Proceedings of the Institute of Acoustics

PSYCHOACOUSTIC QUALITY EVALUATION IN THE CONTEXT OF INTERACTIVE SOUND AND VIRTUAL REALITY

FJ Rumsey Logophon Ltd, Witney, UK

1 INTRODUCTION

Psychoacoustic evaluations of sound quality have historically been mostly integrative (in the form of a single rating), and have often included a hedonic element (related to liking or pleasantness), whether this was explicit or not. The use of reference stimuli has been widespread, and the concept of sonic impairment is inherent in many listening test standards. This is because the idea of fidelity is ingrained in our understanding of good sound reproduction.

It could be argued that the audio-visual industry has already reached a point where no more technical quality (in terms of things like noise, distortion, frequency response) can be squeezed out of sound recording, processing and reproduction systems. That is not to say that all systems are perfect, but at least we know how to do technical quality well now, and any reduction in the quality of audio systems is a conscious design or delivery decision, based on available resources, bit rates, bandwidth and the like. We can choose to make sound quality poor, if we think we can get away with it, but the best available technical quality is probably as good as anyone could expect it to be, or at least further effort in that direction would yield very little return.

Increased attention has been paid in recent years to evaluating descriptive perceptual attributes of sound quality, and to the relationship between these and liking or preference. This paper, though, will concentrate predominantly on integrative evaluations, in order to limit the scope. As sound reproduction became more spatially sophisticated, questions arose about whether single ratings of quality could integrate spatial and other attributes. Spatial audio researchers have emphasised the importance of localization-related attributes, for example, but these may not always be important determinants of overall quality or liking, particularly in entertainment listening by consumers or inexperienced listeners.

Now that current research is concentrated on interactive sound and virtual/augmented reality, the question of what integrative evaluations are meaningful has to be revisited. This is partly because it is much harder for any concept of a reference or ideal to be established when one is no longer dealing with sound *re*-production *per se*. What is correct, good or best, for example, is a major challenge to define when virtual and interactive media lead one into realms of creative and technical possibility that may have few anchors in a person's prior experience. One could say that in many such contexts there is no right way to do things, just a large number of possible choices, each of which might be equally valid.

In the domain of interactive audio systems then, things like listener attention, task, involvement and the effect of other sensory modalities become critical factors. The literature has started to consider quality of experience, plausibility and presence as alternatives to (or perhaps more important than) quality or fidelity in these applications. This paper explores the evolution of integrative psychoacoustic quality evaluation in the light of these developments.

2 EVOLUTION OF INTEGRATIVE QUALITY JUDGMENTS

2.1 Sound quality scales and the idea of sonic impairment

Methods of sound quality evaluation involving single integrative ratings have dominated the field for many years. One of the reasons that integrative sound quality scales have been widely used in listening tests is that they are ostensibly simple to use. If one can come up with a single number that describes the entire listening experience in terms of its quality then it's convenient. The downside is that it requires listeners to decide what's important and somehow to weigh all the different factors that might influence their rating. If more than one aspect of sound quality is varying in the sounds to be rated then listeners may choose to weigh them differently, and the experimenter may get inconsistent results and/or not end up with much understanding of the reasons for the ratings. Such integrative quality ratings are often aggregated across multiple listeners in the form of mean opinion scores, or MOS, acknowledging that they represent the average of a range of opinions on the quality of stimuli.

Integrative sound quality scales tend to allow either absolute ratings or impairment ratings. Impairment scales are common in international standards (e.g. ITU-R BS.1116[1]) because a lot of listening tests were designed to evaluate audio coding systems that subtly (or not) impair the sound quality. Such scales have tended to conflate the ability of the listener to detect impairments with a hedonic judgement about pleasantness or annoyance. For example, the top grade on the BS. 1116 impairment scale is 'imperceptible' whereas the next one down is 'perceptible but not annoying'. A continuous quality scale is used in ITU-R BS. 1534[2], running from 'excellent' to 'bad', but listeners are provided with a reference stimulus because this MUSHRA (Multiple Stimulus with Hidden Reference and Anchors) test methodology is still essentially a test for impairments in comparison to an unimpaired reference. Where there are large spatial differences between stimuli being compared, a problem can arise in deciding on the most appropriate anchor stimuli. That's because the anchors are supposed to define certain quality levels and be used repeatedly across all tests, as a means of anchoring the use of the scale to some pre-determined quality indicators. Normally the anchor stimuli are just band-limited versions of the reference signal, so they are primarily impaired in the timbral domain. If spatially (or otherwise) impaired signals are compared with timbrally impaired anchors the situation could be likened to one of comparing mouldy apples with rancid fish in a taste test, and asking the user to rate how similar they are.

