Spatial audio and sensory evaluation techniques – context, history and aims

Francis Rumsey[1]

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

ABSTRACT

Spatial sound reproduction gives rise to new challenges for those trying to evaluate sensory features contributing to perceived quality. Recent technical developments have enabled the delivery of sophisticated multichannel audio signals to consumers, over links that range very widely in quality, requiring decisions to be made about the trade-offs between different aspects of audio quality. Spatial factors can account for as much as a third of overall ratings of sound quality in listening tests and must therefore be considered seriously in systems and tests that evaluate sound quality. It is therefore important to determine the most important spatial quality attributes of reproduced sound fields and to find ways of predicting perceived sound quality on the basis of objective measurements.

1. INTRODUCTION

The aim of this paper is to establish the context within which spatial audio and sensory evaluation can be discussed in this seminar, and to summarise the current state of the art. Subsequent authors will have the task of developing the ideas introduced here.

Examples of foregoing work will be summarized, in which the importance of the spatial aspects of sound quality is established. This is to show that it is a topic towards which research has been pointing for some time. The difference between reproduced sound and concert hall acoustics gives rise to a proposal that reproduced sound might require some novel approaches to sensory evaluation, or at least that established practices in acoustics do not necessarily provide all the techniques we need to evaluate reproduced sound. In recent years there has been renewed interest in finding out what listeners hear when evaluating spatial audio signals, and a variety of techniques have been adapted from other areas of psychology and sensory science.

It also has to be asked why one should care about spatial audio quality. Do listeners give any weight to it in their overall evaluation of sound quality, and how does it affect their liking for certain types of sound? There is still considerable work to be done to establish such relationships, although some initial evidence is available. There is a need to decide whether a common terminology is needed in this field of expert knowledge in order that those working in the field may share their understanding and compare results. Finally, it is also necessary to consider how 'objective' measurement models can be developed that are capable of predicting listener responses on the basis of physical evaluations of reproduced sound fields.

2. CONTEXT

It is specifically in the context of reproduced sound that we are interested in the topic of spatial quality. Although considerable work has been undertaken in concert hall acoustics, reproduced sound brings its own special set of challenges. The last few years has seen a dramatic increase in the range and variety of spatial audio systems on the market, and the wide acceptance of home cinema has made spatially sophisticated sound a reality in the homes of many consumers. It is becoming common to find some form of spatial audio processing and reproduction in mobile devices and it is possible to deliver multichannel audio to the consumer with a range of qualities that spans the very poor to the excellent.

High technical quality or fidelity, it can be argued, may be taken for granted at this point in the history of audio engineering. Although not all audio devices exhibit the highest technical quality, the technical quality of the best sound reproduction available to the consumer exhibits very low levels of distortion, wide frequency range, flat frequency response and low noise, with specifications that match or exceed the limits of human perception. 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. Spatial quality, on the other hand, has some way to go before the curve could be said to be asymptotic to some ideal.

Spatial audio coding systems aim to reduce the data rate required to deliver multichannel audio over communications networks and it would be useful to have a means by which their effect on spatial reproduction quality could be reliably evaluated. Objective measurement systems that aim to evaluate the 'mean opinion scores' of audio coding systems do not currently take the spatial aspect of such processes into account, even though, as discussed below, it is possible for spatial fidelity to account for as much as 30% of the mean opinion score. As scalable coding systems become increasingly common as a means of delivering spatial audio content over networks with different bandwidths, it is likely that trade-offs will have to be made between different aspects of sound quality when deciding how to scale the process. It will therefore be important to know the relationship between spatial aspects of sound quality and others aspects of the same, as well as to know the weight of spatial quality attributes in the overall quality score.

An understanding of such matters will also be of vital importance in the field of computational auditory scene analysis (CASA) and its partner virtual reality (VR), in which reliable perceptual descriptors and physical correlates of spatial scene attributes are needed for parametric representation and synthesis.

