SYSTEMATIC EVALUATION OF PERCEIVED SPATIAL QUALITY

JAN BERG[1,2] and FRANCIS RUMSEY[1,3]

[1]*School of Music, Luleå University of Technology, Sweden*
[2] *Interactive Institute, Piteå, Sweden*
jan.berg@tii.se
[3]*Institute of Sound Recording, University of Surrey, Guildford, UK*
f.rumsey@surrey.ac.uk

The evaluation of perceived spatial quality calls for a method that is sensitive to changes in the constituent dimensions of that quality. In order to devise a method accounting for these changes, several processes have to be performed. This paper shows the development of scales by elicitation and structuring of verbal data, followed by validation of the resulting attribute scales.

INTRODUCTION

Evaluation of the audio quality a listener perceives is an important issue for all those involved in audio recording and reproduction. The variety of tools in audio work facilitates advanced processing of the audio signal. The aims of these processes differ, but they all in some way, sooner or later give rise to the question of in which way the audio quality is influenced by them.

The use of multichannel techniques for both recording and reproduction has given many listeners the opportunity to listen to surround sound. New systems open possibilities to new experiences for the members of the audio community, both technically and creatively. Surround sound systems are no exception. But what are these experiences? What can a listener perceive? How can the perceived sensation be described? How can it be 'measured'? The answer to these questions could be expressed as the audio quality of a surround sound system. In two-channel stereo, a listener can perceive locations and distances of sound sources. These are descriptions of where sources are positioned; perhaps even how wide they appear to be. Such properties of the perceived sound are referred to as spatial characteristics, or all together as spatial quality. How can we describe spatial quality? How can we evaluate it? What forms the spatial quality of a surround sound system?

These questions are addressed in this paper, where a novel method for evaluation of spatial audio in surround sound systems is presented, and also some results of using such a method.

1 BACKGROUND

This section will provide a summary of issues relating to audio quality evaluation and spatial quality. The aim is to put the research work into its context, starting with a general discussion on sound quality and the fundamentals of objective and subjective evaluation, followed by the definition of spatial quality.

1.1 Sound quality and the recording-reproduction chain

In general, all work concerning recording, post production, distribution and reproduction of sound sooner or later involves a quality evaluation process to answer questions like: "Which of these systems affects the audio quality in the most desired way? Are the audio quality goals of this production met? How can we evaluate differences between audio systems or between recording techniques?" Even if an intentional evaluation is not arranged, those dealing with sound production are likely to constantly reflect on their work. Most reproductions of sound are aimed to present some sort of artistic performance (e g music) to a listener, and the total appreciation/sensation of such a reproduction is obviously affected by the performance itself. If the artistic value in the form of the '(musical) performance quality' itself is not considered, the quality of the chain from recording via the intermediate steps to reproduction then remains to be evaluated. The word "quality" in this context does not preclude descriptive or attitudinal judgements, although it may implicitly have attitudinal connotations to some. This chain is henceforward referred to as the recording-reproduction chain, *Fig 1*.

The recording-reproduction chain of sound utilising reproduction by means of loudspeakers or headphones comprises a number of stages, all where the audio signal is subjected to some form of processing/alteration, either deliberately for achieving certain effects, or undeliberately as a result of the inherent properties of the different stages. Some inherent properties might also be deliberately used by audio engineers for the creation of certain effects, e g the so-called tape compression occurring when an analogue tape is partly saturated by the audio signal. Examples of the stages and the means of alteration are:

- Recording: the room/hall; microphone directional characteristics; microphone positioning; recording equipment in general.
- Post-production: levels of the microphones used; spectral equalisation, dynamic processing, reverberation.
- Distribution: coding for different media, e g Compact Disc (CD), Digital Audio Broadcast (DAB), the Internet, Digital Versatile Disc (DVD), etc, by use of different coding algorithms, e g ISO/MPEG 1 and 2, Dolby Digital, Digital Theatre Systems (DTS) etc.
- Reproduction: Different modes, such as headphones or loudspeakers, 2-channel stereo or 5.1 surround; types of transducers; the reproduction environment (the room); virtual surround; spatial enhancement processes.

Whether the sound source and/or the listener should be considered as parts of the recording-reproduction chain is a matter of definition. However, in this work the chain does not include these.

The digitalisation of audio has, compared to e g storage by analogue means, yielded improvements of electrically measurable quantities of the systems, such as their signal-to-noise ratio, frequency response and total harmonic distortion [1]. Regarding these quantities, Rumsey [2] notes: "Although improvements may still be made in these domains, the technical quality curve is becoming asymptotic to the ideal and product development is in a region of diminishing returns." Altogether, these observations suggest that other measures of quality than the purely technical/electrical ones can be considered.

1.2 Objective and subjective evaluation of audio quality

When evaluating audio quality, two main approaches are found. On one hand, evaluation methods employing measures of physical quantities of the sound, such as sound pressure level, total harmonic distortion, reverberation time, etc are used. These are sometimes referred to as 'objective' methods. On the other hand, when the task is to evaluate the perceived quality of sound, the question of what is perceivable by a listener is crucial. To address perceived audio quality in general, methods where human listeners (subjects) express their judgements on sound stimuli have been utilised in numerous experiments, e g [3, 4, 5, 6, 7, 8, 9]. These methods are often referred to as 'subjective', thus implying that human judgement of the sounds is involved.

Both subjective and objective methods are of interest in audio research, especially when they are used in parallel. If a relationship between a sensation perceived by listeners and a physical quantity could be established, manipulation of the latter's magnitude could be used for controlling the strength of the perceived sensation. An example of this is the variation of the output power of an amplifier for controlling the perceived loudness of a loudspeaker reproduction of sound.

1.3 Perceived total audio quality – MOS tests

The predominant way of assessing perceived (subjective) audio quality, especially for digital low bit-rate audio codecs (=coders and decoders), is utilisation of scales encompassing all aspects of audio quality in a single judgement. This way of addressing the total audio quality is often referred to as 'MOS tests', from the Mean Opinion Score scale used in codec tests, like the ITU-R BS. 1116-1 [10]. Such scales give information on the basic audio quality, but they do not indicate in any detail what features of the audio reproduction contribute to this. Therefore, these scales are likely to be less sensitive to certain aspects of the perceived audio quality, and thereby less useful for dedicated evaluation applications.