2.2 Spatial and timbral domains

As sound systems became more spatially sophisticated the evaluation of overall quality became increasingly complicated. Attempts were made to divide quality evaluation into spatial and timbral sub-domains, in a bid to force listeners to distinguish between changes each domain[3]. This was not a new problem, though, as indeed Letowski had identified spatial and timbral attribute hierarchies in his MURAL (MUlti-level auditoRy Assessment Language) [4]. During the QESTRAL project my research group attempted to predict spatial and timbral quality (or fidelity – see below) ratings from various objective metrics, and looked at the relationship between these and mean opinion score (as evaluated by expert listeners) or preference (as evaluated by naïve listeners)[5]. We found that trained listeners regarded the stereo imaging of sound source locations as an important predictor of overall quality, but naïve listeners (representing ordinary consumers, if you like) hardly seemed to notice this aspect of a system's performance when judging whether it sounded good or not. In this case training or experience had led experts to be able to hear a particular feature and also to believe it mattered. In both cases, though, it was found that timbral fidelity contributed much more than spatial fidelity to overall quality or preference.

2.3 The historical primacy of fidelity

The concept of fidelity has been either consciously or unconsciously burned into ideas about sound quality evaluation from an early stage. There has been an assumption that there is a correct way to reproduce a recording, and that a sound system should reproduce an artist's intention or a natural

sound as faithfully as possible. The idea of a reference reproduction is well embedded in the hearts and minds of most audio engineers.

Fidelity relates to faithfulness of reproduction, and in audio engineering it concerns the extent to which technical equipment is capable of accurately capturing, storing and reproducing sounds. This is almost certainly because sound recording started off as a means of relaying sonic events that actually happened—it was a means of capturing 'reality', so to speak. Fidelity in that context should really be a measure of the similarity of what is reproduced to some notional original sound. There has been a tendency, even so, to include value judgments in concepts of fidelity.

Floyd Toole, for example, describes his concept of fidelity in a 1982 paper on listening tests, and states that in addition to rating various aspects of sound quality 'listeners conclude with an overall "fidelity rating" intended to reflect the extent to which the reproduced sound resembles an ideal. With some music and voice the ideal is a recollection of live sound, with other source material the ideal must be what listeners imagine to be the intended sound'[6]. Fidelity is thus defined in relation to a memorized or imagined ideal reference. Toole's fidelity scale (see Fig. 1) is really enabling a hybrid of a value judgment and a faithfulness judgment. It assumes that listeners know what is correct reproduction, and that what is correct is good.

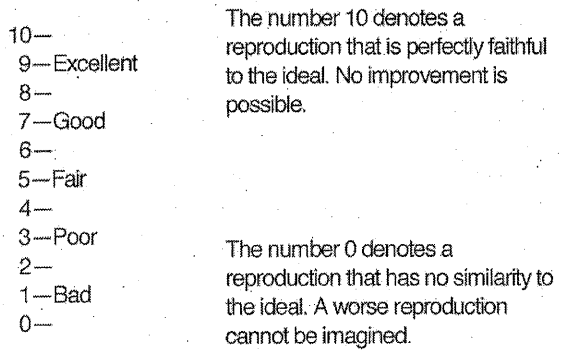

Figure 1 Floyd Toole's fidelity scale

Gabrielsson and Lindstrøm[7], on the other hand, define fidelity as 'the similarity of the reproduction to the original sound…the music sounds exactly as you heard it in the same room where it was originally performed', but acknowledge the difficulty in judging this when listeners do not know what the music sounded like in reality (such as in studio-manufactured pop music).

There are alternative paradigms to this, though. The above idea of fidelity assumes a model of sound quality where there is some ultimate reference against which to compare one's judgment of sound. In the traditional world of recording and reproduction of acoustic signals we usually assume that either the listener has some inherent, learned knowledge about what is correct, or we play them something that we say is correct, against which they will compare various sounds. Interestingly this idea doesn't apply so clearly in some other domains of sensory evaluation such as food and drink, where products are manufactured in order to deliver a particular sensory experience to the consumer. While there may be reference stimuli for particular sensory attributes (e.g. saltiness) in food tasting, designed to calibrate expert tasters' ratings, one rarely finds references (or even the concept of them) for integrative judgements of quality or fidelity. (For an extensive discussion of these and other issues related to sound quality evaluation in different contexts, see the book *Sensory Evaluation of Sound*, edited by Nick Zacharov[8].)