3. WHAT IS SPATIAL QUALITY?

It is hard to get much further in this paper before discussing what is included within the domain of spatial sound quality. A review of the literature relating to spatial quality evaluation in its broadest sense reveals a subtle but crucial division between two different concepts of the term spatial. Put simply it relates to the distinction between 'attributes of spaces' and 'spatial attributes'. In much of the literature relating to concert hall acoustics or the acoustics of enclosed spaces, the attributes that are used to evaluate 'spatial' quality are often parameters that relate to the qualities of the space in question, such as reverberance, warmth, intimacy, and so on. Zacharov and Koivuniemi [1] and Berg [2] review a number of the terms that arise from such studies, and it is clear that only some of them are really what this author would term 'spatial attributes', and that could be related to the evaluation of sound reproduction.

In [3] we attempted a definition of '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'. This definition is not entirely satisfactory, though. In [4] we have also described the search for valid spatial attributes as being primarily concerned with 'the three-dimensional nature of sound sources and their environments', which is possibly closer to the mark. Both these attempts at definitions imply that we are concerned with those perceptual constructs that relate to directionality, size, (height), depth and width, of reproduced sources, groups of sources and acoustical environments. In other words we are concerned to describe and evaluate the three dimensional characteristics of the components of a spatial audio scene that is reproduced using loudspeakers or headphones. In certain contexts there may also be higher level attributes to consider, such as spatial naturalness, presence, envelopment or immersion.

4. DOES SPATIAL QUALITY MATTER?

If spatial audio quality does not matter to listeners then there is little point in going much further, so it seems important to present some evidence that it does matter.

Within the field of concert hall acoustics, Korenaga and Ando [5], for example, found that interaural cross correlation (IACC) was one of four important parameters affecting listeners' preferences for concert hall seats. IACC has been consistently linked in the literature with the spatial perception of source width and envelopment.

In reproduced sound Toole [6], for example, found that loudspeaker fidelity ratings and spatial quality ratings were quite highly correlated (r = 0.7) but did not quantify the relative contributions of the different quality factors he tested to the overall fidelity ratings. Gabrielsson and Lyndstrom [7] also evaluated a number of attributes in terms of their validity for describing Perceived Sound Quality (PSQ). They found both spatial and timbral attributes to be valid and moderately highly correlated with PSQ, but did not attempt to quantify their relative importance. In [8] we reported some observations about the relationships between basic audio quality, spatial and timbral fidelity. These showed that timbral fidelity and two spatial fidelity scales were correlated at a relatively low level (0.33 for frontal spatial fidelity and 0.26 for surround spatial fidelity). It was observed that basic audio quality seemed to be more strongly influenced by timbral fidelity than by spatial fidelity but that spatial fidelity contributed an important component.

Recently [9,10] we published two papers in which we attempted to quantify the contribution of spatial fidelity to overall judgments of reproduced sound quality, for both experienced and naïve listeners. Since these experiments were conducted in the context of consumer home cinema involving 5.1-channel surround sound that had been altered in quality by band-limitation and downmixing, we evaluated spatial quality on two spatial fidelity scales, namely 'frontal spatial fidelity' and 'surround spatial fidelity'. These enabled listeners to compare the spatial similarity of the front component and the surround component of the spatial scene with an unimpaired reference reproduction. They also evaluated timbral fidelity and a regression model was developed to show the relationship between the different fidelity ratings and ratings of 'basic audio quality' (BAQ) that had been gathered previously. The spatial and timbral degradations judged by the listeners were designed to have comparable effects on the overall 'information rate' of the audio signal in the digital domain, so could be said to have some comparability in terms of their demands on delivery bandwidth. The outcome of the analysis showed that spatial fidelity contributed a substantial component of the overall BAQ judgement, as shown in the following equation:

$$BAQ = 0.80\,Timbral + 0.30\,Frontal + 0.09\,Surround - 18.7$$

The above equation was derived from ratings given be experienced listeners, and suggests that although timbral fidelity dominates the BAQ rating, frontal spatial fidelity has an important part to play, with less significance attributed to surround spatial fidelity. Overall, though, the spatial component contributed around a third of the overall BAQ rating.

