Subjective Assessment of the Spatial Attributes of Reproduced Sound

Francis Rumsey

Institute of Sound Recording, University of Surrey, Guildford, UK

Sound systems delivering enhanced spatial reproduction compared with conventional two-channel stereo demand appropriate subjective testing techniques to assess their performance. Sound quality assessment methods of the past 30 years have been mainly concerned with attributes other than spatial, concentrating primarily on timbral and distortion issues when assessing the qualities of loudspeakers or low bit-rate codecs. Spatial attributes, considered an important contributor to overall reproduced sound quality, can be determined by a variety of experimental and statistical means. They may also be related to physically measurable parameters of the sound field. Early work in this field and examples of applications in experimental projects are described and discussed.

0 INTRODUCTION

As sound recording and reproduction systems become capable of increasingly greater spatial sophistication, and numerous alternative methods are offered to enhance the spatial quality of reproduction, there arises a pressing need to develop more advanced subjective testing techniques to assess the performance of such methods. What constitutes subjective ‘quality’ in spatial reproduction, and what factors govern listener preference for the spatial aspects of reproduced sound? Can a clear link be established between subjective attributes and corresponding objective parameters governing spatial reproduction?

This paper reviews previous work in this relatively sparse field of subjective assessment, drawing on closely related studies in timbre perception, loudspeaker testing and concert hall acoustics, all of which have received considerably more attention to date than the spatial attributes of reproduced sound. Current work related to spatial attributes, being carried out as part of EUREKA Project 1653 (MEDUSA), is described, together with proposals for the adoption of methods from related fields of psychology and product quality testing.

The importance of an informed and careful use of any subjective attributes in listening tests is stressed in this paper, as is the need to distinguish clearly between those attributes that are dictated by the source environment, those that are largely governed by the recording technique or programme material, and those that are influenced by the reproduction system and listening room. Subjective attributes relating to spatial reproduction cannot be used in practice without careful reflection on the nature of the practical problem being studied.

While sound engineers may be interested in psychophysical experiments that aim to measure the limits of human perception, and may have uses for such tests in some cases, there is perhaps a greater interest amongst the audio community in experiments that relate closely to issues of subject preference and sound quality than to absolute thresholds of perception. Letowski [1] made a convenient distinction in this respect, between subject-orientated and object-orientated testing. The former are probably best described as tests to find out more about the perception mechanisms of the human subject, whereas the latter use the perceptual capacities of subjects to rate some attribute of the object under test (a process some have likened to using the subject as a ‘quality meter’).

It is recognised that the whole area of quality assessment is fraught with difficulty, and is possibly less easy to justify than the ‘harder’ forms of science, but that is no reason to dismiss it as irrelevant. As Stephens [2] pointed out: “nothing stops research more effectively than the belief that a kind of measurement is impossible”. There is sufficient available evidence from numerous fields of product quality assessment, often completely unrelated to sound, that reliable, meaningful and repeatable results can be obtained, provided that
appropriate methods are adopted, and care is taken in the analysis and application of results. The application of such techniques to some areas of sound quality assessment is in its infancy compared with fields such as the food and beverage industry, but that should only act as a spur to encourage research that moves the audio industry forwards. Since sound is as much a consumer commodity as any other product these days, there are strong arguments for determining what factors have the greatest effect on consumer preference, and how quality is judged by consumers of sound systems and consumers of recorded material. Spatial aspects of sound quality are probably at least as important as timbral quality or distortion artefacts, but have received much less attention to date than either of these two. Now that many other aspects of sound quality are reaching the limits of human perception it is perhaps time to investigate spatial issues more closely. In 1985, Toole [3] concluded that “assessments of stereophonic spatial and image qualities were closely related to sound-quality ratings” — this in relation to loudspeaker tests, but to date there have been relatively few attempts to isolate any more detailed spatial attributes in reproduced sound than the all-encompassing ‘spaciousness’.

1 SUBJECTIVE EVALUATION OF SPATIAL ATTRIBUTES — OVERVIEW

1.1 Spatial attributes related to quality or preference ratings

An attribute is essentially a characteristic quality of an object that one may use in describing it. ‘Warmth’ or ‘coldness’ are attributes of many tangible objects, for example, and could be rated by subjects on an appropriate scale, or objects could be compared with each other. In food and beverage assessment one might use attributes such as ‘salty’ or ‘sweet’. One of the best known examples of a complex attribute analysis is Nobel’s wine aroma wheel [4], which is used widely in the industry. In spatial audio one will probably be interested in attributes relating to such factors as image sharpness, listener envelopment, source spaciousness and source distance.

For a limited range of objects in a similar category it may be possible to define a collection of common attribute scales that are sufficient to describe the principal differences and similarities between them. One might even be prepared to accept that an equation could be developed whereby subject preference could be related to a weighted sum of certain attribute ratings. While the sceptic might say ten different people would have ten completely different views on the relationship between attribute ratings and overall preference or quality judgments, there is in fact surprising correspondence between subjects in some such cases, particularly when the subjects are experts and have developed appropriate discriminatory abilities. A good example of the correspondence between overall preference and weighted spatial attribute gradings of multichannel sound systems can be found in Nakayama [5].

The use of expert panels is a contentious area, particularly in respect of quality judgment or preference tests. It is likely that expert subjects will be using a different side of the brain to non-experts, leading them to reasoned, analytical conclusions based on their training, whereas non-experts may react in a more unreasoned emotional fashion [6]. Which answer do we want? Since the consumers of audio are largely non-experts there is an argument for using them as subjects, but, as Bech [7] has pointed out, one must use a lot more of them to arrive at a similar confidence interval in the result as their ability to discriminate between stimuli is poorer and their results less repeatable. In any case non-experts become more expert as they become familiar with the methods and stimuli.