1.4 Subsets of perceived audio quality

Bech [11] showed that a total auditory impression perceived by a listener could be assumed to consist of a number of auditory attributes subsequently combined

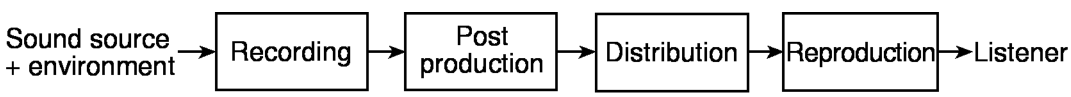

Fig 1: The recording-reproduction chain

together to form this impression. This implies a relationship between subsets of the auditory perception and the perceived total audio quality. The total conception of perceived audio quality was shown by Gabrielsson [12] to consist of different subsets, or perceptual dimensions as he referred to them, each relating to specific perceived properties of the sound. Letovski [13] suggested in his MURAL model that the auditory image is composed of timbre and spaciousness. Without claming completeness, a generic model for the components of perceived total audio quality may include:

- Timbral quality (Relating to the tone colour, or: "the sensation whereby a listener can judge that two sounds are dissimilar using other criteria other than pitch, loudness or duration", as defined by Pratt and Doak [14].)
- Spatial quality (Relating to the three-dimensional nature of the sound sources and their environments.)
- Technical quality (Relating to distortion, hiss, hum, etc.)
- Miscellaneous quality (Relating to the remaining properties.)

In general, these subsets might be evaluated either objectively or subjectively, if suitable methods for identifying the subsets are found. An example is timbre, which has been analysed by several authors [15, 16]. However, from now on this paper is focused on the perceived spatial quality.

1.5 Perceived spatial quality and attributes

In order to assess the subset "perceived spatial quality", this has to be distinct from other possibly existing subsets. The authors defined this as

> "the three- dimensional nature of sound sources and their environments" [17].

Mason and Rumsey [18] defined the "spatial impression" as

> "the auditory perception of the location, dimensions, and other physical parameters of a sound source and the acoustic environment in which the source is located".

Zacharov *et al* [19] performed an evaluation of audio systems, where the spatial sound quality was described as

> "… all aspects of spatial sound reproduction. This might include the locatedness and the localisation of the sound, how enveloping it is, its naturalness and depth".

These definitions imply that perceived spatial quality comprises those perceptual constructs that relate to the sensations of directionality, size, depth and width, of reproduced sources, groups of sources and acoustical environments.

The descriptors used for symbolising sensations or perceptual constructs are often referred to as attributes [20]. Nunnally and Bernstein [21] made the following distinction:

> "The term 'attribute' … indicates that measurement always concerns some *particular* feature of objects. One cannot measure objects – one measures their attributes." … "The distinction between an object and its attributes may sound like mere hair-splitting, but it is important. First, it demonstrates that measurement requires a process of abstraction."

Hence, perceived spatial quality of sound constitutes of a number of sensations described by, or embodied in, spatial attributes. Previous work on identification and employment of perceived spatial attributes are summarised in the next section. A schematic presentation of the relation between perceived total audio quality, subsets of this (e g spatial quality and timbral quality) and attributes of the subsets are shown in *Fig 2*.

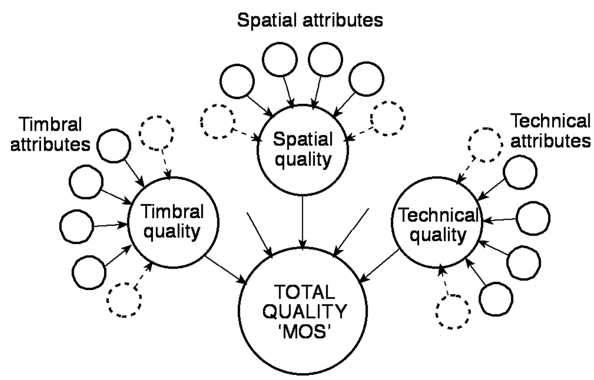

Fig 2: Relations between total audio quality, and its subsets and attributes.

2 SUMMARY OF PREVIOUS WORK

In [22], encompassing the publications [17, 23, 24, 25, 26], Berg reviewed previous work on evaluation of spatial sound quality. The review showed that reproduced audio as well as live sound is perceived multidimensionally and that sensations relating to the spatial features of the sound could be identified, that is spatial quality, as defined by the authors, exists in the context of reproduced audio and that it could be perceived. It was also showed that the perceived spatial quality is made up of different sensations describable by attributes. This is implied by the fact that, in certain contexts, listeners show consistent trends in their judgements of spatial audio reproductions. The different approaches used for assessing and analysing spatial quality all had weaknesses as well as strengths, and these will be discussed in this summary. The section is concluded by the implications of the observations on the previous work.

2.1 Evaluation methods

The most common approach was to use words or phrases to represent sensations. These verbal representations were used for construction of scales for the listeners to make judgements on. (In some other work, not relating to spatial features of sound, multidimensional methods not primarily relying on verbal approaches was used, e g Grey [15], who used Multidimensional Scaling [27] in scaling of musical timbres.) When verbal descriptors in spatial audio work seemed to work less efficiently, i e the results showed low consistency across listeners, it was mainly attributable to the definition of the terms used. The less precise the descriptor was, the more confusion it raised among the listeners. In some cases, the scales used did not enable the listeners to discriminate between the sounds they listened to. The explanation for this confusion could be found both in unclear definitions as well as in too wide descriptions that embraced many sensations, the latter indicating multidimensionality within a single attribute. An extreme example of multidimensionality was the 'MOS' tests, where all perceptual dimensions were included in the judgement. In addition, translations of verbal descriptors were addressed as a source of uncertainty.

The decision on which attributes were perceivable, if they were unidimensional and how these should be verbally represented was in several cases made by the experimenters themselves, directly or indirectly, assuming that these matters were known. Other approaches to generation of verbal descriptors or the search for attributes included participation of people, often experienced listeners, with or without pre-knowledge of the purpose of the experiment, from whom information was elicited. The different forms of descriptor generations were interviews, collection of word lists and panel discussions.

If elicitation of verbal descriptors was used, in the elicitation phase, not all of the studies used the same sound excerpts as those later intended for assessment.

Some studies successfully used standard multivariate techniques [28] (principal component analysis, factor analysis, product-moment correlation, cluster analysis) to check for redundant attributes, find relationships between attributes and to bring the size of the attribute set down to a practically applicable number of attributes for coming experiments.

Rating of preference also occurred in the work, both in reproduced audio and in concert hall acoustics. In some cases it was conducted together with rating on attribute scales with the intention to find predictors for preference in the attribute set.

Regarding the scales, verbal anchors in combination with numbers, integers or decimals, seem to be the most common type and were used by Gabrielsson *et al* [8], Zacharov [29] and Olive [30]. However, the anchors used in the different experiments differed from one another. To find the types of anchors and scales that possess the property of being superior to others is intricate, but there are some indications that intermediate verbal anchors could affect the scale linearity [20].

Alternative approaches, like graphical methods, were also found, e g work by Mason *et al* [31] and Ford *et al* [32].