It would seem reasonable, summarizing even this small number of studies, to suggest that spatial audio quality is important to listeners and is capable of contributing a large enough proportion of overall quality ratings to be taken seriously.

5. WHAT IS UNIQUE ABOUT REPRODUCED SOUND?

Although work undertaken in concert halls is a useful starting point for the evaluation of reproduced sound, there are some good reasons why reproduced sound has some special requirements. Although many of the features of natural environments and spatial listening may be present in reproduced sound, there are a number of unique properties of each, and the cognitive tasks, context and concepts involved may be somewhat different.

In concert halls spatial attribute evaluation had to do with the effects of room reflections on the spaciousness of sources and reverberation. In reproduced sound this may also be interesting, but there are many other factors such as the panning and processing of multiple sound sources, the creation of novel spatial effects and the possibility to place sources anywhere around the listener.

In classical music recording and other recording genres where a natural environment is implied or where a live event is being relayed it is often said that the aim of high quality recording and reproduction should be to create as believable an illusion of 'being there' as possible. This implies fidelity to a remembered reference in terms of technical quality of reproduction, and also fidelity in terms of spatial quality. Others have suggested that the majority of reproduced sound should be considered as a different experience from natural listening, and that to aim for accurate reconstruction of a natural sound field is missing the point – consumer entertainment in the home being the aim.

The ability of spatial sound systems to recreate accurately localised sources in all three dimensions is regarded by many as the 'holy grail' of stereophonic reproduction and the evaluation of perceived sound source location is often the only consideration in perceptual experiments. If true identity were possible between recording environment and reproducing environment, in all three dimensions and for all listening positions, then the ability of a recording–processing–reproducing system to render accurate images of all sources (including reflections) would be the only requirement for spatial fidelity. The need for subjective testing would be eliminated as a result and there would be no need for a discussion such as this. True identity however, is not currently possible and may never be, for a variety of practical and technical reasons. Neither is it necessary to render every reflection accurately in order to obtain a perceptually convincing impression of diffuse reverberation, for example, enabling complexity reductions to be made in practical spatial audio rendering systems [11,12]. Real spatial audio signal chains, from original source to listener, always involve trade-offs and design compromises of one sort or another, which makes subjective testing and comparison necessary and desirable.

The primary aim of most commercial media production is not true spatial fidelity to some notional original sound field, although a mixing engineer might choose to create spatial cues that are consistent with those experienced in natural environments. In a large number of commercial releases there is no natural environment to imply or recreate and one is dealing with an artificial creation that has no 'natural' reference or perceptual anchor. Here the acoustic environment implied by the recording engineer and producer is a form of 'acoustic fiction' or 'acoustic art'. Reproduced sound also enables the introduction of effects not encountered in natural listening, such as out of phase elements. It also brings with it the challenge to evaluate highly complex and changing spatial audio scenes, containing elements that may not have a direct parallel in natural listening or which may be mutually contradictory (dry and reverberant sources could be combined within a single mix, for example).

Even if a reproduced spatial scene is unnatural, unfamiliar or fictitious, it is possible to *compare* versions of spatial reproduction (or scene renderings in VR terms), such as might arise from using different recording techniques, forms of signal processing or reproduction configurations. One can describe their *relative* quality and/or character in terms of differences in magnitudes of clearly defined attributes. It is also possible to talk in terms of desirable and undesirable, or appropriate and inappropriate, spatial qualities. One must also bear in mind the possibility for reproduced sound to be 'hyper-real' – that is having spatial cues that are exaggerated or not naturally occurring. As virtual environments and augmented reality become more common, our concepts of naturalness may be forced to change – after all naturalness is mainly related to familiarity.