There can be a strong correspondence between objectively measurable parameters and subjective preference. To take a concrete example from the field of concert hall acoustics, Ando [8] was able to develop a psychoacoustically optimised Listener Preference Index from his subjective experiments by weighting the gradings of four measurable factors: loudness, reverberation time, delay time to strongest early reflection, and interaural cross-correlation (IACC). It is clearly desirable to be able to relate subjective quality attributes to corresponding objective parameters that can be modified by the equipment designer or the sound engineer, in order that products can be optimised for the market.

Sound quality assessment is often made on very basic scales such as the Mean Opinion Score scale specified in standards like ITU-R BS.1116 [9]. This scale is often used to assess degradation of a signal with respect to an unimpaired reference, and has been used widely for tests on low bit-rate audio codecs. Whilst it is useful as a basic indicator of audio quality, it tells the user little about what is impaired, whether the impairment is spatial, temporal, timbral, or some other factor. While such standards mention the possibility of assessing more detailed attributes such as ‘front image quality’ and ‘impression of surround quality’, few have attempted tests using these and it is not yet clear that these particular spatial attributes are independent or of particular value.
Considerable research has been undertaken in the field of loudspeaker quality assessment, using multiple attribute scales, preference ratings, perceptual distance experiments and advanced statistical tools. These have tended to concentrate on attributes that relate to timbre, more closely than anything else, although occasional use is made of spatial terms and the methods may be adapted for spatial purposes (see below). Considerable work has also been done in the field of concert hall acoustics, using a mixture of attributes, some of which are spatial in nature.

1.2 The need to determine and assess subjective spatial attributes

The need for more advanced subjective assessment of the spatial attributes of reproduced sound is clear when one considers the increasing use of spatially sophisticated audio systems. In addition to conventional mono and stereo sound, a range of alternatives are now possible, encompassing multi-channel surround sound, binaural sound for virtual reality and ‘3D’ or transaural stereo for loudspeaker reproduction based on head-related transfer functions (HRTFs). There are also numerous lossy data reduction systems for audio that use spatial processing compromises in some cases, and which may affect imaging and spaciousness characteristics of signals, particularly at very low bit rates. The applications for such technology are wide-ranging, from multimedia games through conferencing systems and virtual reality to high quality home cinema and sound installations.

Methods of recording and mixing sound for such formats also differ widely, and, while a reasonable degree of experience now exists with two channel stereo, there is little clear understanding of how to optimise recording techniques for the more recent spatial sound systems. Without reliable subjective methods for comparing systems and recording/processing methods from a spatial quality standpoint, there will be no objective information on which to base design and operational decisions.

The majority of subjective experiments that have been conducted on spatial sound systems to date have related to the accuracy with which source localisation can be represented. These experiments are paralleled by theoretical models which aim to determine the ability of systems to reconstruct original source wavefronts or approximate to localisation models of human hearing, e.g. [10], [11]. It is noted that while localisation is clearly an important subjective factor it should not necessarily be regarded as the primary or the only factor of importance in the design of sound systems or recording techniques. Other attributes may be equally or more important. There is no research of which the author is aware, for example, that attempts to determine the relative subjective importance of imaging accuracy and spatial impression in sound reproduction. Anecdotal and some concrete evidence suggests that aspects of spatial impression may be a more important factor determining listener preference for reproduced sound fields, as perhaps demonstrated by the popularity of spaced microphone techniques in two-channel recording.

In general, greater emphasis has tended to be placed on the phantom imaging attributes of two-channel stereo reproduction than it has on the spatial impression, as if the only purpose of stereo reproduction were to recreate source localisation cues as accurately as possible. Indeed this view informs well-known learned papers on the subject [12]. In fact two channel stereo reproduction is only a highly limited compromise solution to the problem of delivering spatial sound to the listener, and many trade-offs may have to be made between imaging accuracy and other spatial attributes if one is to achieve a reasonable impression of spatial quality from only two loudspeakers. Recent papers, e.g. [13], have helped to increase the attention paid to spaciousness by those involved with sound reproduction. It could be argued that the correct reconstruction of all possible source images, including reflections, leads to optimal spatial reproduction but, as Gerzon [14] indicated, an impractically large number of channels would be required.

Although some approaches are more successful than others in recreating ‘natural’ auditory cues, the outputs of most spatial sound systems and recording techniques represent greater or lesser degrees of compromise when compared with the localisation information available in natural listening. One can normally only expect to present an approximation to the soundfields encountered in natural listening using reproduced sound (if indeed that is the aim), and that only for a limited range of listener positions. It is the spatial effects of the compromises inherent in that approximation that are the subject of interest here, as well as the complex issue of how to evaluate the numerous examples of ‘artificial’ sound that have no direct origin or parallel in natural acoustic environments.

2 RELEVANT METHODOLOGIES FOR THE DEFINITION AND APPLICATION OF SPATIAL ATTRIBUTES

2.1 Defining the attributes

The various methods used for defining attribute scales seem to split roughly into three groups: those that aim to arrive at a common set of attributes for grading by all panel members, those that
are based on free choice or individualised scales, and those which use some form of multidimensional analysis based on similarity/difference relationships between stimuli. There are distinct advantages in the first from the experimenter's point of view because common scales enable the results from multiple subjects to be statistically analysed together and some inferences drawn regarding the preferences of the general population. The second group of methods, though, is claimed to have advantages of lack of bias and enables personal reflection on the qualities of the items under test, specifically avoiding subject training. The third has the problem of interpretation and application in practice.

