Spatial audio quality perception (part 1): impact of commonly encountered processes

ROBERT CONETTA[1,2], **TIM BROOKES**[1], *AES Member*, **FRANCIS RUMSEY**[1,3], *AES Fellow*
robertc@sandybrown.com t.brookes@surrey.ac.uk fjr@aes.org

SŁAWOMIR ZIELIŃSKI[1,4], **MARTIN DEWHIRST**[1], *AES Associate Member*, **PHILIP JACKSON**[1], *AES Associate Member*
slawek.Zieliński@live.co.uk martin.dewhirst@surrey.ac.uk P.Jackson@surrey.ac.uk

SØREN BECH[5], *AES Fellow,* **DAVID MEARES**[6], **SUNISH GEORGE**[1,7], *AES Associate Member*
sbe@bang-olufsen.dk sunish.george@iis.fraunhofer.de

[1]*University of Surrey, Guildford, UK*
[2]*now at Sandy Brown Associates LLP, UK*
[3]*now at Logophon Ltd. - Oxfordshire, UK*
[4]*now at the Technical Schools, Suwałki, Poland*
[5]*Bang & Olufsen a/s, 7600 Strüer, Denmark,*
[6]*DJM Consultancy, West Sussex, UK, on behalf of BBC Research, UK*
[7]*now at Harman Becker Automotive Systems GmbH, Germany*

Spatial audio processes (SAPs) commonly encountered in consumer audio reproduction systems are known to generate a range of impairments to spatial quality. Two listening tests (involving two listening positions, six 5-channel audio recordings and 48 SAPs) indicate that the degree of quality degradation is determined largely by the nature of the SAP but that the effect of a particular SAP can depend on programme material and on listening position. Combining off-centre listening with another SAP can reduce spatial quality significantly compared to auditioning that SAP centrally. These findings, and the associated listening test data, can guide the development of an artificial-listener-based spatial audio quality evaluation system.

0 INTRODUCTION

A desire exists to create or reproduce increasingly real and immersive soundfields or listening experiences [1][2][3][4][5]. This can be observed in the functionality of current consumer products (e.g. surround sound 'home-cinema' systems, DVD video and audio appliances, gaming consoles). Mobile devices such as MP3 players, mobile phones and tablet computers are becoming increasingly popular and have the potential to deliver binaurally enhanced spatially immersive environments via headphones [6][7]. Furthermore, broadcasters can now deliver spatially enhanced multi-channel audio scenes in the form of matrixed 5.1 surround sound via high definition (HD) television broadcasts [8][9].

Multi-channel audio codecs are often used to reduce bandwidth requirements but they can have detrimental effects on perceived spatial audio quality [10]; this is particularly apparent under the most band-limited delivery conditions (e.g. online streaming) and where storage space is limited (e.g. mobile phone MP3 players). The delivery format of audio programme material is often different from the rendering (reproduction) format: audio is delivered in a format that suits the transmission technology (e.g. HD broadcast, DVD) and can be reformatted for replay over any of a number of reproduction systems (e.g. 2-channel stereo, 5.1); the upmixing and downmixing techniques used for such reformatting can further degrade quality [11][12][13], as can changes made by the consumer to intended loudspeaker positions. Degradations could include changes to source-related attributes such as perceived location, width, distance and stability, and changes to environment-related attributes such as envelopment and spaciousness [6].

The desires, technologies and consequences outlined above motivate the development of an efficient and effective method for assessing perceived spatial quality, for research, for product development and for quality control. The costs (in terms of time and money) of maintaining a listening panel, and assessing audio quality by formal listening tests, can be prohibitive [14]. A computer model of quality perception could act as an

'artificial listener'. An artificial-listener-based perceptual evaluation system, while perhaps not completely replacing assessment by human listeners, could however provide an indication of likely perceived audio quality where human assessment would be impractical or impossible.