6. WHAT DO LISTENERS HEAR WHEN EVALUATING SPATIAL AUDIO QUALITY?

In order to explore the percepts arising in the minds of listeners when comparing spatial audio stimuli, a number of recent studies have used structured elicitation techniques. These enable listeners to take part in the definition of scales that they may subsequently use to evaluate sound quality. Such studies also ensure that any such scales are directly related to the stimuli in question. Such techniques are based in psychological methods used for other applications, such as Repertory Grid Technique [4], on sensory evaluation approaches used in the food industry, such as Descriptive Analysis [13], and on techniques from psycho-linguistics, such as Perceptual Structure Analysis [14]. Because spatial audio stimuli are not always easily described using verbal language, some researchers have experimented with the use of graphical languages and response formats, as an interesting alternative [15,16].

Interestingly, although perhaps not entirely surprisingly, and despite the wide range of complex stimuli employed in these experiments, the perceptual attributes elicited from subjects show a remarkable degree of similarity, suggesting that a common underlying perceptual structure for spatial audio quality evaluation may not be too elusive. There exists, of course, the difficulty of inter-language translation of terms and the problem of knowing whether one subject is describing the same feature as another, but simply using different words. However, many of these techniques have ways of getting at the underlying perceptual similarities between terms, by looking at common rating patterns among them, for example. A summary of all these attributes will not be attempted here, but some interesting features of the different results will be highlighted.

Firstly it can be seen that listeners commonly make a lot of references to width-related attributes, but they need a means to distinguish between the widths of individual sources and those relating to the distribution of a group of sources, which we may call an ensemble [17]. Similarly, there is a need for them to be able to separate the discussion of reverberant environments from that of sources. The width of an environment can be perceived separately from that of a source [18].

Secondly, listeners pay a lot of attention to envelopment-related attributes, weighting them highly in relation to enjoyment, naturalness and 'presence'. They distinguish, however, between being enveloped by sources and being enveloped by diffuse reverberant sound [2].

Thirdly, although they can detect distance-related features in a reproduced sound scene, the concept of depth is often problematic for listeners – it being difficult to perceive the depth of a source, an ensemble or a whole scene [19]. Whether this is due to some limitation of spatial audio reproduction systems, to limitations in the response format or to limitations in the perceptual mechanism is not clear. Some distance-related attributes can be perceived even with mono reproductions (because of cues such as direct to reverberant ratio), whereas a true perception of depth seems to require truly three-dimensional reproduction and perception. One only has to consider the difference between a mountain scene depicted on a postcard, in which the relative distances of different mountains can be determined by means of perspective cues and 'flat' information, and the real experience of standing in a mountain scene where the depth of the scene opens up before one and a true sense of three-dimensional depth is perceived. This has to do with parallax cues, binocular vision and the various stereoscopic cues that arise when presented with a truly three-dimensional stimulus. The former could be considered as enabling the judgment of the relative distance of objects, whereas the latter might be considered as enabling the experience of a deep scene. There is, therefore, perhaps a difference between the ability to judge the three dimensional positions of objects and the experience of spatial naturalness and reality. The postcard is a pale imitation of the real thing, but retains some of its perceptual cues.

Fourthly, localization quality crops up quite regularly as a factor to consider. Subjects often comment on the ease or difficulty of localizing sources – something that may be a key differentiator between good and bad spatial audio systems. This often turns out to be closely related to attributes such as source width and 'focus', but is is in fact different from source width in particular. For example, Lee and Rumsey [20] showed that while source width and 'locatedness' were quite highly correlated when looked at over all stimuli in an experiment on microphone crosstalk in multichannel recording, they were not so when examined in stimulus subsets. It seems that subjects 'lock on' to information such as the starting transient when localizing sources, but more to the ongoing information when evaluating their width. Although there is a difference between locatedness and width, wide sources may also be judged to be difficult to localize. This raises the important issue of clarity in definition of terms.

Finally, the concept of naturalness arises regularly in subjects' responses [2]. Although this is in something of a different category to dimensional scales such as width or depth, or to 'spatial experience' scales such as envelopment, it says a lot about the importance of a subject's internal natural listening reference when evaluating spatial audio. Listeners seem sensitive to situations when reproduction could be deemed unnatural. This is probably related to the degree of familiarity with the cues concerned and the extent to which they resemble those in everyday acoustical environments.

