is the elicitation of personal constructs, which was accomplished by interviewing subjects, one by one. However, this way of collecting the personal constructs was quite time-consuming. Therefore, a solution employing some means of automation was sought. As a response to this, the OPAQUE (Optimisation of Perceived Audio Quality Evaluation) project was initiated.

In its current form the OPAQUE is a prototype of a computer application for elicitation, grading and analysis of personal constructs. The system comprises three interface screens that facilitate:

- Elicitation of personal constructs by comparison of triads of stimuli selected from the stimulus set
- Grading of all stimuli on the elicited constructs
- Data analysis by reduction of the data set through grouping of the graded original constructs

The sequence toggles between the elicitation process and the grading process until the desired number of triads has been presented to the subject. Plots of screen examples are in Figures 3 to 5.

In brief, the different parts of the interface function as follows:

6.1. Elicitation

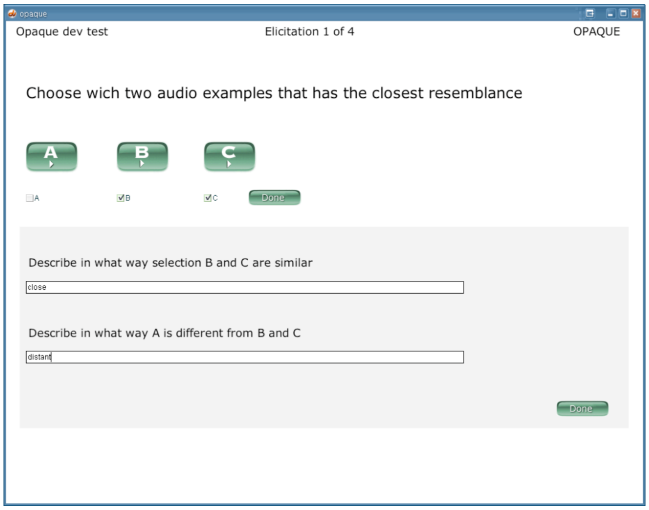
Three out of the total number of stimuli under test are randomly selected and assigned to the three playback buttons. The subject is instructed to indicate which two of them are more similar and thereby different from the third. When the indication is done by the subject, two text input fields are displayed, where the subject enters phrases describing the perceived similarity and difference respectively. When the subject has completed these operations, the grading process commences. For every new elicitation, a stimuli triad that has not previously occurred is presented to the subject.

6.2. Grading

The grading screen comprises a scale at which endpoints the text from the two text fields is displayed. Thus, a bipolar scale is formed. The sound stimuli are represented on the screen by an icon each. Each icon works as a playback button for the associated stimulus. The icons are movable along the bipolar scale and can thus be placed at a scale position that corresponds to the subject’s judgement of the stimuli on that construct.

6.3. Data analysis

The data reduction in the current software version is performed by means of cluster analysis. The resulting cluster is represented by a dendrogram that enables the experimenter to get a visual representation of the data structure, i.e. how the original constructs are related. Constructs that are similarly perceived are then grouped together.



Opaque dev test Elicitation 1 of 4 OPAQUE

Choose wich two audio examples that has the closest resemblance

□A ☑B ☑C

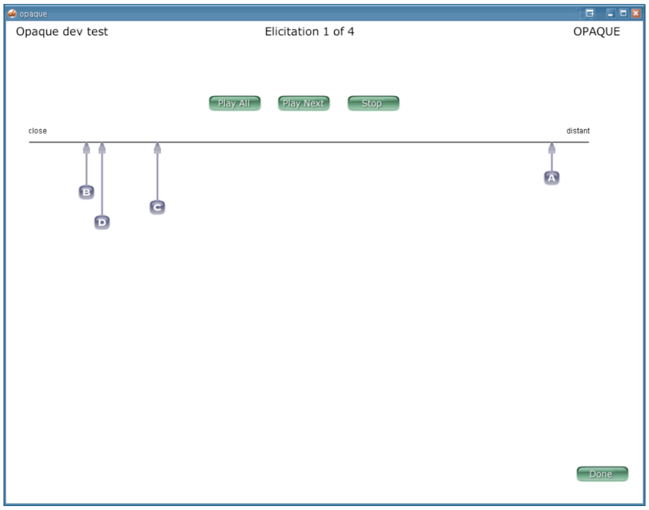
Describe in what way selection B and C are similar

close

Describe in what way A is different from B and C

distant

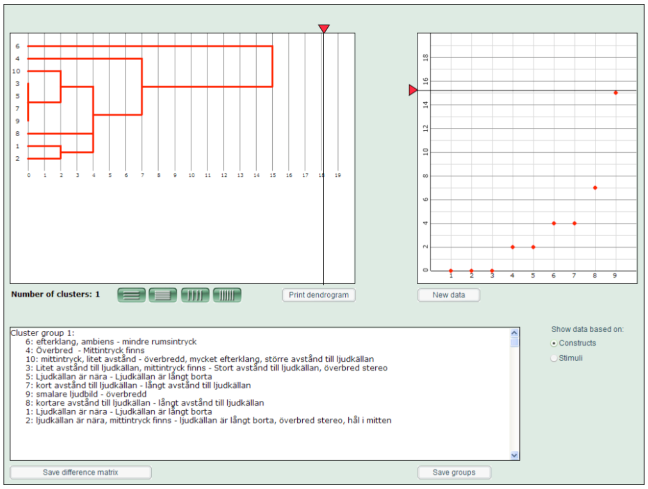
Fig 3: Elicitation screen



Opaque dev test Elicitation 1 of 4 OPAQUE

Fig. 4: Grading screen

The decision on how many groups that contain similar variables is supported by the agglomeration distance plot (Fig. 5, upper right part of the screen). By inspecting the plot, the number of groups of variables can be determined by identification of the point where the slope of the plot makes a clear change. The interface allows for a dynamic change of the number of groups, in order to find the optimal data structure.