Current standard algorithms for evaluating perceived sound quality (e.g. PEAQ [15]) focus on impairments to timbral quality such as audio coding distortions, noise and bandwidth reductions, and do not account for the contribution of spatial attributes[1]. Since the development of PEAQ, Choi *et al* [17], George [18] and Seo *et al* [19] have created spatially-aware sound quality models but these only consider the degradations resulting from a limited selection of spatial audio processes (SAPs).

The QESTRAL (Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener) project aimed to develop an artificial-listener-based evaluation system capable of predicting, for real or virtual multi-channel loudspeaker reproduction, the perceived spatial quality degradations resulting from a wider range of SAPs. Metrics and extraction algorithms for a number of spatially-relevant audio features (informed by the body of research in binaural auditory modelling that aims to predict the perception of specific spatial attributes) have already been developed [20][21][22]. The experiments reported in the current paper aim to determine, by way of two listening tests, the degree of perceived overall spatial quality degradation resulting from SAPs commonly encountered in consumer audio reproduction systems, and to determine the influences of listening position and source material on that degradation. The intention is: (i) to build a quality-annotated database of processed and unprocessed programme items; and (ii) to gain qualitative insights into the effects of SAPs on quality. In a follow-up paper these findings and the quality-annotated database will be combined with the previously-developed metrics to build a regression model of perceived spatial audio quality.

0.1 Spatial Quality

Spatial audio quality is a global attribute comprising a number of lower level attributes [23]. Past studies by Berg [24], Berg & Rumsey [25], Choisel & Wickelmaier [26], Koivunmiemi & Zacharov [27], Rumsey [6], Rumsey *et al* [28] and Zacharov & Koivunmiemi [29, 30] have identified a number of these lower level attributes (e.g. source location, width, depth, envelopment). However, in order to avoid exclusion of potentially-important factors, the current study is not limited to specific previously-identified attributes but, instead, defines spatial quality as the global attribute encompassing any and all perceived spatial differences between a reference recording and a processed version.

[1] An adaptation to enable PEAQ to evaluate degradations to spatial quality is under consideration [16]

1 DESIGN OF LISTENING TESTS

Two listening tests were conducted to achieve the aims stated above. In each test listeners were required to rate the perceived spatial quality of each of a number of test stimuli, as compared to a reference stimulus. Each test stimulus was a SAP-degraded version of the reference stimulus against which it was compared. For each test stimulus, the average of all its quality ratings was sought for the quality database. The following sections explain the reasons for using two tests and two listening positions, detail the test apparatus, programme items and SAPs, and describe the loudness equalisation applied and the test method employed.

1.1 Use of two tests & two listening positions

It is known from previous studies that off-axis listening leads to image skew [31] and that this skew has a negative impact on overall quality [6]. There have been various attempts to widen the acceptable listening area [32][33] but no previous studies have quantified the impact of off-centre listening on overall spatial audio quality. The QESTRAL system was intended to be able to evaluate spatial audio quality at both on- and off-centre listening positions and so the effects of listening position were investigated. They were considered in two complementary ways, using two listening tests, with the choice of off-centre listening position informed by the previous studies cited above and the likely seating positions in a typical domestic listening room.

In listening test 1, centrally-auditioned SAPs were compared to a centrally-auditioned reference and off-centre-auditioned SAPs were compared, separately, to an off-centre-auditioned reference. Thus, alternative listening positions were treated as alternative test conditions under which to evaluate the effects of a wide range of SAPs. This allowed determination of the extent to which the deleterious effects of SAPs might depend on listening position (e.g. one SAP might degrade a centrally-auditioned signal significantly but for an off-centre listener the same SAP might leave the reference signal quality relatively unimpaired).

In listening test 2, centrally-auditioned SAPs and off-centre-auditioned SAPs were both compared to a centrally-auditioned reference. Thus, off-centre listening was, in effect, treated as an additional SAP combined with the SAP under test. This allowed examination of the resulting compound quality degradation (e.g. moving off-centre might significantly degrade the perceived spatial quality of one particular SAP, but might make little difference to the quality of another SAP).