7. WHAT ATTRIBUTES MATTER AND WHAT SOUNDS GOOD?

There is work still to be done in this field to determine the contribution of different spatial quality attributes to overall judgements of fidelity, quality and 'liking'. For example, results from concert hall acoustics seem to suggest that people prefer the sound of sources that have been made wider by certain early reflections. However, there also seems to be a tradition among sound recording engineers that tightly-focused phantom images are desirable in stereo sound reproduction. Criticisms are sometimes made of microphone techniques or panning methods that give rise to blurred or broad phantom images [21]. Are we to believe that these opinions actually arise from different phenomena, or that concert hall audiences are different to recording engineers, or that there is some missing factor we are not taking into consideration?

Some evidence for recording engineers being different to the average 'man in the street', at least in terms of their preferences for stereo images, can be gleaned from a recent study [9], in which it was shown that untrained listeners (a large number of people recruited 'off the street') tended to give no significant weight to 'frontal spatial fidelity' in their preference for different versions of a surround sound reproduction of typical programme material. These listeners tended to be most impressed by enveloping, surrounding reproductions, and had little or no concern for changes in phantom source imaging in the frontal arc. The trained listeners (who were students on a course in sound engineering), responded in the opposite sense, giving most weight to frontal spatial fidelity and little to surround spatial fidelity. They might be argued to have been biased by their training to prefer stereophonic images with precisely located front sources, and care less about the 'wow' factor of novel sounds coming from the surround channels. Both groups gave significantly higher weight to timbral fidelity than to spatial fidelity in their overall judgement of sound quality. Despite the differences between the groups in terms of the weight given to frontal or surround fidelity, the proportion of the overall judgement attributable to spatial quality changes in general was roughly the same in both cases, at roughly one third of the total.

Lee [22] found that scenes having phantom images that had been made wider by interchannel crosstalk were not consistently preferred by experienced listeners. It seemed to depend on the type of source material concerned. This suggests that such preferences are highly likely to be context dependent. It is also possible that the lack of a visual stimulus in reproduced sound gives rise to the need for a different quality of spatial localization, in order to be able to identify sources. Whereas wide sources may be pleasant in a concert hall because it is possible to construct a convincing scene by means of the visual mechanism, more precisely focused sources may help to compensate for the lack of the visual element in sound-only presentations.

8. WHAT REMAINS TO BE DONE?

Among the tasks that remain to be fully tackled in the field of spatial audio quality evaluation is that of agreeing about important attributes and the means that can be used to evaluate them. Although language consensus is not vital for work to continue, it would enable experiments to be compared more easily if there was a common set of attributes and definitions thereof. This is complicated by the need to evaluate complex and varied scenes and stimuli, but the consensus already implicit in the results from independent experiments indicates that there is reasonable potential for agreement. The evaluation of changing or moving spatial audio scenes also presents a number of unique challenges, as most work done so far has concentrated on static scenes.

It also seems to be important to determine the level of detail at which it is necessary to evaluate spatial audio quality. For some experiments a simple spatial 'mean opinion score' might be adequate, while for others it might be necessary to decompose this into a more sophisticated set of ratings on a number of attribute scales. Furthermore, if we are to stand a chance of being able to predict factors such as listener preference or 'liking' on the basis of expert ratings of descriptive quality attributes, then a reliable means of accounting for the context dependencies of such matters needs to be devised.

The most challenging task of all is to develop 'objective' measurement models that enable the prediction of perceptual results on the basis of physical measurements. Extant perceptual models such as that standardized in ITU-R BS.1387 (PEAQ) do not take into account spatial degradations in audio quality, for example. It will therefore be necessary to develop reliable measures for the attributes deemed to be important, which may involve the use of sophisticated scene decomposition algorithms if complex programme material is to be evaluated.