2.2 Semantic approaches resulting in common scales
The method known as Quantitative Descriptive Analysis (QDA) [15] involves the selection of approximately 12–15 panel members from some 25 putative members, based on their discriminatory ability and other factors relating to the product category in question. A descriptive language is then developed under the guidance of a panel leader. The scales thereby developed are then used in grading sessions, and the results analysed using traditional statistical methods such as ANOVA. It is probable that this method could be adopted for spatial audio testing purposes.

The alternatives to a structured definition of attributes by discussion, such as QDA, usually involve approaches such as factor analysis or PCA, used by Gabrielson and others (described below). Using a wide range of terms arrived at through questionnaires or by expert intuition, subjects are asked to grade a range of stimuli against each of these terms. Factor analysis is then used as a form of information reduction process to extract a smaller number of common quality or sensory attributes which can be labeled by examining the factor weightings applied to different terms and deciding how the factor analysis has grouped the information.

2.3 Reflective definition of attributes and the use of personal scales
Annie Kjeldsen recently alerted the audio world to an alternative approach from experimental psychology that might have validity in the field of sound quality assessment [16]. She describes the method known as Repertory Grid Technique which can be used in the analysis of personal preference. She rightly points out the limitations of semantic differential methods based on pre-determined attribute scales, which are that although the expert panel members you use may all understand the same thing by the terms used, the rest of the world may not. "An obvious limitation of this type of measure," she says, "is that you only get an answer to what you ask".

The repertory grid technique, on the other hand, encourages personal reflection upon the qualities of the stimuli under examination, and definition of a personal set of constructs that differentiate between them. Subjects have been shown to be more reliable when using their own language than that of others. The method usually relies on the comparison of triads of stimuli, with subjects each asked to describe ways in which two of the stimuli are alike and different from the third. A new triad is then presented and the same question asked. This continues until the subject stops providing new answers. A grid is then constructed upon which subjects rate each of the stimuli according to each of the constructs elicited in the previous phase. The constructs are created out of opposing pairs of terms, such as 'loud/soft', 'open/closed', etc. It is possible for the experimenter also to introduce terms considered important for the test in hand, although this moves more towards the 'provided constructs' rather than the 'elicited constructs' domain.

Difficulties with this type of approach are that simple forms of statistical analysis are precluded, since subjects may come up with widely differing constructs. What is possible, though, is to examine the ways in which people interpret their experience, degree of complexity resulting from different stimuli categories, range of differentiation between similar stimuli, and so on.

It is nonetheless possible to envisage a hybrid of the two basic approaches described above. One might be able to use techniques such as repertory grid to encourage subjects individually to differentiate clearly between stimuli, and to awaken them to the attributes of the stimuli by forcing them to compare and contrast items in a structured fashion. The constructs elicited by this method could then be pooled across subjects and common constructs or terms extracted. Discussion, along the lines of a QDA panel, could then result in the agreement on common meanings, and the resulting scales used in tests that would then be open to more conventional analysis.

2.4 Multidimensional scaling (MDS)
MDS, unlike these other methods, relies commonly upon ratings of difference or similarity between stimuli. There may be a number of dimensions in the relationships between stimuli revealed by an MDS analysis that could not be uncovered without this statistical analysis. A primary advantage of MDS is that because subjects are making ostensi-
bly simple judgments that are not dependent upon labeled scales, and are not rating identified factors, there is little chance of bias or distortion owing to differences in understanding of semantic meanings [17]. Of course the result is that a number of dimensions are revealed by statistical analysis that then have to be interpreted, giving rise to a similar problem as with other methods as discussed in the next section. Nonetheless, MDS may be capable of revealing 'hidden meaning' in the data which might otherwise have remained hidden.

2.5 Independence of attributes

When assessing sound quality attributes it is desirable that as much useful information as possible is extracted from experiments. This implies the need for attribute scales that are reasonably independent of each other — in other words, attributes should be defined that do not interact, as far as possible. The degree of independence between attribute ratings can be assessed using statistical techniques such as correlation analysis, wherein low correlation between variables is indicative of greater independence. Attributes that are truly independent are said to be orthogonal, and can be represented as orthogonal dimensions on a multi-dimensional scalogram. A simple (fictional) two-dimensional example using spatial terms is shown in Figure 1, where image focus is shown on one dimensional attribute is one which, perhaps obviously, only describes one unique facet of the sound in question in an unambiguous fashion. A good example of the difference between independence and one-dimensionality can be found in a recent experiment on surround sound synthesis algorithms by Rumsey [18]. It was found that the attributes 'front image' and 'spatial impression' were reasonably independent of each other, having low correlation (at least for two channel material), but that the 'spatial impression' attribute was subject to wide differences in grading between listeners. Listeners were nonetheless consistent with themselves and reliable in repeating their gradings.

Based on informal feedback from listeners and personal listening it was hypothesised that this was due to the attribute 'spatial impression' being itself multidimensional, having a number of sub-attributes such as 'envelopment', 'source width', 'image depth', 'stability', 'phasiness', and so on (although the precise nature of these sub-attributes has yet to be determined). Almost certainly the attribute 'spatial impression' was not one dimensional, but it was fairly independent of 'front image quality' (which was graded much more consistently, but which is also suspected to consist of more than one sub-attribute). This is discussed further below.