The concept of fidelity becomes almost redundant in the context of some types of sound design, too, because one is dealing with manufactured sounds that never existed in natural life. In such a case one might be more concerned with whether the sound matches or complements some intended application, context or other sensory mode (such as vision or touch). It's also important to bear in mind that some commercial audio-visual experiences are intentionally designed to be 'hyper-real'—

that is exaggerated or enhanced versions of natural experience, for creative impact, improved experience or artistic intention. One can readily question the relevance of fidelity in an age of interactive sonic products that can be altered or created on the fly by consumers who have no notion of a correct version of that product, or where multiple ways of rendering the product are available.

3 FROM QUALITY TO QUALITY OF EXPERIENCE

There has been a general recognition in recent years that the term quality is rather broad and ill-defined, leading various people to attempt to define it more carefully, or to introduce more specific terms. Authors have variously tried to drill down into the definition of quality, looking into quality of service, quality of system, or quality based on experiencing, and this seems particularly important if one is to stand a chance of developing reliable methods for evaluating interactive and immersive audio systems. Possibly one of the most comprehensive taxonomies of sound quality evaluation came from Blauert and Jekosch in 1997[9], where they attempted to define *product* sound quality in the following terms: 'Product-sound quality is a descriptor of the adequacy of the sound attached to a product. It results from judgements upon the totality of auditory characteristics of the said sound—the judgements being performed with reference to the set of those desired features of the product which are apparent to the users in their actual cognitive, actional and emotional situation.' This clearly sets out the need to consider the issues of context and user expectation in relation to what is desired for a particular purpose.

Quality of Experience is a term that gets bandied about widely, and dangerously it can mean different things to different people. There seems to be general agreement, however, that it has something to do with the degree of enjoyment a person experiences when using something. In a Qualinet White Paper on Definitions of Quality of Experience[10] a working definition is given thus: 'Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state.' This would appear to bring various contextual dependencies to bear on the evaluation of quality, and makes explicit the idea that the user's expectations should be acknowledged.

In his paper of 2007 on quality taxonomies for auditory virtual environments Andreas Silzle[11] attempted to determine a set of auditory quality features that contribute to overall QoE, but as Schoeffler et al point out in the paper cited below this does not take into account those factors that are directly related to the user, such as mood, age and experience. Silzle's 'quality elements' are essentially technical in nature, whereas his 'quality features' are essentially auditory attributes.

In a paper on the evaluation of spatial sound systems, Schoeffler, Conrad and Herre attempted to translate the concept of QoE to the problem of evaluating music reproduction[12]. They came up with the term overall listening experience (OLE), which they say 'can be seen as QoE in the context of listening to music', and is used to describe the degree of enjoyment while listening to music. 'When listeners rate the overall listening experience, they are asked to take every factor into account that influences their enjoyment while listening to music. Possible factors of influence might include the song, lyrics, audio quality, listener's mood and the single-/multi-channel system.' While one might initially struggle to see the difference between OLE and the ITU definition of Basic Audio Quality as including 'any and all differences' between the reference and the stimulus in question, the important factor here is that OLE explicitly concerns listener enjoyment.

Tim Walton's 2018 PhD thesis[13] on quality of experience of next generation audio systems discusses the various factors that influence QoE, adapting a diagram from Möller and Raake[14] to show how human, system and context factors all overlap and come to bear (Fig. 2). He points out that there is still a problem with using just a mean opinion score for QoE when evaluating sophisticated modern audio systems, because 'QoE is multi-dimensional and is dependent upon the context and the use'. He argues that mixed methods involving the relationship between individual quality features or attributes and overall quality or preference are more likely to shed light on the reasons for global

assessments. This is strongly supported by numerous other studies, in the audio field and in other disciplines, where the most interesting insights are often those that attempt to explain how low level factors affect higher level constructs such as preference or enjoyment in particular contexts. It's perhaps not enough to know how much something is liked or enjoyed—knowing what makes it so is the useful thing to the system designer.