1.2 Listening test apparatus

The listening tests were conducted at the University of Surrey's Institute of Sound Recording (IoSR) in a listening room compliant with ITU-R BS.1116-1 [34] requirements. Bang and Olufsen Beolab 3 loudspeakers (Frequency response: 50Hz to 20kHz [35]) were used and were concealed from the listener by an acoustically transparent but visually opaque curtain. The high-quality listening room and loudspeakers were chosen in order

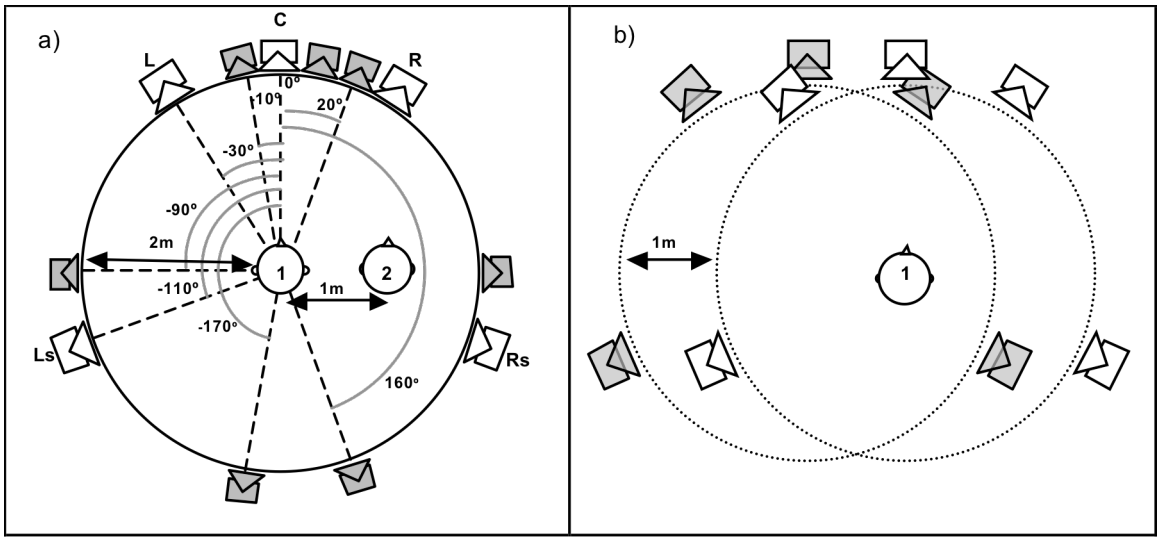


Figure 1. (a) Listening and loudspeaker positions for listening test 1: core 3/2 array (white) labelled L, C, R, Ls and Rs. (b) Listening and loudspeaker positions for listening test 2: core 3/2 array (white); off-centre array (grey).

that the reproduction system should be as transparent as possible, so that the most significant degradations to the programme material would be due to the SAPs under test. A room with a poorer acoustic, or lower-quality loudspeakers, would constitute an additional SAP. This could be considered in a future investigation and the effects on quality incorporated into a future version of the QESTRAL system.

For listening test 1 the core playback system comprised 5 loudspeakers arranged in 3/2 stereo configuration according to the requirements described in ITU-R BS.775-1 [11]; additional loudspeakers were employed for SAPs that required them (Figure 1). Listening test 2 employed two 5-channel loudspeaker systems, one as a reference system with a central listening position (LP1) and one to provide an off-centre listening position (LP2) for comparison. Prior to each test all channel gains were calibrated individually to produce the same sound pressure level, at the centre of the corresponding loudspeaker system, using a pink noise test signal.

1.3 Programme material

SAPs were applied to six 5-channel audio recordings (Table I). These programme items were chosen to span a representative range of ecologically valid audio recordings, likely to be listened to by typical audiences of consumer multi-channel audio, while also covering typical genres and spatial audio scene types. For example, the content of programme item 1 (TV/sport) is mixed to represent a scene suitable for a television sports broadcast with multi-channel audio. There are two commentators panned slightly left and right of the front centre position where the television set would likely be. Audience applause and ambience can be heard from 360º around

Table I. Programme items used in listening tests 1 (items 1–3) and 2 (items 4–6).