In many experiments the attribute scales are defined by the experimenter, using his or her knowledge of the subject and intuition concerning the factors of interest. This is arguably valid as an approach, and indeed the experimenter is perhaps the most likely person to be able to define the factors of interest, but the chances of those scales being truly independent is limited. Whether or not it is necessary for attributes to be orthogonal is open to conjecture. While it is mathematically neat for the dimensionality of space perception to be reduced to as few dimensions as possible, it is also important that the scales or dimensions defined are meaningful. The scales proposed for use in loudspeaker testing, such as those suggested in IEC 268-13 [19], are almost certainly not orthogonal, for example, but they are meaningful to the audio engineer.

Using multidimensional scaling (MDS), as described in section 2.4, it is possible to determine a number of dimensions on which stimuli can be plotted, based on judgments of their perceptual similarity. While these dimensions represent the main elements of variance in a similarity matrix and enable one to map stimuli in a 'perceptual space', they do not necessarily lead to the identification of the fundamental orthogonal attributes of the quality under examination because the dimensions arrived at through MDS are open to
considerable interpretation. Usually other information is needed to make sense of the dimensions revealed, and the labels given to the dimensions (if any) will usually be based on the results of other experiments such as semantic differential or other descriptive adjective-based methods. Some experimenters in timbre research, such as Grey [20], avoid the use of verbal descriptors for dimensions almost completely.

It is suggested, therefore, that while orthogonality/independence of attributes is desirable, it is by no means the only issue of importance in the use of attribute scales for the spatial assessment of reproduced sound. While it is possible that there exist a number of fundamental, orthogonal and incontrovertible quality dimensions of spatial sound perception appropriate for use with reproduced sound, it is unlikely that a conclusion will be reached concerning their identity in the near future. The equivalent dimensionality of 'timbre space' has been hotly debated in timbre research for some thirty years or more, without satisfactory conclusion, yet there are numerous researchers around the world using a variety of attribute scales for subjective experiments on sound timbre, each with differing degrees of usefulness and applicability. This is summarised well by Plomp [21], when he points out, for a timbral experiment using nine stimuli:

"in this example, based upon a specific set of stimuli, three factors alone appeared to be sufficient to describe the differences satisfactorily. This number cannot be generalised... It is also possible to select nine stimuli which would require, for example, five dimensions to represent their timbres accurately."

Spatial subjective assessment is at a very early stage in its development compared with timbre, loudspeaker or codec quality impairment tests, and even earlier compared with work in the food and beverage industry. It is therefore likely to be several years, perhaps many tens, before a degree of consensus begins to emerge among those working in the field concerning what attributes are important and what not. It is almost certain to be a case of 'horses for courses', with attributes appropriate to the problem in hand being chosen by a variety of recognised methods.

2.6 Sets of attributes and their relationship to different modes of natural and reproduced sound

It is interesting to consider whether or not the attribute scales appropriate for the assessment of reproduced sound are the same as those that might be used to describe sounds in natural listening, whether they are a subset of the 'natural' set, or whether they have only some common features. Figure 2 shows some possible arrangements of sets of attributes, the correct organisation of which can only be determined by experiments that have only recently begun. The author favours the likelihood that there is a large area of overlap between natural and reproduced sound, but that some attributes may be unique to each.

McKinnie [22], in an experiment conducted as part of the MEDUSA (Eureka 1653) project to determine a set of adjective scales for use in subsequent experiments on reproduced sound, arranged for a number of subjects to experience a wide range of different natural acoustic environments in a structured fashion, asking them to write down non-technical words that described the spatial, three-dimensional nature of sounds they experienced. Upon their return, a structured discussion was held during which the results were shared and meanings of adjectives were agreed upon. Words which appeared similar were discussed as pairs to determine whether or not they meant the same thing. Scales were then constructed out of pairs of antonyms. The subjects listened to examples of 4- and 5-channel surround recordings and were asked to rate each of the recordings and distinctive sounds within each recording using the rating scales defined. Between each item subjects discussed the usefulness of the scales with relation to the task.

While it is not appropriate here to divulge the precise results of this experiment, it is clear that common agreement was reached with reasonable ease on seven scales that should be used. Some of the scales related to attributes of sources themselves and others more to places or environments. This is a crucial distinction as most complex audio reproductions consist both of spatial cues relating to specific sources, often described in terms of 'imaging' attributes, and cues relating to background environments (e.g. reverberation, reflections), often described using 'spaciousness' terms. Only one of the scales derived from the natural sound experiences was found not to be useful in evaluating reproduced surround sound, but all the others were valuable in some way. It was not claimed that the attributes identified were in any way universal or that others might not be needed for certain types of reproduced sound, only that these were adequate for the recorded examples and natural environments presented.

One of the problems of using scientifically well-defined, independent perceptual attributes when assessing sound reproduction is that there may be questions one wishes to ask as a sound engineer that relate to the overall balance or some technical
problem with the reproduction. One may wish to solicit judgments from subjects concerning the front/back balance, or narrowness/width of front sound stage. Technical terms such as 'phasiness' are also useful when experts want to comment on sound reproduction, since they are meaningful, but they almost never arise in natural listening and would probably not arise in a word list experiment involving non-technical language. Again it is a matter of choosing the scales to be used in an informed manner, and ensuring that all subjects understand what they are grading. Independence and one-dimensionality may have to be sacrificed to usefulness and meaning of the results in many practical situations, particularly when one is more interested in assessing products or systems than one is in learning more about the minutiae of human perception.