Cluster group 1:
6: efterklang, ambiens - mindre rumsintryck
4: Överbred - Mittintryck finns
10: mittintryck, litet avstånd - överbredd, mycket efterklang, större avstånd till ljudkällan
3: Litet avstånd till ljudkällan, mittintryck finns - Stort avstånd till ljudkällan, överbred stereo
5: Ljudkällan är nära - Ljudkällan är långt borta
7: kort avstånd till ljudkällan - långt avstånd till ljudkällan
9: smalare ljudbild - överbredd
8: kortare avstånd till ljudkällan - långt avstånd till ljudkällan
1: Ljudkällan är nära - Ljudkällan är långt borta
2: ljudkällan är nära, mittintryck finns - ljudkällan är långt borta, överbred stereo, hål i mitten

Fig. 5: Analysis screen

6.4. Pilot experiment

The software was tested in a pilot experiment [16], where four subjects evaluated stimuli processed through different reverberation algorithms. The experiment resulted in a number of attributes derived from the subjects' personal constructs. The attributes were:

- Room/hall size, reverberation
- Source distance
- Presence (ability to give the impression of being present in the room)
- Width
- Treble level

The experiment showed that the OPAQUE software could be used for finding attributes of a stimulus set. The attributes revealed in the pilot experiment mainly referred to spatial quality. Similar attributes were previously encountered in other studies, which to some extent reinforces the findings in this experiment.

From the subjects' point of view, the software seemed to be simple to use. This might enable inexperienced listeners to participate in listening tests in the future without intensive training and/or detailed instructions.

The conclusion of the pilot experiment was that the software is a functional tool for attribute generation.

7. DISCUSSION

In the series of experiments, attributes derived from subjects' personal constructs have shown to produce statistically significant results. A similar approach, but using another elicitation and attribute definition method, was tested by Koivuniemi & Zacharov [8]. They also came up with attributes that produced statistically significant results. In their work, attributes similar to the ones presented above were encountered.

The methodological approach outlined by the series of experiments has shown to be functional for the evaluation of spatial audio quality. A prototype of a computer implementation, OPAQUE, has been tested with promising results.

Current and coming applications including audio are likely to strive for an enhanced spatial audio quality. As the traditional means of spatial audio reproduction (e.g. stereophonic home systems, cinemas, home theatre systems) to a large extent depends on physically displaced transducers, the implementation in mobile devices will pose a challenge due to their relatively small physical dimensions. The spatial sensations have to be created using advanced signal processing, possibly generating new artefacts. Therefore, methods for evaluation of the spatial quality will be an even more important part of product development to ensure that the desired quality goals are reached.

8. ACKNOWLEDGEMENTS

The author wishes to thank the following: Dr Francis Rumsey for valuable support during the preceding research work; Oscar Lovnér and Andreas Renhorn for recording of speech and music used in the early experiments; Jonas Ekeroot for software programming in the early experiments; Johnny Isacson and Mats Liljedahl for OPAQUE software programming; Stefan Lindberg for preparation of the audio clips used as stimuli in the pilot experiment; Members of the EUREKA project 1653 for discussions on methodology. This project was carried through by Luleå University of Technology and Sonic Studio, Interactive Institute in cooperation.

9. REFERENCES

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

[2] ITU-R (1996): **Recommendation BS. 1116, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems**. International Telecommunication Union.

[3] ITU-R (2003): **Recommendation BS.1534-1. Method for the subjective assessment of intermediate quality levels of coding systems**. International Telecommunication Union.

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

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

[6] Rumsey, F (1998): Controlled subjective tests on 2-5 channel surround processing algorithms. Presented at **AES 104th Convention, Amsterdam**. Preprint 4654. Audio Engineering Society.

[7] Bech, S. (1998): The influence of stereophonic width on the perceived quality of an audiovisual presentation using a multichannel sound reproduction system. **J. Audio Eng. Soc**. **46**, pp 314-322.

[8] 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.

[9] Berg, J. (2002): **Systematic evaluation of perceived spatial quality in surround sound systems**. Doctoral thesis, 2002:17. Luleå University of Technology, Sweden.

[10] Kelly, G. (1955): **The psychology of personal constructs**. Norton. New York.

[11] Berg, J. and Rumsey, F. (2006): Identification of quality attributes of spatial audio by repertory grid technique and other methods**. J. Audio Eng. Soc**. **54**, 5 pp 365-379**.**

[12] ITU-R (1993): **Recommendation BS.-775, Multichannel stereophonic sound system with or without accompanying picture**. International Telecommunication Union.

[13] Berg, J. and Rumsey, F. (2000) In search of the spatial dimensions of reproduced sound: Verbal Protocol Analysis and Cluster Analysis of scaled verbal descriptors. Presented at **AES 108th Convention, 19-22 February, Paris**. Preprint 5139.

[14] Berg, J. and Rumsey, F. (2001): Verification and correlation of attributes used for describing the spatial quality of reproduced sound. In *Proceedings of the* **AES 19th International Conference on Surround Sound, 21-24 Jun**. pp 233-251. Audio Engineering Society.

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

[16] Berg, J. (2005): OPAQUE – a tool for the elicitation and grading of audio quality attributes. Presented at **AES 118th Convention, Barcelona, Spain**.