No.	Genre Type	Scene Type	Description
1	TV Sport	F-F	Excerpt from Wimbledon (BBC catalogue). Commentators and applause. Commentators panned mid-way between L and C, and C and R. Audience applause covers 360º.
2	Classical Music	F-B	Excerpt from Johann Sebastian Bach – Concerto No.4 G-Major. Wide spatially-continuous front stage including localisable instrument groups. Ambient surrounds with reverb from front stage.
3	Rock/Pop Music	F-F	Excerpt from Sheila Nicholls – Faith. Wide spatially-continuous front stage, including guitars, bass and drums. Main vocal in C. Harmony vocals, guitars and drum cymbals in Ls and Rs.
4	Jazz/Pop Music	F-B	Excerpt from Max Neissendorfer & Barbara Mayr – I've Got My Love To Keep Me Warm. Live music performance. Wide front stage. Ambience from room and/or audience in rear loudspeakers.
5	Dance Music	F-F	Excerpt from Jean Michel Jarre – Chronology 6. Very immersive. Sources positioned all around the listener. Some sources are moving.
6	Film	F-B	Excerpt from Jurassic Park 2 – The Lost World. Dialogue in C. Ambience, sound effects and music in L, R, Ls, and Rs.

the listening position. This recording represents a typical F-F (foreground-foreground)[2] scene type where each audio source is either close or clearly perceivable [36]. In comparison, programme item 2 (classical music) is a classical recording with a different mix style, typical of many recordings from this genre, where the front three loudspeakers (i.e. left, centre and right) contain a wide spatially-continuous mix of the orchestra while the rear or surround loudspeakers contain ambient or reverberant energy. This recording represents a typical F-B (foreground-background)[3] scene type.

Table II SAP groups used in listening tests 1 and 2

Group	Process type
1	Down-mixing from 5 channels
2	Multi-channel audio coding
3	Altered loudspeaker locations
4	Channel rearrangements
5	Inter-channel level misalignment
6	Inter-channel out-of-phase errors
7	Channel removal
8	Spectral filtering
9	Inter-channel crosstalk
10	Virtual surround algorithms
11	Combinations of group 1–10 SAPs
12	Anchor recordings

1.4 Spatial audio processes evaluated

Forty-eight different SAPs were chosen for evaluation, to create a large number of stimuli, exhibiting a wide range of typical impairments to spatial quality. The selection was informed by discussions amongst the QESTRAL project group, by previous related studies [12, 37, 38], and by the results of specific pilot studies [22]. The chosen SAPs can be divided into 12 groups (Table II). Table XI in the Appendix gives full descriptions.

It is possible that some SAPs may enhance, rather than degrade, spatial quality but informal pilot evaluations by the authors indicated that, for the selections employed in this study, processed stimuli were never of a higher quality than the corresponding unprocessed reference. Within this study, therefore, the unprocessed reference stimuli were considered to be of optimal quality. If the results from the formal tests include processed stimuli rated at 100 % quality then this will be revisited.

[2] F-F (Foreground-Foreground) denotes Foreground programme material (e.g. speech, musical sources) in the front loudspeakers and Foreground material in the rear.

[3] F-B (Foreground-Background) denotes Foreground material in the front loudspeakers and Background material (e.g. reverberation, applause) in the rear.

1.5 Stimulus loudness equalisation & playback

The stimuli (SAP and programme item combinations) were loudness equalised using a listening panel. Each listener was asked to adjust playback gain to make all unprocessed reference stimuli equally loud, and then to make each processed stimulus equally loud to the corresponding original unprocessed reference. The means of the resulting gain adjustments were applied to the experiment stimuli. Overall playback gain was kept constant across all trials, having first been adjusted to provide a comfortable listening level. Thus, all stimuli were equally loud and measured 75–80 dB $L_{AEQ(1\text{-}3mins)}$.