Spatial attributes appropriate for assessing mono reproduction may be different from those appropriate for two-channel stereo, which may again be different from those appropriate for multichannel surround. There are also reproduction modes such as binaural listening on headphones, and transaural reproduction on loudspeakers, each of which may give rise to the need for some unique attribute scales. As with the large overlap between natural and reproduced sounds suggested in Figure 2, it is probable that there are numerous common attributes useful for describing different modes of reproduced sound, but there may be some that are unique to certain modes. Figure 3 shows a possible way of visualising the sets of attributes, some of which are common to all modes of reproduction, some to only a few, and some only to one.

2.7 Attributes of sources and environments – different kinds of spaciousness?

The rough distinction that may be made between scales describing sources and scales describing environments is probably useful in instructing subjects and interpreting results. It would probably be nonsensical, for example, to ask subjects to

---

**Figure 2** Some possible arrangements of natural and reproduced attribute sets

**Figure 3** A possible arrangement of attribute sets for different modes of reproduced sound
rate all aspects of a complex sound reproduction using the scale 'pinpoint/vague'. Such a scale is relevant to the judgment of individual source images, such as instruments in an orchestra or voices in a radio play, but probably not to the sound of the hall or studio in which the sources are sounding. Nonetheless there are many cases in which the distinction is much less clear, since it is often not possible to separate sources and environments, or foregrounds and backgrounds. The spaciousness associated with a source image may be perceived as part of the source itself, linked directly to it, in which case it may be quite reasonable to describe a source rather than an environment as having some form of 'spaciousness'.

There is an important link here with auditory streaming, as discussed by Bregman [23], and with what Griesinger [12] has termed CSI (Continuous Spatial Impression), ESI (Early Spatial Impression) and BSI (Background Spatial Impression). Griesinger asserts that when a direct sound source is continuous and cannot be split into separate events, the interaction with reflected energy and the interaural fluctuations that result can give rise to a sense of full envelopment or spaciousness that appears to be connected to the sound (CSI). ESI, on the other hand, is related to separable sound events that form a foreground stream, where reflected energy arrives during the sound event and within 50 ms of its end. ESI, it is claimed, is not fully enveloping and is perceived as occupying much the same location as the source itself. It is the spatial impression experience in small rooms. In large spaces, or in other situations where much reflected energy arrives more than 50 ms after the ends of sound events, BSI results. Spatially diffuse reflected energy of this type results in good envelopment, but BSI is not bound to the source that created it. In assessing reproduced sound using typical listening rooms with short reverberation times, the BSI is almost certain to be provided by the recording rather than the room. BSI can probably be assessed subjectively using terms that relate more to environments, whereas CSI and ESI may require hybrid terms that relate to the spaciousness of sources.

There is an interesting parallel here with work in concert hall acoustics that has subsequently been applied to reproduced sound. Two clear types of spatial impression are identified by Morimoto [24], classed as ASW (apparent source width) and LEV (listener envelopment). The former is related to the space that a source image appears to occupy, is clearly linked to the source itself and has the same general location as the source. It is strongly correlated with the physically measurable parameter he defined in earlier papers as DICC (related to Interaural Cross Correlation, or IACC) which is a measure of the degree of correlation in the sound at the two ears of the listener, comparing early and late sound. LEV, on the other hand, is related closely to late lateral sound energy, and corresponds to the envelopment of the environment. It is possible that Griesinger's ESI corresponds to Morimoto's ASW, and that his BSI corresponds closely to Morimoto's LEV. In terms of sound reproduction, Morimoto found that ASW and LEV were largely independent, it being possible to affect ASW by only manipulating sound fed to the front loudspeakers in a surround sound array to vary the measured DICC parameter. ASW seemed to be unaffected by the direction and amplitude of reflections from side and rear loudspeakers, and independent of front/back sound level ratio, provided DICC was kept constant.

What is not clear is the reason for the apparent contradiction arising from the implication that large values of ASW are a good thing in concert hall acoustics and the view often expressed by sound engineers that precise imaging is a good thing. Pinpoint imaging would seem to be the opposite of large ASW. The answer is probably related to the substantial differences in spatial experience between concert hall acoustics and sound reproduction. With conventional two-channel stereo the degree of envelopment is usually limited compared with natural environments, and all the reverberation in the recording is presented in front of the listener. Possibly precise imaging is subjectively more important in such a case as a means of separating foreground sources spatially from background reverberation. The situation may change somewhat as we make more use of multichannel reproduction which has a more 'natural' distribution of spatial information.

2.8 Origin of spatial features in reproduced sound

When conducting experiments to assess various spatial features in reproduced sound it is clearly important to be aware that those spatial features could originate from a number of different stages in the signal chain. One must be clear about whether one is attempting to assess the effects of (a) the source environment; (b) differences in microphone or processing technique used when generating the programme material; (c) the effects of mode of reproduction (e.g. headphones, loudspeakers, binaural, transaural, etc), (d) the effects of differences in transducers and their layout (e.g. different styles and positions of loudspeakers), (e) the influence of the listening room.

It probably goes without saying that one must know something about the nature of the pro-
Programme material one is using to assess a reproduction system or method of signal processing, otherwise it is very difficult to know whether any spatial feature is inherent in the source material or a real artefact of the system under test. The spatial nature of programme material is strongly dependent upon the recording technique used to generate it. There are ways of removing the influence of the programme items themselves from the results, by running tests, say, which simply compare the ratings of attributes of pairs of reproduction conditions, but the issue is important if one intends to make any absolute judgments or single stimulus ratings of spatial quality. One is often interested in the interactions between programme items and mode of reproduction, so it is desirable to include the programme item as a factor in the experimental design.