2.2 Attributes

A number of attributes relating to the spatial quality of sound were found. Certain attributes seem to occur in almost every study. As indicated by Beranek [33] and Shaw and Gaines [34], terminology used by different individuals in different contexts may contain differences in interpretation between individuals (listeners). This should be observed when comparing verbal responses in general. Due to this, no extensive comparison of previous work's findings regarding the attributes themselves is intended at this point. However, some trends are observable.

The review of previous work showed that some of the spatial attributes encountered in concert halls also were perceivable in surround sound systems, whereas other spatial attributes only occurred when sound was reproduced through audio systems.

Predominant both for live sound and reproduced sound was the attribute "spaciousness". Other terms used for describing this are "envelopment", "spatial impression" and "ambience". In concert hall applications, this seems to be composed of "listener envelopment" and "apparent source width". In general, descriptions of different forms relating to the width of the source, the scene, the stage, etc, were often occurring. Theories of what physical measure best would predict different aspect of these were given, but since the relationship with physical parameters is not the scope of the research work reported in this paper, those are not considered here.

Some references on the auditory event's extension and/or its distance from the listener were also found.

In concert halls acoustics, there was no occurrence of descriptions of source location. This may have a simple explanation in that the visual cues from the stage totally dominate the perception of location, and thereby suppress the auditory importance of localisation – people can see the source's position. In audio systems, where no visual cues relating to the source are present and where the system's ability to render source locations accurately is limited, auditory localisation becomes more important. Consequently, attributes relating to localisation were found in some studies.

The work on audio systems contained descriptions of sensations not encountered in live sound. Examples of these were "disturbing sounds", "abnormal effects" and "penetration".

It was also noted that attributes relating to movement/positional changes were less frequent. An obvious explanation is that most of the sound sources encountered were stationary. What made this point interesting is that if one set of sounds is used to define the attributes, and these attributes are used for judgements on another sound set, some attributes may no longer be valid, or some perceived sensations may not be represented among the attributes.

What was not seen in the previous work was a division of the auditory scene into single components, or to use Bregman's [35] term, streams, in order to aid the listener to focus on singularities. In a scene with more than one perceived sound source, it was not possible to rate the different sources within the scene independently on any attribute, as no separate scales for each source were provided.

2.3 Implications for evaluation of spatial quality

In some work, it was shown that a group of listeners shared a common interpretation of the spatial quality of audio reproductions. Evaluation of a subset of audio quality was accomplished by instructing the subjects to focus on different parts of the auditory event. The instruction comprised verbal descriptors used as tools for directing the subject's attention towards the desired feature of the sound. Possible problems connected with the use of verbal methods were found to be either unclear or too general descriptions, misinterpretations by listeners, irrelevancy of the scales and translation of the descriptors.

The approaches utilised in previous work all have different features, but no one of them alone encompassed all of the following characteristics:

- Elicitation of perceived spatial sensations in the form of verbal responses from other (experienced) people than the experimenter.
- The use of multichannel reproduction/surround sound systems.
- When eliciting verbal responses, the use of stimuli later included in the evaluation.
- Systematic collection and structuring of verbal responses.
- Generation of attribute scales based on the elicited perceived spatial sensations.
- The scales employed explicitly referring to different parts of the reproduced sound, as independent sources or the environments of these.
- The attribute scales generated by these steps being used for rating stimuli.

As noted above, work where these points have been employed all together in assessment of spatial quality has not been encountered.

When taking this fact into account, the implications for evaluation of perceived audio quality in general, and of spatial quality in particular, are that a continued work within this field should consider the characteristics above in the design of evaluation methods. It is also reiterated that the instruction and the set of descriptions are crucial for the result of the evaluation, and that the process in which the descriptors are derived and defined must be given careful consideration.

This is the background and justification for the evaluation method described in this paper.

3 DEVELOPMENT OF THE METHOD

The theoretical and experimental work performed in order to investigate the applicability of a method was reported by Berg in [22]. The method described in the publications [17, 23, 24, 25, 26] can briefly be summarised in the following points:

- Verbal data relating to a relevant stimulus set was elicited from subjects.
- The data was structured by means of different techniques to find principal structures in the data set.
- These structures were analysed for their meaning and they were subsequently used for deriving attributes.
- Attribute scales were defined.
- Evaluation was performed by experienced listeners auditioning a number of sound stimuli which were rated on the attribute scales.
- The ratings were analysed for statistically significant differences between the stimuli and to determine the validity of the derived attribute scales.
- The evaluation was repeated on a new set of stimuli by using attributes from the previous experiment supplemented by attributes from a new elicitation.

A summary of each phase in the work with its main content and its significance is given in this section.

3.1 Spatial attribute identification – generation and analysis of verbal data

The instigating work [17] addressed the use of semantic scales and alternative approaches to attribute identification and scaling. Issues discussed were how to reach a sufficient degree of communality in semantic scales, training of listeners and the problem with scales based on provided constructs. Comments were made on the nature of 'expert' knowledge and knowledge elicitation in general, as well as on the correspondence between perceived sensations and physical quantities. Different methods for handling multidimensional data were utilised.

The area of spatial attributes was considered as being under-investigated and the search for a method for construct generation with few assumptions about the outcome of an experiment resulted in an approach based on the repertory grid technique (RGT) being introduced as a tool for knowledge elicitation. The technique, used in personal construct psychology, was comprehensively explained, whereupon an experiment utilising features of the RGT was performed and analysed. The experiment's aim was to determine whether an approach based on RGT could produce relevant descriptors for reproduced audio. A number of subjects, experienced in listening to reproduced sound, listened to a wide variety of recorded sounds (outdoor environment, speech, pop music, etc), recorded and reproduced using different numbers of channels and recording techniques. The subjects made comparisons of the sound stimuli, in order to detect differences and similarities between the stimuli. During this process, the subjects were interviewed using a technique based on the RGT, and they were able to come up with a large number of personal constructs describing their auditory experiences during the listening. Every subject rated the different stimuli on his/her own recently elicited personal constructs. The ratings of the constructs were analysed, firstly for the individual listener and secondly, for the group as a whole. The results showed that a common set of attributes existed within the subject group. The multidimensional nature of perceived spatial quality was also confirmed. The broad attribute classes found were:

- Authenticity/naturalness
- Lateral positioning/source size
- Envelopment
- Depth

3.2 Structuring of descriptive verbal data

After conclusion of the work in [17] it was observed that a number of attributes were attitudinal, in other words expressing emotional responses or preference. If more attributes of a descriptive (and attitude-free) character were to be found, the presence of attitudinal constructs would possibly constitute 'noise' in the data. This could make it more difficult to discover patterns among the descriptive constructs. Therefore, the aim of this phase, reported in [23], was to extract information that was more detailed on the descriptive nature of the constructs elicited in [17], and thereby possibly find additional attributes.