1.6 Listening test method

Pilot studies investigating the magnitude of perceptual differences between stimuli led to the choice of a *multi-stimulus with hidden reference and anchors* (MUSHRA) test method [39]. Listeners were presented with 8 stimuli at a time and instructed to rate the spatial quality of each stimulus compared to an unprocessed reference programme item. Listeners listened to the stimuli and recorded their responses using a graphical user interface (GUI) designed to reduce assessment scale biases inherent in listening tests (Figure 2) [40][22]. The GUI was presented on a laptop situated at the listening position. The full instructions given to each listener, including a definition of spatial quality, are provided in the Appendix. It is acknowledged that listeners, although instructed to consider only spatial attributes, might also have considered timbral and other attributes. It will be important to take this possibility into consideration when the collected data are used to build a spatial quality model.

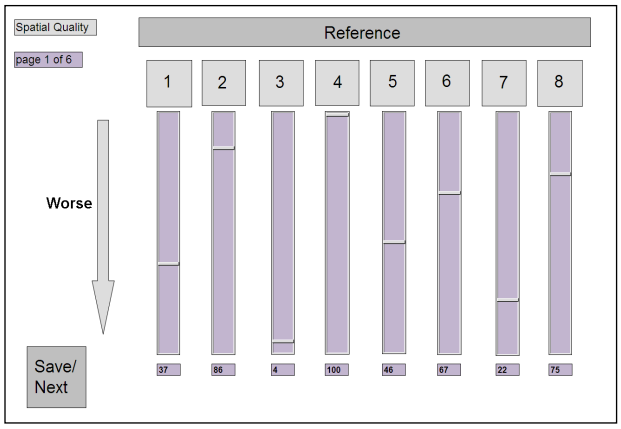


Fig. 2 Graphical user interface used in listening tests 1 and 2

Quality ratings were recorded as integers from 0 to 100. These are reported in later sections of this paper as percentages but it should be noted that they can only be considered as such within the context of the chosen scale end-points: the lowest anchor and the unprocessed reference. If a stimulus has a quality rating of 0 % then this indicates that no other version of that programme item presented in the experiment was perceived as having a lower quality; it does not indicate that quality could not

possibly be lowered further. Similarly, if a stimulus has a quality rating of 100 % then this indicates that no other version of that programme item presented in the experiment was perceived as having a higher quality; it does not indicate that quality could not be possibly be improved.

A full factorial experimental method was used so the listeners assessed every stimulus in every condition over 4 sessions, at each listening position. The presentation order of the stimuli within each session was randomised. Each session consisted of the test and a repeat of the test, and lasted approximately 30 minutes. Before commencing each session, listeners completed a familiarisation trial to enable them to hear, and practise the assessment of, each stimulus. Fourteen listeners from the IoSR (Tonmeister and post-graduate students) with training in technical/critical listening and prior experience as listening test subjects, took part in listening test 1 and seventeen took part in listening test 2. Due to the exploratory nature of the experiment, listeners were not specifically trained for it: it was important that they should interpret and rate the spatial quality of what they heard freely [25].

In accordance with the MUSHRA test method, 3 hidden indirect audio anchors, chosen to lie at the top, middle and bottom of the test scale, were employed. These anchors were included on every test page in order to encourage listeners (without their knowledge) to use the rating scale more consistently from page to page and from test to test, and to reduce range equalisation bias and centring bias [40]. They also allowed each listener's discrimination ability to be checked (see Sections 2.1 and 3.1). The listeners were not informed of the anchors' presence. The anchors are detailed in Table III. The high anchor was the unprocessed reference recording. The mid and low anchors were degraded using processes (representative of those used to generate the test stimuli) that a series of pilot studies [22] showed to produce appropriate levels of spatial degradation.