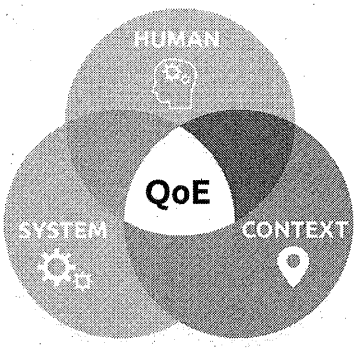

Figure 2 Factors influencing Quality of Experience. (Walton 2018, adapted from Möller and Raake)

It would seem, then, that the field is moving closer to accepting that the roles of user expectations, listening context and emotional response need to be made more explicit when attempting to develop integrative evaluation methods that offer genuine insights into users' experience of emerging interactive and immersive audio systems.

4 FROM FIDELITY TO AUTHENTICITY OR PLAUSIBILITY

The most significant paradigm shift that is taking place at the moment results from a move away from the reproduction of recordings and towards the rendering of interactive experiences. Put another way, until recently a recording would be made of some live event or studio session, the balance and sound of which a producer and an engineer would agree and fix between them. It would then be released as an artistic entity that the creators might hope would be reproduced as accurately as possible. If you didn't like the result there wasn't a great deal you could do about it except not buy it, or buy it and tweak the bass and treble controls a bit. Today, however, there is a shift towards more flexible rendering of media content (e.g. using object-based audio), a wide range of possible listening contexts (mobile, in-car, living room, cinema), the possibility of interaction by the user and others, and numerous options for spatial rendering, ranging from binaural synthesis on headphones to multichannel immersive loudspeaker formats of many channels. This inevitably challenges ideas of a reference reproduction or the possibility that the content creator can fix exactly how something is expected to sound. How something sounds will depend greatly on the sophistication of the rendering system used at the listener's end, and how the listener chooses to set various parameters. Although it could be argued that this has always been the case (a conventional recording could be reproduced on a small radio or a splendid hi-fi system, say), developments such as object-based audio and advanced spatial rendering make the number of listener-end variables considerably greater, including varying the balance and spatial positioning of content elements. In the case of virtual or augmented reality there are only options or pathways through a curated environment or scenario, and the factors contributing to the overall quality of experience are many and various.

In such novel situations many versions of the overall experience could have equal validity. The content rendering system used to deliver the user experience becomes a particularly critical part of the signal chain. It is the quality of this, combined with the sophistication of the content coding system, the means of delivering stimuli to the listener, and the way in which user interfaces alter the sound, which will determine to a large degree the quality of the overall auditory experience. There is increasing interest in using reference-free or 'ideal point' evaluation methods in these contexts.

An interesting example of this new world of sound rendering is an experimental sci-fi drama, The Vostok-K Incident, created by BBC R&D as a pilot or 'taster' to show how flexible spatial rendering could be used to deliver advanced immersive sound to the home[15]. The idea is that listeners can connect various networked devices in the home to act as additional loudspeakers in different locations, in order to enhance the experience of the drama. Careful authoring ensures that the more devices you add the more interesting the experience becomes, and additional content is revealed, but there is no fixed way of experiencing the production. Everyone who hears it will experience it slightly differently.

When thinking about alternatives to the idea of fidelity in such contexts it seems important to distinguish between the media content and any coding, rendering or reproduction systems used to deliver it. Here we will primarily concern ourselves with the audio systems involved, rather than the content. In a recent PhD thesis[16], Chris Pike has discussed alternative approaches to evaluating the perceived quality of binaural technology (such as the loudspeaker virtualization methods used for rendering immersive audio formats over headphones, or direct binaural rendering of audio objects). He refers to work in which alternatives to quality or fidelity are discussed, and which may be more appropriate in modern contexts where many more variables are involved, and where there is no unique reference. These matters are discussed in relation to experiments by various authors that have attempted to determine the degree to which listeners can identify whether something simulated or 'virtual' can be distinguished from something 'real'. 'This relates to the goal of *authenticity*, requiring the simulation to be perceptually identical to reality', says Pike. Often the methods used here are indirect—in other words listeners have either to choose which of a pair of stimuli is virtual or real, or choose the odd one out from a group of alternatives. This is in contrast to asking listeners to rate some attribute on a scale (which can give rise to various response biases).