This was done by separation of the constructs classified as descriptive from those being attitudinal by parts of an existing application of verbal protocol analysis (VPA). The analysis showed that two-thirds of all personal constructs from [17] were descriptive, whereas the remaining constructs were attitudinal. The personal constructs categorised as being descriptive were subjected to a cluster analysis, where the objective was to arrange similar descriptive constructs together in construct groups. The data on which the clustering was made consisted of the ratings made by the subjects in [17]. A number of construct groups were found. These construct groups were analysed for what sensation(s) the constructs within each group described. The analysis of the descriptive construct groups resulted in a number of new attributes and also some distinctions previously encountered:

- Localisation, left – right and front – back
- Depth/distance
- Envelopment
- Width
- Room perception
- Externalisation
- Phase
- Source width
- Source depth
- Detection of background noise
- Frequency spectrum

3.3 Attitudinal constructs

The initial purpose of [24] was to map attitudinal constructs on the descriptive constructs to explore which of the descriptive attributes contribute to certain attitudinal responses. The division of the constructs into descriptive and attitudinal categories were taken from [23]. The work comprised subdivision of the attitudinal constructs into an "emotive/evaluative" class and a "naturalness" class. During the work, several ways to describe attitudinal features of a sound were discovered. Constructs in the naturalness class showed to be made up of three groups:

- Natural/normal/real (or its opposite, unnatural/not common)
- Technical device involved (loudspeaker, microphone, recording)

- Feeling of presence (in the room or at the venue or its opposite, absence)

The emotional/evaluative group contained:

- Positive/negative (approval, disapproval)

The ambiguity in classifying verbal data was also noted, as some terms interpreted in [23] as attitudinal constructs in this analysis actually showed to be descriptors of spectral components, e g “sharp”, ”dull”.

The analysis showed that an enveloping sound gave rise to the most positive descriptors and that the perception of different aspects of the room was most important for the feeling of presence. Good localisation showed not to be most important for the feeling of naturalness.

It was also suggested for future work that when subjects are encouraged to describe what they perceive, either by free verbalisation methods or with more stringent questionnaires, a better understanding of the elements to which they are referring in a complex sound field is needed.

3.4 Validation of attribute scales

Next step was a validation experiment [25]. Its aim was to investigate the applicability of a selection of attributes encountered in the previous papers [17, 23, 24] within the context of partly new stimuli. The hypothesis tested was: If the attribute scales were relevant for evaluation of the spatial quality of a set of reproduced sounds, the subject group would be able use the scales to differentiate between some or all of the stimuli in the experiment at a significant level. Hence, the scales would have sufficient common meaning to the group,

The attributes defined in the preceding works were analysed for their applicability to the experiment. Some attributes were omitted as a result of this. The omitted attributes were “externalisation”, “phase” and “technical device”, since they were a result of the use of phase reversed signals in [17], and no phase reversed signals occurred in this subsequent experiment. “Externalisation” (to perceive sound as coming from outside one’s head in contrast to “internalisation” where the sound is perceived as coming from within the head) was also considered as being a dichotomous attribute hard to grade on the linear scales intended for the experiment.

The selected 12 attributes were compiled to a list with associated descriptions of their meanings. Based on the experience of the previous publication [24], the chosen attributes were divided into different classes depending on the part of the auditory scene to which they were relating. The stimulus set contained different reproduction modes of new sounds as well as sounds previously used for elicitation of personal constructs in [17]. A group of subjects, who had not participated in the previous experiment, rated the stimuli on the attribute scales provided.

The attributes all enabled the subject group to produce judgements significantly indicating differences between the stimuli. Hence, it was concluded that the attributes originally emerging from an elicitation of personal constructs conveyed an utilisable common meaning to the group of subjects. The strongest relationships between attributes were found between “envelopment” and the attributes “naturalness”, “presence” and “preference”. It was also observed that some attributes, e g “room size” and “room level” were less sensitive to different modes of reproduction.

For future experiments, it was proposed that measures should be taken to diminish the spatial difference between stimuli, in order to test the stability and sensitivity of the attributes. Another suggestion for future work was to re-iterate the elicitation process, since more knowledge exist about stimuli and perceivable dimensions, with the aim of finding more and new attributes.

An example of output from the validation experiment is shown in *Fig 3*, where four sound sources were reproduced in different modes: 5.1, 2-channel stereo and phantom mono) and, in addition to that, one of the sources was reproduced in one-speaker mono. In this graph, the attribute *envelopment* was assessed by the subject group.

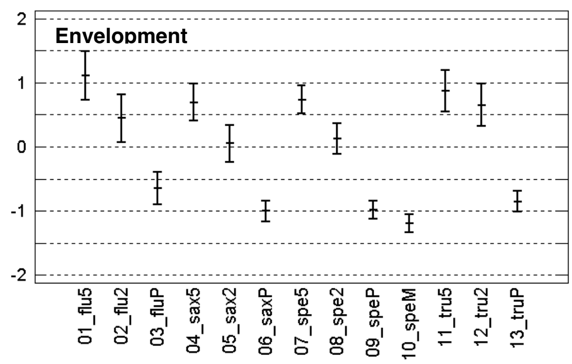

Fig 3. Mean values and 95% confidence intervals for the attribute “envelopment” for four sound sources (flute, saxophone, speech and trumpet) reproduced in 5.1 (5)*, 2-channel stereo* (2)*, phantom mono* (P) *and centre-speaker mono* (M)

3.5 Final experiment

In [26], the purpose was to test the method and the attributes used in the work up to this point. A new set of stimuli, intended to embrace smaller differences in spatial quality than the sets previously used by the authors, was created. The stimulus set consisted of two events, a single source (a viola) and a dual source (vocal + piano), each recorded in a hall with five different 5-channel microphone techniques, thus yielding 10 stimuli, all reproduced by means of an ITU-R BS. 775 5-channel system.

The stimuli were used in an elicitation experiment, where a number of personal constructs were elicited. As a result of the elicitation, a few new attributes were found. Furthermore, some attributes from the previous experiment [25] were removed from the list because they either were regarded as referring to non-spatial percepts, or were inconsistently used in [25]. The new observations led to amendments of the attribute list from [25]. The new list was tested for its significance for the stimuli used in the elicitation experiment. This was performed by a group of subjects, who rated all stimuli on the attributes included in the attribute list.

Also in this experiment, all the attributes enabled the subject group to produce judgements significantly indicating differences between the stimuli. An example is shown in *Fig 4*. It was discovered that this experiment as well as the previous one [25] contained data suggesting that the perception of room properties are perceived on two dimensions, one relating to sensation/impression of presence, and another relating to judgement of certain room characteristics, like the perceived room size and the level of the reflected sound in the room. This is shown by the plot in *Fig. 5*, where the room attributes were subjected to factor analysis and two factors were extracted.