Table III Anchor recordings used in listening tests 1 and 2

Anchor	Anchor description
Anchor recording A	High anchor: unprocessed hidden reference
Anchor recording B	Mid anchor: audio codec (80 kbs)
Anchor recording C	Low anchor: mono down-mix reproduced asymmetrically by the rear left loudspeaker only

2 LISTENING TEST 1 RESULTS & DISCUSSION

Listening test 1 compared SAP-degraded audio to unprocessed reference stimuli, at both central and off-centre listening positions. The SAPs employed in this test are indicated in Table 11 and were applied to programme items 1–3. The following sections investigate the degree of perceived degradation and the factors affecting it. The intention is: (i) to identify any data relating to unreliable listeners or a lack of inter-listener consensus, since these data would be unsuitable for inclusion in the database which will be used in the development of the quality evaluation system; and (ii) where there is consensus among reliable listeners, to learn more about the relationships between SAP, programme item, listening position and quality.

2.1 Data screening

Prior to results analysis each listener's responses were assessed, so that the unreliable data (i.e. data from a listener who lacked discrimination ability or consistency) could be removed. A listener's discrimination ability was established using a one-sided t-test to determine if their scores, throughout the listening test, for Anchor recording A were significantly different ($p < 0.05$, degrees of freedom = 95) from the instructed value of 100; if they were not then that listener was deemed able to successfully identify that recording. A listener's consistency was assured if the RMS difference between their scoring of initial and repeat presentations of each SAP stimulus was less than 15 %. Although lower thresholds have been used in other studies [41] a higher threshold was chosen here due to the difficulty of the task.

The complete data sets of four of a total of 102 listeners were removed.

2.2 Analysis of Variance

After screening, the distributions of the SAP scores were assessed for normality using the Kolmogorov-Smirnov test (Field [42] cites this as being the most important test to guide choice of analysis technique). This showed 55 % of the data to be normally distributed, indicating that parametric testing would be most suitable [ibid.]. A univariate Analysis of Variance (ANOVA) was conducted, with the independent variables included as fixed factors, to investigate the main effects of the independent test variables (SAP, listening position, programme item, session and listener), and their first-order interactions, on perceived spatial quality (dependent variable) ($r^2 = 0.908$). The results for the variables of interest are presented in Table IV. Session was found to have no significant effect.

Table IV ANOVA: significance and effect size of independent variables and interactions in listening test 1

Independent variable	Significance (p)	Partial-eta-squared	F
SAP	<0.001	0.891	1,865
Listener x SAP	<0.001	0.413	12.29
Programme item x SAP	<0.001	0.234	34.77
Listening position x SAP	<0.001	0.111	28.28

2.3 Influence of spatial audio process

SAP had the largest effect on spatial quality ($p < 0.001$, partial-eta-squared = 0.891). Figure 3 shows means and 95 % confidence intervals for all SAPs (including the hidden anchors), averaged across both listening positions and all programme items and listeners. The mean scores and confidence intervals for the SAPs cover the entire range of the test scale and have 95 % confidence intervals narrower than 10 points (10 %) of the scale.

Overall, groups 1–10 predominantly created small (quality scores of 75 % plus) to moderate (quality scores 50 % to 75 %) impairments to the perceived spatial quality. However, some SAPs in groups 1 (down-mixing), 9 (crosstalk) and 10 (virtual surround) produced large changes to inter-channel relationships (sometimes to the extent that the resulting auditory image was perceived as being in-head, as with SAPs 29 and 37 for example) and reduced quality severely (quality scores less than 50 %). Many SAPs in group 11 (combinations of 1–10) also created severe impairments. This is not surprising as these SAPs compound the degradation created by two different processes.

In group 2 (multi-channel audio coding), only the lowest bit-rate process achieved a mean score of less than 50 % (and even then not significantly so). The SAPs in groups 3 (altered loudspeaker locations), 4 (channel rearrangements), and 8 (spectral filtering) also reduced quality by small to moderate amounts, again with no mean scores significantly less than 50 %. No group 3 SAP produced a mean quality score significantly below 70 %.

The smallest impairment to spatial quality was created by SAP 1 (3/1 down-mix) but, in general, groups 5 (inter-channel level misalignment), 6 (inter-channel out-of-phase errors) and 7 (channel removal) seemed least capable of degrading quality (with no score significantly below 75 %).