Later in his thesis Pike is interested in evaluating the performance of non-individual binaural rendering, and decides to use 'plausibility' as the basis for evaluation. He refers to Lindau and Weinzierl's work[17], where instead of authenticity they aimed for the goal of plausibility, defined as '...a simulation in agreement with the listener's expectation towards a corresponding real event'. Pike explains that plausibility can be judged by a 'real or virtual?' style of question, but in this case listeners make comparison with an internal (remembered, imagined) reference for the character of a real event, rather than direct comparison to a real event. Again the method of evaluation is to some extent indirect, in that listeners are not asked to rate the degree of plausibility of stimuli, but it is inferred using signal detection theory from yes/no answers about whether the sound in question was simulated or not. Pike also refers to a study by Bergstrom et al[18], which looks at the idea of plausibility in the context of VR applications. In this case plausibility was defined in terms of 'the illusion that events in the virtual environment are really happening, and is distinguished from place illusion or "being there"'.

It's interesting to compare the definitions of fidelity used in earlier work mentioned above to those of authenticity and plausibility used more recently. From Toole, for example, we have a fidelity rating '...intended to reflect the extent to which the reproduced sound resembles an ideal ...the ideal is a recollection of live sound... what listeners imagine to be the intended sound'. From Pike we have authenticity as 'requiring the simulation to be perceptually identical to reality'. From Lindau and Weinzierl we have plausibility as '...a simulation in agreement with the listener's expectation towards a corresponding real event', and from Bergstrom 'the illusion that events in the virtual environment are really happening'. Either the listener's expectation or ideal is regarded as the standard against which stimuli are compared, and/or 'reality' or some recollection of natural listening is an implied reference. New ideas of authenticity and plausibility are perhaps not so different to old ideas of fidelity. It is not surprising, therefore, that we find naturalness-related terms cropping up time and again in experiments where listeners are asked to describe what they hear when comparing different spatial rendering systems. Whether something sounds real or natural, rather than artificial or simulated, seems to be critical. Listeners have strong sense for 'weirdness' or 'unnaturalness', and this may be related to ideas of the uncanny valley in robotics and computer animation[19]. This clearly has relevance to technology such as that used in augmented or extended reality, where seamless integration between real and artificial audio-visual worlds may be aimed for.

5 ATTENTION, INTERACTION, TASK-DEPENDENCY AND OTHER SENSORY MODALITIES

When considering integrative psychoacoustic evaluation in the context of highly interactive applications such as computer games and virtual reality there are a number of additional challenges to contend with. In such cases the listener may also be a player, participant or doer, the attention may be divided many ways, sound quality issues may not be foremost in one's awareness, and changes to sound rendering may result from direct interaction with a scene. There is also the question of how to evaluate sound quality that is dependent on audio-visual interaction.

Back in 2003, Zielinski et al[20] looked in to the effect of division of attention between evaluation of multichannel audio quality and involvement in a visual task (playing a computer game). It's pointed out that most audio-visual interaction experiments reported in the literature to date had involved the effects of passive watching of visual content on audio quality evaluation. The experiment reported in that paper therefore attempted active involvement of the listener in a visual task, and a computer game was used as the means of controlling attention. The hypothesis was that listeners would be more tolerant towards audio quality impairments when actively involved in a game. For static audio quality impairments it was found that the involvement in such a task could significantly (but relatively little) affect the results for some listeners and conditions. Kassier et al[21] went on to look at the effect of evaluating time-varying audio degradations while involved in a game task, finding similar results. A small but significant effect was noticed across all listeners, with the biggest effect amounting to about 15% higher ratings of impaired audio quality when involved in the game task. (Presumably the game players didn't notice the audio degradations so much because they were otherwise occupied.)

More recently, in a paper describing the evaluation of real-time binaural renderers for virtual reality applications, Olli Rummukainen[22] and his colleagues point out that audio quality evaluation for virtual reality (VR) must acknowledge that there's a visual element, sometimes haptic feedback, and that users probably have a strong mental reference for how natural interaction in VR should look and sound based on life experience. Reference quality stimuli are hard to identify in VR because the true reference is probably the real world. It's pointed out that if we define an explicit audio reference in such a multimodal context then it may bias participants to concentrate on something that turns out not to be particularly important for the overall quality of experience.