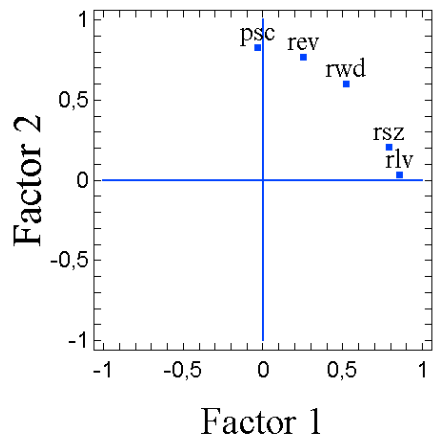

Fig. 5: Factor loadings of room attributes only. Two factors were extracted. Rotation: Varimax. Factor 1 represents a judgement dimension, while factor 2 relates to a sensation/impression dimension.

Considering the dimensionality of the whole data set, the attributes seem to be perceived mainly in the dimensions “source width”, “distance to the source” and “sense of presence in the room/hall”.

In the discussion, it was pointed out that an elicitation of constructs performed without constraints on the elicitation process generally could produce any type of constructs, also non-spatial ones, e g those referring to the frequency spectrum of the sound.

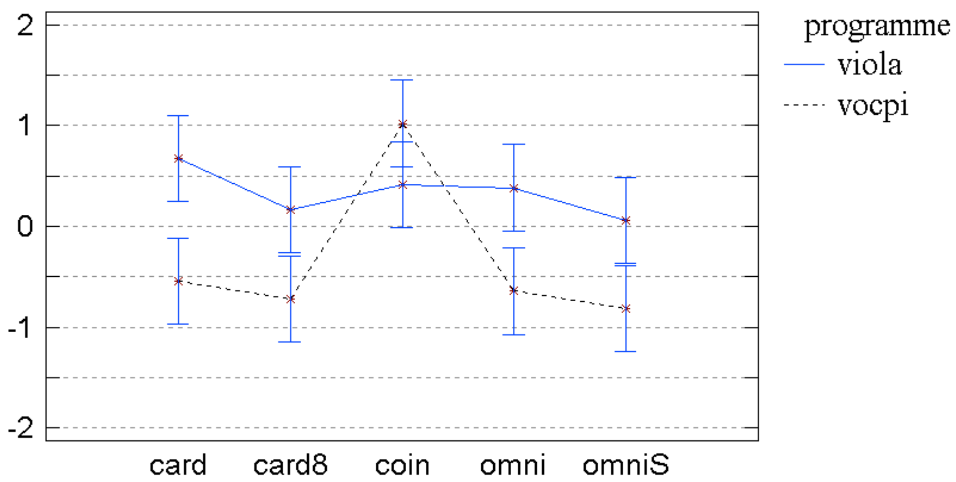

Fig 4: Mean scores and error bars for attribute “localisation1”in the final experiment where two programme types (viola and vocal + piano) were recorded by five different 5-channel microphone set-ups. Reproduction was made through a ITU-R BS. 775 5-channel system. The attribute refers to the ability to pinpoint the perceived location of the music instruments. As one example of the output from the experiment, a significant improvement in localisation of the piano for the coincident microphone technique (coin) is visible in the graph.

4 A METHOD FOR EVALUATION

The work contained in the publications in the foregoing section described a method for the evaluation of perceived spatial quality in multichannel systems. The application of the method resulted in a number of attributes being defined and validated as operational for evaluation within a certain context. In order to accentuate the method, this and the attributes resulting from it are summarised below.

4.1 Observations on the evaluation method

A number of observations were made throughout the work. Some of them were explicitly stated in the publications, whereas other may have been more implicitly suggested. In a condensed form, the observations of importance for the work were:

- Elicitation using a technique based on the repertory grid technique was successful for collecting information on perceived differences and similarities between stimuli [17, 26].
- Structuring, by means of multivariate methods, of the elicited descriptors (personal constructs) rated by the elicitee, indicated perceptual patterns that were used for interpretation of perceived sensations encountered [17, 23, 24].
- Verbal protocol analysis aided the structuring by sorting descriptors into different categories for separate analysis [23, 24].
- It was possible to construe relevant attributes with associated descriptors from data acquired through elicitation and structuring [17, 23, 24, 25, 26].
- The attributes resulting from previous experiments gave significant results when used by a new group of listeners and for partly new stimuli, which validated the attributes as suitable descriptors for spatial quality of reproduced audio in the current domain [25].
- Certain attributes were defined to apply only to parts of the auditory scene, e g the source(s) and the space [25, 26]. This avoided confusion among the subjects about to what the attributes referred.
- Conditions for the modification and discarding of attributes were applied [25, 26]:
  - inapplicability to the context of spatial audio
  - inapplicability to linear scales (if such are used)
  - low listener consistency in rating
- The attributes also produced significant results despite a change in domain – from stimuli differing in modes of reproduction to stimuli recorded with different surround sound microphone techniques [26].

4.2 Framework of the evaluation method

The work within this study implicitly conveys a method for evaluation of perceived quality of spatial audio. This method comprises a number of successive steps, starting with a context definition and ending with the actual evaluation of sound stimuli. In this section, the method is presented in a step-by-step fashion where the different steps and their results in this study are emphasised. This is also depicted in the block diagram in *Fig. 6*.

1. *Context definition and purpose of test:* The purpose was defined: investigation of what listeners could perceive in terms of spatial attributes in general, when listening to a variety of audio reproduction systems, including surround sound.
   Results: Stimuli requirements were defined: a set of stimuli was selected that embraced a variety of recording and reproduction techniques, as well as spatially processed recordings. Listeners with experience in listening to reproductions of sound were recruited.
2. *Selection of stimuli:* Six different sounds were either recorded or copied from existing recordings. Each sound was manipulated, either by means of the recording or reproducing technique, or electronically processed, with the purpose of creating spatial differences. This yielded three versions per source.
   Results: A stimulus set comprising a total of 18 (6 sources x 3 versions) were produced.
3. *Elicitation and rating of constructs:* The subjects compared the three versions of a sound source and were encouraged to verbally express the perceived differences and similarities between them, using a structured knowledge elicitation technique. These descriptions formed the personal constructs that were documented. Each subject subsequently made ratings of a selection of the stimuli on all of his/her own constructs.
   Results: About 350 elicited constructs and associated ratings were produced.
4. *Structuring of constructs:* Cluster analysis, principal component analysis and verbal protocol analysis were made on the ratings in order to separate attributes into descriptive and attitudinal classes, as well as reduce the dimensionality and remove redundancy.
   Results: 15 construct groups were identified. Eleven of them were categorised as descriptive and four as attitudinal.
5. *Definition of attributes and construction of scales:* The construct groups were analysed for their content. Appropriate descriptions for the attributes found in the construct groups were formulated. (Attributes generated and defined by the experimenter, provided attributes, may also be added, but were not in this study.) Rating scales were defined.
   Results: A set of attribute scales, in the form of a written definition was created.