The anchor recordings (group 12) were all scored in their expected locations. Anchor recording A, the unprocessed reference, was scored at 100 %. (NB. The confidence intervals for this group are small due to the anchors appearing on every test page and therefore being assessed many more times than the other SAPs)

2.4 Influence of listener

The interaction between listener and SAP had the second largest effect on perceived spatial quality ($p < 0.001$, partial-eta-squared = 0.413) and this suggests that there was a difference in opinion or lack of consensus between listeners with respect to the qualities of certain stimuli. Further experimental work might provide insights into the reason(s) for this lack of consensus (listener reliability was validated in Section 2.1 and so this will not be a factor) but, for the purpose of the analysis presented in this paper, it will be sufficient to identify the stimuli concerned. A number of statistical and visual analysis

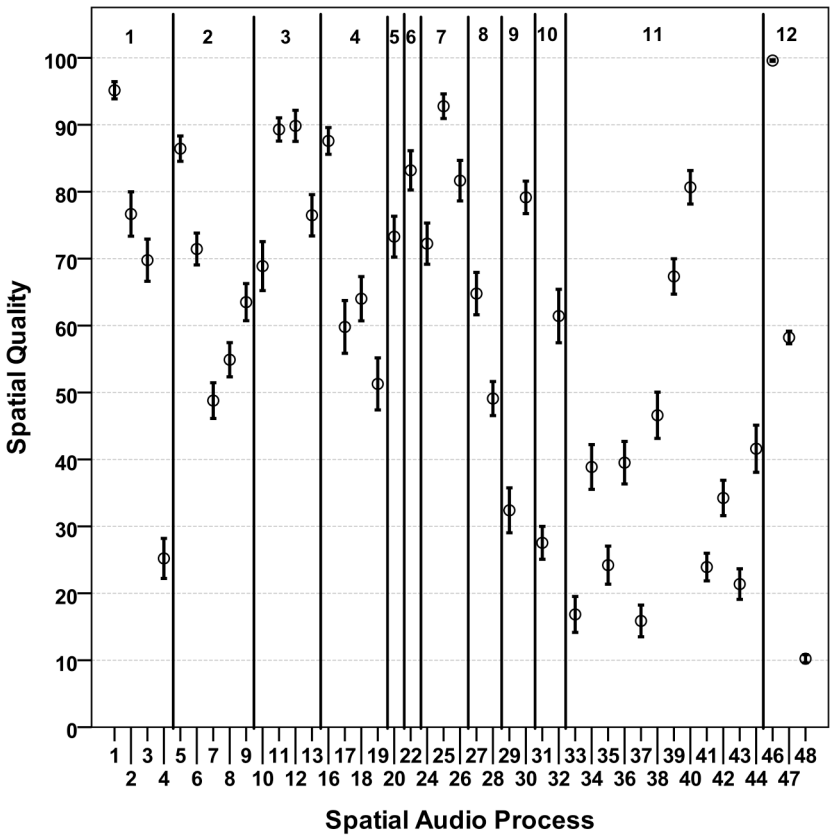


Figure 3. Mean spatial quality scores for each SAP in listening test 1, averaged across programme item type, listening position and listener.

techniques—including the Kolmogorov-Smirnov test, modality, standard deviation, data range and kurtosis (z-score) test—were used to identify stimuli, which have scores exhibiting a multi-modal, wide or platykurtic distribution (Figure 4a); for comparison, a stimulus with statistically normal and reliable average score is depicted in Figure 4b. Score averages for stimuli producing platykurtic distributions will not be meaningful or reliable; therefore the effects of the corresponding SAPs on spatial quality cannot be defined. Consequently, results relating to stimuli where this effect is observed—where the standard deviation of the data distribution is greater than 20, the data range is greater than 75 % and kurtosis score is greater than -1—should not feed into the development of a quality evaluation system. Combinations of programme item, SAP and listening position identified as having unreliable average scores are listed in Table V, which shows that 11 % of the data should be removed. In cases where the distribution was non-normal but leptokurtic, and the other tests had been passed, the median value will be taken to be a reliable average score.