29 Selection of programme material
One must exercise care in the selection of programme items for listening tests, in order to have material with intentionally selected characteristics, just as one would in any other form of listening test. Items might be chosen for their precise imaging, for example, or because they use a particular recording technique, or some other parameter considered important. There is also the thorny issue of what to do about the vast range of recorded material that is created in the studio, as opposed to 'natural' material that purports to be a record of an acoustic performance. There is no 'norm' or memorised anchor for such artificial material, and it is therefore hard to know how to categorise it in terms of spatial qualities. This is an ongoing issue that will not be solved easily, and some might question whether it needs to be solved, but it is proposed that those spatial attributes that govern listener preference in so-called 'natural' balances may also have relevance when assessing 'artificial' balances. Only experiments will confirm this hypothesis.

Of particular importance in the case when one is asking subjects to grade certain well-defined attributes is the need to ensure that the programme material is not so complex or so varying over time as to make it difficult or impossible for the subject to know what he or she is supposed to be grading. Past moving sources, or scenes that change their nature during the course of the programme extract are likely to be very hard to grade reliably. It is probably most sensible, unless one has a particular reason to the contrary, to choose material that has a reasonably static character, possibly involving only single sound sources, when one is attempting to vary a specific experimental parameter and assess its effect on the subjective result. The problem with doing otherwise is that error variance in the resulting data may be considerable owing to the complexity of the task for the subject, and confusion over the nature of the task. One could end up measuring three or four different factors in one grading, and not really knowing which one affected the result.

It is worth noting in this connection that some experiments designed to identify television picture quality attributes use static pictures rather than moving scenes for precisely the same reason. (This does not apply to low bit rate moving picture codec tests where image movement is one of the main factors giving rise to distortion artefacts.)

3 REVIEW OF SELECTED EXPERIMENTS INVOLVING SPATIAL ATTRIBUTES OF REPRODUCED SOUND

3.1 Localisation experiments
It is proposed to leave to one side the considerable number of projects that primarily involved tests on localisation. That is partly because these are first and foremost perception experiments rather than quality attribute experiments, although it is recognised that there is really no clear dividing line between the two except conceptually or philosophically. Most experiments involving localisation have mainly been concerned with the ability of sound systems to encode and reconstruct source image positions in two or three dimensions, which on the surface is an admirable aim, but which may or may not have a great deal to do with perceived sound quality (see above).

3.2 Subjective assessment of multichannel reproduction
One of the few examples of spatial subjective quality tests carried out during the previous intense period of interest in multichannel surround reproduction is the work of Nakayama et al [5]. He studied the subjective effects of 1–8 channel reproductions in an anechoic chamber using recordings made in a concert hall with unidirectional microphones in the same arrangement as the reproducing loudspeakers. Other microphone arrangements such as an MS pair and a close multimic balance were also used. The microphone array was used at three different distances from the orchestra.

Two different approaches were used in the subsequent subjective assessment, in which 13 different speaker arrangements ranging from 1 to 8 channels were presented. In a single-stimulus experiment listeners made a preference judgment on a seven-point scale, ranging from 'very good' to 'very bad'. In a paired-stimulus experiment listeners were asked to judge the similarity between
stimuli, also on a seven-point scale ranging from 'just the same' to 'quite different'. A distance scale for preference was constructed from the quality judgments and the similarity data were converted to similarity distances between all combinations and subjected to multidimensional analysis (MDA).

In these experiments it appeared that no semantic scales were used to grade particular attributes, but one must conclude that adjectival ratings were used at some point because the authors proceed to interpret three perceptual dimensions revealed by MDA (corresponding to about 77% of the total variance in the similarity matrix) "based on vector analysis of the adjective space of the reproduced sound". It appears that this work was reported in another place. The axes they identify are interpreted as (a) 'depth of image sources', (b) 'fullness', (c) 'clearness'. An examination of their results suggests that 'fullness' is very similar to what others have called 'envelopment', as it is heavily loaded for reproductions involving more loudspeakers to the sides and rear of the listener, and weak for two-channel frontal stereo. It appeared to be greatest in a four channel reproduction when the side loudspeakers were located between about 50 and 60 degrees off front centre (two front speakers at ±15°). 'Depth of sources' seems in fact to be more like 'nearness' or 'closeness' of sources when one reads the authors comments, providing a good example of the difficulties of language and translation in comparing such results with others. It was changed greatly as the recording position of the microphones was moved closer to the orchestra, as one might expect. 'Clearness' was found to relate closely to the measured concert hall acoustics parameter D50 (Definition or Deutlichkeit), and is thus clearly an indication of direct to reverberant ratio.

They also formulated an equation that related the quality ratings of listeners to the three attributes by weighting the factors appropriately, based on a least-squares solution which fitted values from the three scales to the observed quality values. Their equation suggests that 'fullness' ('envelopment') was weighted most strongly in this equation, followed by 'depth of sources', followed by 'clearness', which is most revealing.

The authors' concluding remarks are worth noting with regard to the problem of assessing 'non-natural' recorded material.