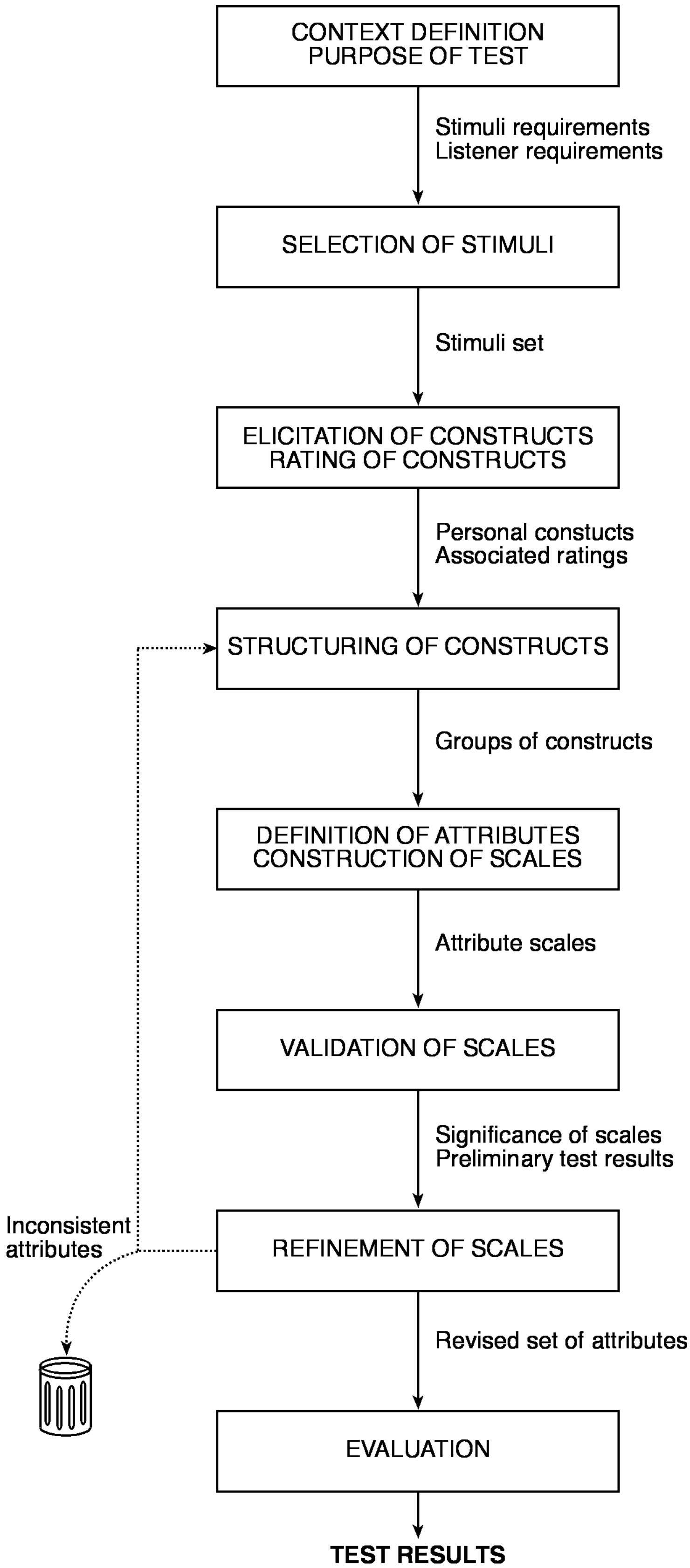

Fig. 6: Evaluation method (block diagram)

6. *Validation of scales:* A (pilot) experiment for testing the significance of the attribute scales as well as the experimental design was performed, with the objective of testing if a, in this case a new, group of subjects were able to make statistically significant judgements on the scales.
   Results: Statistical measures of the significance of attributes, e g in the form of F ratios from an analysis of variance, were acquired. This resulted in all attribute scales being found significant.
7. *Refinement of scales*:* An analysis of the results of the validation experiment was performed. Because of this, some attributes were removed. In addition, another condition for discarding attributes was employed as well; two attributes were considered as non-spatial. In the event of a discovery of an in-consistent attribute (i e an attribute not giving rise to significant judgements or showing high error variance), there was also the possibility to return to the structuring stage in this model to analyse if this particular attribute could have been less ambiguously defined. Some amendments were done, relating to which part of the auditory event the attributes referred to, e g "envelopment" was split in "source envelopment" and "room envelopment"
   Results: A revised set of attribute scales was defined.
8. *Evaluation:* A new experiment, similar to the validation experiment, was performed. The data was analysed and it showed that the scales employed were functional as means of finding significant differences between the stimuli.
   Results: Differences between the stimuli on the attribute scales were found.

*) In the last experiment executed in this study, the stimuli selection was altered. To focus on differences occurring in a surround sound system, 5-channel recordings were used with differences in recording technique and in the number of sources. To ensure that this had not introduced new perceived sensations that were not accounted for by the original attributes, a new elicitation was performed prior to the experiment. The original attribute set was mainly confirmed as a result of the new elicitation.

4.3 Attributes

The attributes resulting from this study have been modified through the series of experiments due to new insights and additional elicitations. The attributes used in the last experiment [26] were all found to be significant for differentiating the stimuli. *Table 1* shows the attributes and their associated descriptions. Apart from these attributes, others have been used, and a list of all attributes collected up to the validation experiment is found in [25]. Some additional observations were:

- Attributes referring to the space (the room/hall) seem to be judged independently of the type of source in most cases [26].
- The last experiment [26] as well as the preceding one [25] contained data suggesting that the perception of room properties are perceived on two dimensions, one relating to sensation/impression of being present at the venue and another relating to judgement of certain room characteristics, like the perceived room size and the level of the reflected sound in the room
- Considering the dimensionality of the whole data set in [26], the attributes seem to be perceived mainly in the dimensions "source width", "distance to the source" and "sensation of presence in the room/hall". A cautious comparison with research on concert hall acoustics, based on the verbal descriptors suggests that "source width" and Apparent Source Width (ASW) seem to be similar as well as "sensation of presence" and Listener Envelopment (LEV).
- It showed that an enveloping sound gave rise to the most positive descriptors and that the perception of different aspects of the room was most important for the feeling of presence. Good localisation showed not to be most important for the feeling of naturalness [24].