9. REFERENCES

1. Zacharov, N. and Koivuniemi, K. (2001) Unravelling the perception of spatial sound reproduction. In *Proceedings of the AES 19th International Conference*, pp. 272–286
2. Berg, J. (2002) *Systematic evaluation of perceived spatial quality in surround sound systems.* PhD thesis. Luleå University of Technology, School of Music at Piteå, Sweden.
3. Mason, R. and Rumsey, F. (1999) An assessment of the spatial performance of virtual home theatre algorithms by subjective and objective methods. Presented at *108th AES Convention*, preprint 5137
4. 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 *108th AES Convention*, preprint 5139
5. Korenaga, Y. and Ando, Y. (1993) A sound-field simulation system and its application to a seat-selection system. *J. Audio Eng. Soc.* **41**, 11, pp. 920–931
6. Toole, F. (1985) Subjective measurements of loudspeaker sound quality and listener performance. *J. Audio Eng. Soc*, **33**, pp. 2–32
7. Gabrielsson, A, and Lyndstrom, B. (1985) Perceived sound quality of high-fidelity loudspeakers. J. Audio Eng. Soc. **33**, 33–53.
8. Zielinski, S.K., Rumsey, F., Kassier, R., and Bech, S. (2005) Comparison of basic audio quality, timbral and spatial fidelity changes caused by limitation of bandwidth and by down-mix algorithms in 5.1 surround audio systems. *J. Audio Eng. Soc.*, **53**, 3, pp. 174–192, March
9. Rumsey, F., Zieliński, S., Kassier R., & 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, June
10. Rumsey, F., Zieliński, S., Kassier R., & Bech.S. (2005) On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *J.*

Acoust. Soc. Amer., **118**, 2, 968–977, August

11. Schroeder, M. (1987) Normal frequency and excitation statistics in rooms: model experiments with electric waves. *J. Audio. Eng. Soc.* **35**, 5, pp. 307–316
12. Savioja, L., Huopaniemi, J., Lokki, T. and Väänänen, R. (1999) Creating interactive virtual acoustic environments. *J. Audio. Eng. Soc.* **47**, 9, pp. 675–705
13. Bech, S. (1999), Methods for subjective evaluation of spatial characteristics of sound. In *Proceedings of the 16th AES International Conference on Spatial Sound Reproduction*, April 10–12
14. Choisel, S. and Wickelmaier, F. (2005) Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound. Presented at *118th AES Convention*. Paper 6369
15. Ford, N., Rumsey, F. and Nind, T. (2005) Communicating listeners' auditory spatial experiences: A method for developing a descriptive language. Presented at *118th AES Convention, Barcelona, 28–31 May*. Paper 6481
16. Usher, J. and Woszczyk, W. (2004) Visualizing Auditory Spatial Imagery of Multi-channel Audio. Presented at *AES 116th Convention*. Paper 6054
17. Rumsey, F. (2002) Spatial quality evaluation for reproduced sound: terminology, meaning and a scene-based paradigm. *J. Audio Eng. Soc.* **50**, 9, pp. 651–666
18. Mason, R., Brookes, T. and Rumsey, F. (2004) Spatial impression: measurement and perception of concert hall acoustics and reproduced sound. Presented at *International Symposium on Room Acoustics: Design and Science, Hyogo, Japan, 11-13 April*. Paper S03
19. Martens, W. (1999) The Impact of Decorrelated Low-Frequency Reproduction on Auditory Spatial Imagery: Are Two Subwoofers Better Than One? In *Proceedings of the 16th AES International Conference*, Paper 16-006
20. Lee, H-K. and Rumsey, F. (2005) Investigations on the effect of interchannel crosstalk in 3 channel microphone technique. Presented at *118th AES Convention, Barcelona, 28–31 May*. Paper 6374
21. Theile, G. (2001) Natural 5.1 music recording based on psychoacoustic principles. In *Proceedings of the 19th AES International Conference.* Paper 1904
22. Lee, H-K. (2006) Effects of interchannel crosstalk in multichannel microphone techniques. *PhD thesis, University of Surrey*