"Needless to say, the present study is concerned with the multichannel reproduction of music played only in front of the listeners, and proves to be mainly concerned with extending the ambience effect... In other types of four-channel reproduction the localisations of image sources are not limited to the front. With regard to the subjective effects of these other types of reproduction, many further problems, those mainly belonging to the realm of art, are to be expected. The optimisation of these might require considerably more time to be spent in trial, analysis and study."

3.3 Perceived quality of sound reproducing systems
Some of the most well-known and in-depth investigations into perceived quality of sound reproducing systems were conducted by Gabrielsson and others. For example, Gabrielsson and Sjören [25] conducted a range of experiments aiming, among other things, 'to find out and interpret the meaning of relevant dimensions entering into perceived sound quality'. They conducted tests on headphones, loudspeakers and hearing aids, in mono on loudspeakers and stereo on headphones. Subjects were asked (a) to rate stimuli on a large number of adjective scales that had previously been selected by a group of sound engineers from a longer list; (b) to rate the similarity between pairs of stimuli; (c) to provide free verbal descriptions of a sample of stimuli.

The adjective ratings were analysed using principal components analysis (PCA) in an attempt to isolate a limited number of quality 'factors'. PCA achieves this by looking for correlations between the multiple adjectival ratings and then offering a limited number of principal factors or components which represent the main perceptual dimensions on which the adjectives seem to be most correlated. The factor weightings given to each adjective show how each 'scored' under each perceptual factor (they extracted three factors in the loudspeaker test and five in the headphone test), and this assists in interpreting the meaning of each factor. While the majority of adjective scales related to timbral and other attributes, a number related at least partially to the spatial attributes of reproduction. Terms such as 'distant/near', 'diffuse', 'closed/shut-up', 'airy', 'confined to a point', 'feeling of room', 'blurred', 'open', could all be considered spatial attributes, and scored high weightings on one of the factors which was interpreted by the authors as "a general quality factor emphasizing clearness/distinctness, feeling of space and nearness in the reproduction". In the headphone experiment one can isolate two factors from the five that may represent spatial attributes: the authors report that Factor II was interpreted as 'clearness/distinctness', and received high factor loadings for adjectives such as 'clear', 'pure/clean', 'true-to-nature' and 'feeling of presence', balanced up with strong negative factor loading for 'diffuse'.

AES 15th INTERNATIONAL CONFERENCE
Factor V is characterised as 'feeling of space', showing a strong negative factor loading for the opposite 'closed/shut-up'. Factors II and V were also found to have a modest correlation (0.45) between them. In the hearing aid tests, the factor 'nearness' came out in one test.

The authors also looked for relationships between listeners' ratings of the two terms 'pleasant' and 'natural/true to nature' and the main factor loadings. In relation to the 'feeling of space' factor these terms appear loaded on the 'open/airy' side, with the 'nearness' factor the balance is in favour of 'near' rather than 'distant' (although not conclusively), and with the 'clearness/distinctness' factor the high loadings are towards the 'clear/distinct' side.

These experiments suggest strongly that spatial attributes are at least one of the main factors determining quality ratings in sound reproduction, and that there is a degree of consensus among listeners as to what spatial attributes are preferred.

3.4 Subjective attributes in audio-visual system tests

Woszczyk et al [26] attempted to define some appropriate attributes that might be used in the subjective assessment of the 'total audio-visual experience' of viewers/listeners. Somewhat arbitrarily, it appears, the dimensions of 'action', 'motion', 'mood' and 'space' were chosen to highlight synergetic links between sound and picture. They all have some implied relationship to spatial factors, with 'motion' and 'space' being the most obvious. In a companion paper, Bech et al [27] use these dimensions in the assessment of audio-visual material, asking subjects to rate aspects of 'quality', 'magnitude', 'degree of involvement' and 'audio-visual balance'. A pilot experiment showed that 'space' was the most sensitive dimension to the changes in chosen A/V factors, so only that factor was subsequently used for further tests. The other terms were possibly rather too difficult to interpret clearly for the majority of subjects.

In a later series of experiments, Bech [28, 29] examines aspects of multichannel surround system alignment and layout on spatial attributes. The questions asked of listeners tended to agglomerate multiple attributes into one category: for example the term 'spatial impression' is defined as including width of the spatial impression in the front (front image width?), depth of the spatial impression in the front (front image depth?) and 'degree to which you feel surrounded by sound' (envelopment?). The weighting of these three factors was left up to listeners to decide. In the case of audio-only reproductions the end-points of the 10-point scale for spatial impression were defined as 'a perfect reproduction of the recording space' (10) and 'no correlation between the reproduced space and the recording space' (0). Such a scale is clearly difficult to use, since the listeners have no knowledge (except an imagined one) of the original recording space, and it cannot be used for artificial balances (e.g. pop music) created in the studio where the original recording space acoustics are not really represented in the reproduction. It is nonetheless interesting to note that Bech also assessed the 'overall reproduction quality' in these experiments, and there appears to be strong correlation between the gradings for this and for quality of spatial reproduction.

Zacharov [30] conducted experiments on the effects of loudspeaker directivity in multichannel reproductions with picture, and separated the attributes somewhat more. For rear surround channels he asked listeners to rate (a) the degree to which they felt enveloped by sound (0 = not very surrounded, 10 = very enveloped), and (b) how detailed are the directional effects? (0 = unclear or fuzzy, 10 = very distinct). This latter question is clearly related more to source definition, and the former more to envelopment. They were also asked to rate spatial naturalness of the presentation, comparing the experience to memorised natural listening experiences. In an experiment on frontal reproduction he considered 'coordination of sound and picture', 'how well do you sense acoustic space of changes thereof?', and naturalness.