5 DISCUSSION

5.1 Attributes

In this study, a number of attributes were defined and validated as operational for evaluation of certain natural sounds in a reverberant environment reproduced through a ITU-R 775 surround sound system. The general validity of the attributes found could be confirmed by comparing them with attributes employed by other authors, like Zacharov and Koivuniemi [36], Toole [37] and Gabrielsson *et al* [8, 38, 39]. Such a comparison is far from straightforward due to a number of difficulties in interpretation: shift in accuracy due to translation, domain of application, types of scales used and overlapping. Some attributes may also be multi-dimensional. Although these problems may exist, tentative similarities on several counts can be observed. Where similarities are implied, these are analysed in the subsections. The result of the comparison is shown in *Table 2*.

5.2 Conclusions

The method described has been shown to produce statistically significant results in evaluation of different modes of spatial reproduction and different microphone techniques. Despite changes of subjects and stimuli, the attributes on which the scales are based seem valid and reliable in the context of evaluating the spatial quality of surround sound reproductions of stationary, naturally occurring sound sources in reverberant spaces, recorded acoustically without using artificial multitrack mixing. This reinforces the strength of attributes originating in constructs elicited from listeners. There are no data available for direct comparison to support the superiority of attributes generated this way, but the fact that all attributes have showed to be highly significant indicates the power of this method.

However, the attributes cannot automatically be expected to work in new contexts, as these might not excite sensations described by the current set of attributes. Examples of contexts not applied in this work are artificial environments, binaural recordings, moving sources and film sound tracks.

The method is new for evaluating spatial quality and comprises several stages of structuring and validation. When more is learned about the contexts and the attributes, certain stages may be omitted or reduced due to an eventual stabilisation of the attribute set.

The evaluation method is also believed by these authors to be expandable to other domains of quality assessments. Since the attribute scales have their origin in the elicitation sessions where the constructs are generated from differences and similarities in the stimulus set, the choice of stimuli substantially influences the scale design.

Nevertheless, the method produces highly significant results and it should be considered viable for the evaluation of spatial quality in surround sound systems.

Attribute	**Description**
Naturalness	How similar to a natural (i.e. not reproduced through e g loudspeakers) listening experience the sound as a whole sounds.
Presence	The experience of being in the same acoustical environment as the sound source, e g to be in the same room.
Preference	If the sound as a whole pleases you. If you think the sound as a whole sounds good. Try to disregard the *content* of the programme, i e do not assess genre of music or content of speech.
Low frequency content	The level of low frequencies (the bass register).
Ensemble width	The perceived width/broadness of the ensemble, from its left flank to its right flank. The angle occupied by the ensemble. The meaning of "the ensemble" is all of the individual sound sources considered together. Does not necessarily indicate the known size of the source, e g one knows the size of a string quartet in reality, but the task to assess is how wide the sound from the string quartet is perceived. Disregard sounds coming from the sound source's environment, e g reverberation – only assess the width of the sound source.
Individual source width	The perceived width of an individual sound source (an instrument or a voice). The angle occupied by this source. Does not necessarily indicate the known size of such a source, e g one knows the size of a piano in reality, but the task is to assess how wide the sound from the piano is perceived. Disregard sounds coming from the sound source's environment, e g reverberation – only assess the width of the sound source.
Localisation	How easy it is to perceive a distinct location of the source – how easy it is to pinpoint the direction of the sound source. Its opposite is when the source's position is hard to determine – a blurred position.
Source distance	The perceived distance from the listener to the sound source.
Source envelopment	The extent to which the sound source envelops/surrounds/exists around you. The feeling of being surrounded by the sound source. If several sound sources occur in the sound excerpt: assess the sound source perceived to be the most enveloping. Disregard sounds coming from the sound source's environment, e g reverberation – only assess the sound source.
Room width	The width/angle occupied by the sounds coming from the sound source's reflections in the room (the reverberation). Disregard the direct sound from the sound source.
Room size	In cases where you perceive a room/hall, this denotes the relative size of that room.
Room sound level	The level of sounds generated in the room as a result of the sound source's action, e g reverberation – i e not extraneous disturbing sounds. Disregard the direct sound from the sound source.
Room envelopment	The extent to which the sound coming from the sound source's reflections in the room (the reverberation) envelops/surrounds/exists around you – i e not the sound source itself. The feeling of being surrounded by the reflected sound. .

Table 1: Attributes in the final evaluation experiment.

It can be concluded that:

- Personal constructs can be elicited from listeners in the context of reproduced audio
- A common perceptual pattern exists, expressed as attributes
- A list of attributes has been derived that are meaningful and valid
- Attribute scales derived from personal constructs enabled a group of listeners to make statistically significant judgements on reproduced sounds.

The implication of these conclusions is that:

- A method utilising elicited personal constructs can be used for finding a set of attributes with sufficient common meaning, thus enabling a group of experienced listeners to make significant judgements of spatial quality in surround sound systems.

5.3 Further work

A number of ways are open for applications of the work so far. Regarding the method, refinements of the different steps are possible. It was noted that some subjects used gestures in their communication of certain sensations perceived, especially those related to source positions and width. A graphical interface for such attributes may be a useful complement to verbal elicitation.

An obvious direction of the work is a continued refinement of the method by expansion of the stimulus set to explore its shortcomings and to see if generalisation of results across a greater variety of sound excerpts is possible.

To find the applicability of the method for product evaluation tasks, the comparison of different system components can be made, e g loudspeakers, spatial enhancers, etc.

When the method has been employed a number of times, the attribute set resulting from new experiments might have reached such level of refinement that mapping of physical parameters onto them is possible. In that case, objective measures can be tested for their feasibility as predictors for certain perceived sensations.

An area not analysed so far concerns listener behaviour and training. Definition of what constitutes an experienced listener in the context of spatial audio may be addressed. Also detection of statistical outliers, the reason for their occurrence and the treatment of them are topics relevant for further work.

Finally, the evaluation method described in this study would be simpler to employ if software could be developed to assist all steps in the process. Such work could also be considered.

6 ACKNOWLEDGEMENTS

A number of people directly or indirectly contributed to the research work summarised in this paper. The authors wish to thank the members of the Eureka project 1653 group for their support and useful comments during the work with the quoted publications. The staff at Swedish

Berg	Zacharov & Koivuniemi	Toole	Gabrielsson *et al*
Low frequency content			
Naturalness	Naturalness	Perspective*	Fidelity
Preference			
Presence	Sense of space	Perspective*	Feeling of space
Ensemble width	Broadness*	Width of the sound stage	
Localisation	Sense of direction	Definition of sound images	
Source envelopment	Broadness*		
(Individual) Source width	Broadness*	Definition of sound images	
Source distance	Distance to events	Impression of distance	Nearness
Room envelopment	Broadness*	Reproduction of ambience, spaciousness and reverberation	
Room size			
Room level			
Room width	Broadness*		
	Sense of movement		
Externalisation (opposite)**	Penetration	Abnormal effects	
	Depth		
		Continuity of the sound stage	

* Denotes multiple occurrences of the attribute in the table's column.
** Attribute found in [40] and reported in [25].