Of interest in this study is that participants were asked to evaluate overall audio quality 'with respect to the visuals shown and your own movements' while exploring a number of different virtual scenes. The ends of the scale were labeled 'bad' and 'excellent'. Quality ratings were done on a projected scale that could be displayed or hidden within the virtual environment, allowing listeners to point with a virtual laser at the sound quality scale and record their opinion at any time while exploring the scene. No explicit reference was provided, but an expert group of listeners was told that their answer might depend on how well the reproduced acoustics matched the visual space, the localization, and things like coloration or other artifacts. A naïve listener group didn't get any specific instructions. The experts were both more critical of sound quality and could differentiate between audio renderers. Naïve listeners on the other hand were thought to have been overwhelmed by the multimodal environment and the quality judgment task, showing limited discrimination. The scene content appeared to affect the audio quality scores, with the moving object version resulting in the highest scores for all renderers and both groups. Under those conditions the logged movement of participants was lowest, suggesting that people were devoting all their attention to tracking the sound-producing object, instead of moving around the space or evaluating audio quality. Differentiation between renderers was best in the full indoor scene where participants could walk around and through a virtual loudspeaker and examine it closely.

Another interesting study on sound quality evaluation in a VR context was done recently by Robotham et al[23]. The authors were interested in whether real-time user interactions with a VR scene, accompanied by relevant audio changes, made a difference to quality evaluation of binaural renderers when compared with prerecorded sequences replayed to users. In this experiment they decided to

get the participants to rate QoE along with specific considerations that they were to bear in mind when rating their experiences. These included overall presence, timbral quality, distortions/artifacts, localization quality, consistency of audio with respect to motion cues, and relationship of audio to visual cues. Offline (prerecorded) tests seemed to result in no significant perceived differences being detected between renderers, whereas real-time interaction tests enabled such differences to be revealed. It was suggested that the realistic scenario afforded during real-time interaction revealed differences in renderer performance, because the subjects' senses were all working together, and the changes in binaural rendering resulted directly from subjects' own movements.

The question of audio-visual congruity and its effect on sound quality evaluation is brought to bear on the evaluation of acoustical scenes in a recent paper by Suárez et al[24]. Here the authors attempted to use a VR simulation of various acoustical scenes as a way of trying to ensure that what the listener saw (in terms of likely room acoustics) matched what they heard. They tried to find out whether having congruent scenes in both auditory and visual modalities affected the rating of a number of the 'quality' metrics discussed so far in this paper, namely basic audio quality, plausibility and QoE. They proposed that a higher degree of plausibility in the congruent scenes would make the evaluation task more natural and efficient, and consequently improve the QoE. Unlike some previous experiments designed to measure plausibility, in this case it was rated directly by listeners on a scale presented in the VR environment. Although some small trends were observed towards higher plausibility in the case of congruent audio-visual scenes, none of the results here turned out to be statistically significant, and the authors concluded that the biggest problem was that none of the listeners had really experienced a VR environment before, and were thought to have been cognitively overloaded by the experience and difficulties with the user interface.

Behavioural approaches may yield more useful insights in some cases than asking people questions or getting them to rate things, as they can concentrate on the task in hand without having to think directly about evaluating audio quality. In an earlier pilot study by Rummukainen et al[25] on audio quality evaluation (of binaural renderers) in virtual reality, some data was gathered about user position and orientation in addition to audio quality ratings, in order to find out whether audio quality affected their movement in the virtual space. It was found, however, that this data didn't reveal anything new about the quality of the renderers tested, and the authors commented that behavioural data might only be useful if the users had a task to complete in VR and their performance depended on the quality of sensory information provided by the system.

Constructing meaningful experiments to evaluate an aspect of sound quality in highly interactive applications is clearly extremely difficult. It runs the risk of experiments being contrived, too difficult, or producing irrelevant results. For example, trying to get subjects to rate sound quality while involved in a virtual reality task requires them to stop thinking about the task in hand and attend to the sound rating, then return to the task. The degree of cognitive load involved in such experiments can be quite high. It is possible, therefore, that indirect methods involving observation of user behavior or task-related outcomes could be more relevant here, but only if those are likely to be affected by some aspect of audio quality. One could imagine, for example, that a critical training task involving fighter pilots locating and dodging incoming missiles on a VR simulator might be strongly dependent on the quality of binaural rendering, whereas performance in immersive entertainment tasks might be much less affected by audio quality. Psychoacousticians might therefore consider taking some lessons from behavioural psychology and observe what people do rather than what they say in response to specific questions. Alternatively some forms of brain scanning or EEG may provide more direct access to people's neurological responses than is possible by asking them questions.