3.5 Subjective tests on 2–5 channel surround synthesis algorithms

In a series of tests on 2–5 channel surround synthesis algorithms, Rumsey [18] attempted to separate the issue of source imaging from that of spatial impression by asking listeners to grade 'front image quality' and 'quality of spatial impression' of original 2-channel source material and synthesised 5-channel versions of the same material. Only 'natural' programme material was employed, where a real-life acoustic context was implied (i.e. orchestral music, choral music, sport, drama). Tests were arranged in the form of paired ratings, wherein the 2-channel and 5-channel versions of the same item were paired. Differences between the two could then be ascertained, thereby allowing for differences in image quality and spatial impression of the original programme items to be taken into account in the gradings of the 2-channel version of the pair.

In these experiments, as in others, the author was faced with the difficult task of trading off accuracy and independence of attributes with complexity of the task. Subject accuracy is known to get poorer
SUBJECTIVE ASSESSMENT OF THE SPATIAL ATTRIBUTES OF REPRODUCED SOUND

Figure 4 Front image quality difference grades for four surround processing algorithms, plotted by subject

Figure 5 Spatial impression difference grades for the same algorithms, plotted by subject
as the task complexity increases, but simple tests often suffer from amalgamating too many attributes under one heading. The answer is clearly to look for one or two things at a time, even if it may limit the breadth of the study.

In the front image rating listeners were asked primarily to consider the focus and stability of individual sources in the front image. Programme extracts were chosen which had different degrees of source imaging accuracy, by employing material based on recording techniques known to give good imaging (e.g. simple coincident pair), as well as those often considered poorer in this respect (e.g. spaced tree). Based on the work of Bech and Zacharov, noted above, 0–10 scales were used with only the end points defined by descriptive anchors. (Intermediate labels have been shown to affect scale linearity and translatability.) In the spatial impression rating listeners were asked to consider 'envelopment', 'stability' and 'naturalness/appropriateness' of spatial impression. Clearly this is also an example of the agglomeration of too many attributes in one rating, but one only learns by experience.

The results showed considerable agreement between listeners with regard to changes of front image quality caused by different algorithms (compared with the 2-channel original), as shown in Figure 4. They are almost universally in the same direction (downwards), although the magnitude varies owing to different uses of the numerical scale. Intra- and inter-listener consistency are quite good in the case of this question, suggesting that listeners understood clearly what was meant by the image quality question.

Changes in the spatial impression, on the other hand, as shown in Figure 5, showed wide variations between subjects, with some grading the 5-channel version in one direction and others in completely the opposite direction. Subjects commented informally that this was due to their difficulty in separating 'envelopment' from 'naturalness/appropriateness' and 'stability'. They commented that while envelopment was often high in the 5-channel case, naturalness was often low. Some therefore had decided to concentrate on degree of envelopment, and others on naturalness. Intra-listener consistency, though, was quite good, suggesting that at least listeners were consistent with themselves once they had decided how to use the scales.

A correlation analysis on the data from all the subjects for the two grading scales showed that the attributes of 'front image' and 'spatial impression' were almost independent for the two-channel original material (correlation = 0.16). For the synthesised five-channel versions the correlation was slightly higher (0.57), but this is probably because algorithms that affected the front image negatively also tended to affect the spatial impression negatively at the same time.

3.6 Listener training and reliability
In all of the tests conducted by Bech, Zacharov and Rumsey described above, listeners were subjected to considerable training and familiarisation with the programme material, the grading scales and the test method. Pilot experiments were conducted to determine possible problems with the tests, and confirm the method. In all cases it can probably be accepted that the listeners were 'experts', since they were often musicians with an interest in sound reproduction, hi-fi enthusiasts or sound engineers, having been well trained in the test procedure, or they were members of a permanent and regularly trained listening panel. Listeners were pre- and/or post-screened for reliability before including their results in the tests. By ensuring that listeners always repeated the tests at least once it was possible to perform individual ANOVAs on their results to determine their personal error variance (consistency in grading the same item with the same grade on different occasions), and to examine the residual distribution of results for normality.

A good indication of the reliability of listeners and the difficulty of the task can be obtained by looking at the overall error variance in results. It is interesting to compare the error variances obtained in experiments of this type as it helps one to determine the number of subjects and replications needed to ensure a certain confidence interval in future tests. There is surprising similarity in the error variances obtained by the three authors in these experiments, with values for multichannel material varying between 0.8 and 1.2. (All authors used ten-point scales.) This is considerably higher than the values obtained by Bech [7] after prolonged training of subjects for loudspeaker listening tests, in which case he obtained values of 0.5 or less after six experiments. This was the result of considerably more training than took place for the spatial tests reported here, and it is possible that the error variance of spatial tests could also be reduced with suitably improved and extended training. In Rumsey's experiment (see above) the error variance was noticeably lower for two-channel material than it was for five (in the region of 0.6–0.7). This suggests that subjects found it easier to grade such material reliably compared with the multichannel versions.
4 SUMMARY
In the foregoing pages a reasonably comprehensive review has been attempted of the issues relating to subjective assessment of the spatial attributes of reproduced sound. While some work has been undertaken in this field over the last 30 years it is suggested that the industry may benefit from a more detailed study of this area, particularly in order to be able to assess and optimise the performance of advanced multichannel and 3D sound systems more effectively.

REFERENCES
6 deBruyn, B. (1998) Personal communication
22 McKinnie (1998) A spatial audio wordlist experiment. Internal report, University of Surrey