Table 2: Comparison of attributes published by some authors

Radio has contributed with general support of whom Jonas Ekeroot is especially thanked for computer programming and for solving a diversity of practical issues. Oscar Lovnér and Andreas Renhorn are thanked for making the recordings of sound stimuli used in the experiments. Students at the School of Music at Piteå have participated in the experimental work, both as performers and as patient subjects. These are also thanked for their contribution. Søren Bech, Willam L Martens, Wieslaw Woszczyk and Anders Ågren are thanked for examining the work.

A number of organisations have made this work possible through funding support, and they are hereby thanked: The School of Music at Piteå, Centek, Acusticum and the Town of Piteå.

7 REFERENCES

1 Watkinson, J. (1994): *The art of digital audio*. Focal Press, Oxford.

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

3 Nakayama, T., Miura, T., Kosaka, O., Okamoto, M. and Shiga, T. (1971): Subjective assessment of multichannel reproduction. *Journal of the Audio Engineering Society* **19**, pp 744-751.

4 Morimoto, M. (1997): The Role of Rear Loudspeakers in Spatial Impression. Presented at *103rd AES Convention*. Preprint 4554. Audio Engineering Society.

5 Hawkes, R. J. and Douglas, H. (1971): Subjective acoustic experience in concert auditoria. *Acustica* **24**, p 235.

6 Martin, G., Woszczyk, W., Corey, J. and Quesnel, R. (1999): Sound source localization in a five-channel surround sound reproduction system. Presented at *107th AES Convention, New York*. Preprint 4994. Audio Engineering Society.

7 Berg, J. (2001):, *Report on the performance of the Coding Technologies AACplus audio codec in comparison with MPEG-4 AAC.* Audio subgroup, ISO/IEC JTC1/SC29/WG11 MPEG 2001/6720. International Organisation for Standardisation.

8 Gabrielsson, A., Lindström, B. and Elger, G. (1983): *Assessment of perceived sound quality of eighteen high fidelity loudspeakers*. Report TA No. 106. Technical Audiology, Karolinska Institutet, Stockholm.

9 Bech, S., Hansen, W. and Woszczyk, W. (1995): Interactions between audio-visual factors in a home theater system: experimental results. Presented at *AES 99th Convention, New York. Preprint 4096.* Audio engineering Society.

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

11 Bech, S. (1999): Methods for subjective evaluation of spatial characteristics of sound. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, pp 487-504. Audio Engineering Society.

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

13 Letowski, T. (1989): Sound quality assessment: concepts and criteria. Presented at *87th AES Convention, New York*. Preprint 2825. Audio Engineering Society.

14 Pratt, R. L. and Doak, P. E. (1976): A subjective rating scale for timbre. *J. Sound and Vibration* **45**, p 317.

15 Grey, J. M. (1977): Multidimensional perceptual scaling of music timbres. *J. Acoust. Soc. Amer.* **61** pp. 122-135.

16 Staffeldt, H (1984): Measurement and prediction of the timbre of sound reproduction. *J. Audio Eng. Soc.* **32**, pp. 410

17 Berg, J. and Rumsey, F. (1999) Spatial attribute identification and scaling by Repertory Grid Technique and other methods. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, 10–12 Apr. pp 51-66. Audio Engineering Society.

18 Mason, R. and Rumsey, F. (2000): An assessment of the spatial performance of virtual home theatre algorithms by subjective and objective methods. Presented at *108th AES Convention*. Preprint 5137. Audio Engineering Society.

19 Zacharov, N. and Huopaniemi, J. (1999): Results of a Round Robin Subjective Evaluation of Virtual Home Theatre Sound Systems. Presented at *AES 107th Convention*. Preprint 5067. Audio Engineering Society.

20 Rumsey, F. (1998): Subjective assessment of the spatial attributes of reproduced sound. In *Proceedings of the 15th International conference on audio, acoustics and small spaces*, pp 122-135. Audio Engineering Society.

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

22 Berg, J (2002): *Systematic evaluation of perceived spatial quality in surround sound systems.* PhD thesis. Luleå University of Technology, Sweden.

23 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.

24 Berg, J. and Rumsey, F. (2000) Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction. Presented at AES 109th Convention, 25-22 September, Los Angeles. Preprint 5206.

25 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.

26 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.

27 Borg, I and Groenen, P. (1997): *Modern multidimensional scaling.* Springer-Verlag, New York.

28 Everitt, B. S. and Dunn, G. (1991): *Applied Multivariate Data Analysis.* Edward Arnold, London

29 Zacharov, N. and Koivuniemi, K. (2001): Unravelling the perception of spatial sound reproduction: Techniques and experimental design. In *Proceedings of the AES 19th International Conference on Surround Sound,* pp 272-286. Audio Engineering Society.

30 Olive, S. (2001): Evaluation of five commercial stereo enhancement 3D audio software plug-ins. Presented at *AES 110th Convention, Amsterdam*. Preprint 5386. Audio Engineering Society.

31 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.

32 Ford, N., Rumsey, F., de Bryun, B. (2001): Graphical elicitation techniques for subjective assessment of the spatial attributes of loudspeaker reproduction – a pilot investigation. Presented at *AES 110th Convention, Amsterdam*. Preprint 5388. Audio Engineering Society.

33 Beranek, L. (1996): *Concert and opera halls – how they sound*. Acoustical Society of America.

34 Shaw, M. and Gaines, B. (1995): *Comparing conceptual structures: consensus, conflict, correspondence and contrast.* Knowledge Science Institute, University of Calgary.

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

36 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.

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

38 Gabrielsson, A., Hagerman, B. and Bech-Kristiansen, T. (1991): *Perceived sound quality of reproductions with different sound levels*. Report TA No. 123. Technical Audiology, Karolinska Institutet, Stockholm.

39 Gabrielsson, A., Hagerman, B., Bech-Kristiansen, T. and Lundberg, G. (1988): *Perceived sound quality of reproductions with different frequency responses and sound levels*. Report TA No. 117. Technical Audiology, Karolinska Institutet, Stockholm.

40 Berg, J. and Rumsey, F. (1999): Identification of perceived spatial attributes of recordings by repertory grid technique and other methods. Presented at *AES 106th Convention, Munich*. Preprint 4924. Audio Engineering Society.