6 IMMERSION AND PRESENCE

One of the primary drivers for new methods of quality evaluation in emerging audio systems, apart from interactivity, is auditory immersion. Not only can a user alter their auditory experience by interacting with a system, but they are also often immersed in a sound field that occupies the full sphere around them. Coupled with a strong three-dimensional visual stimulus in VR, for example, the

entire sensory experience offered to the listener can have what some have called 'presence'. A high quality system could be said to deliver a high degree of presence, but the experience of presence is almost certainly as much to do with content as it is to do with the systems used to deliver it. As with some other aspects of the quality evaluation challenge, discussed so far, we are again considering not just what makes something sound good but what makes it seem real.

In a recent paper by Eaton and Lee[26], this terminology is reviewed in the context of virtual reality audio evaluation. They distinguish between passive and active experiences of immersion, the passive version having to do with the experience of being in a space, and the active version having more to do with immersion in a task, or cognitive absorption. (It is possible to talk of 'being immersed in what you're doing' – active – as opposed to 'being immersed in reverberation' – passive, for example.) In other studies the passive version has been termed 'perceptual immersion', and the active version 'psychological immersion'. Eaton and Lee go on to deliver the results of a preliminary survey on factors thought by audio professionals and consumers to contribute to immersion. This was at a relatively early stage at the time of writing, and most of the factors seemed to be concerned with existing perceptual auditory factors such as envelopment and localization.

Both presence and immersion seem to be widely described in the literature as having to do with 'the sense of being there'. It could be argued that a sensation of immersion is a pre-requisite for experiencing the higher level construct of presence. It's interesting to return to Chris Pike's reference to the Bergstrom study mentioned earlier, where plausibility is discussed in relation to presence. Pike said that 'plausibility is described as the illusion that events in the virtual environment are really happening, and is distinguished from place illusion or "being there". Both are said to be components of presence.' (For a useful review of presence in immersive VR see *The Sound of Being There* by Rolf Nordahl & Niels Nilsson[27].)

7 CONCLUSION

There has been a gradual shift in emphasis in the recent past away from sound reproduction of fixed recordings and passive listening, towards flexible rendering of audio content and active immersion, interaction or engagement. This shift is only a trend, though, and there is still a role for the former even if the latter is occupying a lot of research attention and consumer interest. The result of this trend is that existing integrative approaches to audio quality evaluation have needed to be reviewed and revised, so as to enable meaningful questions to be asked, or observations made, and answers provided that are relevant to immersive and interactive applications.

It has been argued that the concept of fidelity to some real or remembered ideal has been inherent in definitions of integrative sound quality evaluation for many years. This is only possible if such an ideal can be clearly exemplified or defined by an experimenter, or conceptualized clearly by a listener. It has also been suggested that defining or conceiving of such an ideal reference becomes increasingly hard in immersive and interactive applications where many outcomes could have equal validity. In such contexts experimenters are trying alternatives that either emphasise the purely hedonic response to an experience (degree of delight, annoyance, enjoyment), or that attempt to measure the degree to which an experience seems to be plausible or real. It is argued that the latter still has an important anchor in listeners' internal standards for what sounds natural or 'true to life', but the reference here is inevitably an internal one rather than one that can easily be provided as a stimulus.

Measures of audio system quality that relate to immersion and presence are related to ideas about plausibility and realness, but there is a danger of mixing up perceptual versions of these concepts with psychological or cognitive ones. The perceptual believability delivered by immersive and interactive audio systems, sometimes termed place illusion or being there, is perhaps only a necessary pre-requisite for a form of cognitive presence and engagement with an experience. The former may be evaluated in terms of auditory attributes and quality judgements, but the latter is related to mental state and behaviour.

Behavioural, observational and neurological approaches have yet to come fully into their own in audio quality evaluation. While there are examples of experiments that attempt to infer things about audio quality indirectly, from data about what people do or how they respond physically for example, these are not yet widespread. Where the application context is purely an entertainment one it seems unlikely that such approaches would have much to offer, but where an aspect of sound quality is critical to user performance in work-related tasks they seem likely to offer valuable insights.

8 REFERENCES

1. ITU-R Recommendation BS.1116-1, 1994. Methods for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems. International Telecommunications Union, Geneva.
2. ITU-R Recommendation BS.1534-1, 2003. Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems. International Telecommunications Union, Geneva.
3. F. Rumsey, S. Zieliński, R. Kassier & S. Bech. On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. J. Acoust. Soc. Amer. 118(2) 968-76. (Aug 2005)
4. T. Letowski, Sound quality assessment: concepts and criteria. Presented at the 87th AES Convention. Preprint 2825. New York (1989)
5. F. Rumsey, S. Zieliński, R. Kassier & S. Bech. Relationships between experienced listener ratings of multichannel audio quality and naive listener preferences. J. Acoust. Soc. Amer. 117(6) 3832-40. (June 2005)
6. F. Toole, 1982. Listening tests: turning opinion into fact. J. Audio. Eng. Soc. 30 (6) 431–445. (June 1982)
7. A. Gabrielsson & B. Lindström. Perceived sound quality of high fidelity loudspeakers. J. Audio. Eng. Soc. 33 (1) 33–53. (Jan/Feb 1985)
8. N. Zacharov, (ed.) Sensory Evalution of Sound. CRC Press, Boca Raton. (2019)
9. J. Blauert and U. Jekosch. Sound-Quality Evaluation – A Multi-Layered Problem. Acta Acustica united with Acustica. 83 (5) 747–753. (1997)
10. Qualinet White Paper on Definitions of Quality of Experience. Output from the fifth Qualinet meeting, Novi Sad, Version 1.2 Novi Sad (March 2013)
11. A. Silzle. Quality Taxonomies for Auditory Virtual Environments. Presented at 122nd AES Convention. Paper 6993. Vienna, Austria (2007)
12. M. Schoeffler, S. Conrad and J. Herre. The influence of the single/multi- channel-system on the overall listening experience. Presented at the AES 55th Conference on Spatial Audio. Paper 5-4. Helsinki, Finland (2014)
13. T. Walton. The Quality of Experience of Next Generation Audio: Exploring System, Context and Human Influence Factors. PhD thesis. Newcastle University. (2018)
14. S. Möller and A. Raake (eds.) Quality of experience: Advanced concepts, applications and methods. Springer. (2014)
15. BBC Research and Development. The Vostok-K Incident. https://www.bbc.co.uk/taster/pilots/vostok and https://vostok.pilots.bbcconnectedstudio.co.uk) (2018)
16. C. Pike. Evaluating the Perceived Quality of Binaural Technology. PhD Thesis. University of York. (2019)
17. A. Lindau and S. Weinzierl. Assessing the Plausibility of Virtual Acoustic Environments. Acta Acustica united with Acustica 98 (5) 804–810. (2012)
18. I. Bergstrom et al. The Plausibility of a String Quartet Performance in Virtual Reality. IEEE Trans. Visualization and Computer Graphics 23 (4) 1352–1359. (2017)
19. M. Mori. The uncanny valley. IEEE Robotics and Automation. 19 (2) 98–100. (2012)
20. S. Zieliński, F. Rumsey, S. Bech, B. de Bruyn and R. Kassier. Computer Games and Multichannel Audio Quality – The Effect of Division of Attention between Auditory and Visual Modalities. Presented at the 24th AES International Conference: Multichannel Audio, The New Reality. Paper 8. Banff, Canada (2003)

21. R. Kassier, S. Zieliński and F. Rumsey. Computer Games and Multichannel Audio Quality Part 2 – Evaluation of Time-Variant Audio Degradations Under Divided and Undivided Attention. Presented at the AES 115th Convention. Paper 5856. New York (2003)
22. O. Rummukainen et al. Influence of visual content on the perceived audio quality in virtual reality. Presented at the 145th Convention. Paper 10128. New York (2018)
23. T. Robotham et al. Online vs. Offline Multiple Stimulus Audio Quality Evaluation for Virtual Reality. Presented at the 145th Convention. Paper 10131. New York (2018)
24. A. Suárez, N. Kaplanis, S. Serafin and S. Bech. In-Virtualis: A Study on the Impact of Congruent Virtual Reality Environments in Perceptual Audio Evaluation of Loudspeakers . Presented at the AES International Conference on Immersive and Interactive Audio. Paper 67. York (2019)
25. O. Rummukainen et al. Audio Quality Evaluation in Virtual Reality: Multiple Stimulus Ranking with Behavior Tracking. Presented at the AES International Conference on Audio for Virtual and Augmented Reality. Paper P2-4. Redmond (2018)
26. C. Eaton and H. Lee. Quantifying Factors of Auditory Immersion in Virtual Reality. Presented at the AES International Conference on Immersive and Interactive Audio. Paper 103. York (2019)
27. R. Nordahl and N. Nilsson. The Sound of Being There: Presence and Interactive Audio in Immersive Virtual Reality. In The Oxford Handbook of Interactive Audio. OUP (2014)