Table V Stimuli producing unreliable average scores in listening test 1 (refer to Table XI for descriptions)

Listening position	Programme item	Spatial audio process
1	1	7, 32, 33
	2	7, 17, 19, 22, 27, 34, 36, 38, 44
	3	6, 7, 19, 32, 44
2	1	18, 19, 20
	2	4, 19, 27, 29
	3	4, 8, 19, 27, 29

2.5 Influence of programme item

The interaction of programme item type with SAP had a significant effect on perceived spatial quality ($p < 0.001$, partial-eta-squared = 0.234). This indicates that certain SAPs degraded spatial quality more for some programme items than for others. Therefore, in the development of a spatial quality evaluation system, SAP scores obtained from one programme item should ideally be considered separately from those obtained from another. A one-way ANOVA using programme item as the factor was used to determine which SAPs exhibited this effect ($p < 0.05$), and these are listed in Table VI.

Table VI SAPs producing significantly different scores for different programme items in listening test 1 (refer to Table XI for descriptions)

Listening position	Spatial audio process
1	1, 2, 3, 5, 9, 10, 11, 12, 13, 16, 17, 18, 19, 22, 24, 25, 26, 32, 33, 34, 38, 39, 40, 42, 44, 46, 47
2	1, 2, 3, 5, 9, 10, 11, 12, 13, 16, 17, 19, 22, 24, 26, 30, 32, 34, 39, 40, 41, 42, 43, 46, 47

In many cases the difference in the perceived spatial quality between programme items can be accounted for by differences in spatial scene-type. For example, a far smaller impairment resulted when a 3.0 down-mix was applied to programme item 2 (classical music) than when it was applied to items 1 (TV/sport) and 3 (rock/pop). This is because the rear channels of item 2 contained only background ambient or reverberant information, which was included to enhance the spaciousness or presence in the recording. This background content was diffuse and not very localisable, and so down-mixing it into the front channels did not create an overly degrading impairment. This is different from programme items 1 and 3 whose

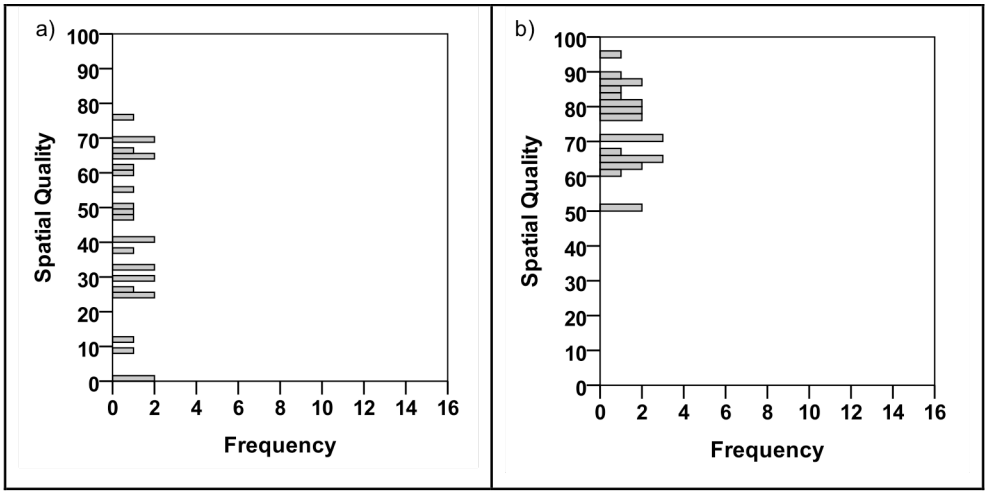


Figure 4. (a) Stimulus producing a platykurtic score distribution. (b) Stimulus producing a statistically normal score distribution. (Distributions categorised by tests described in §2.4.)

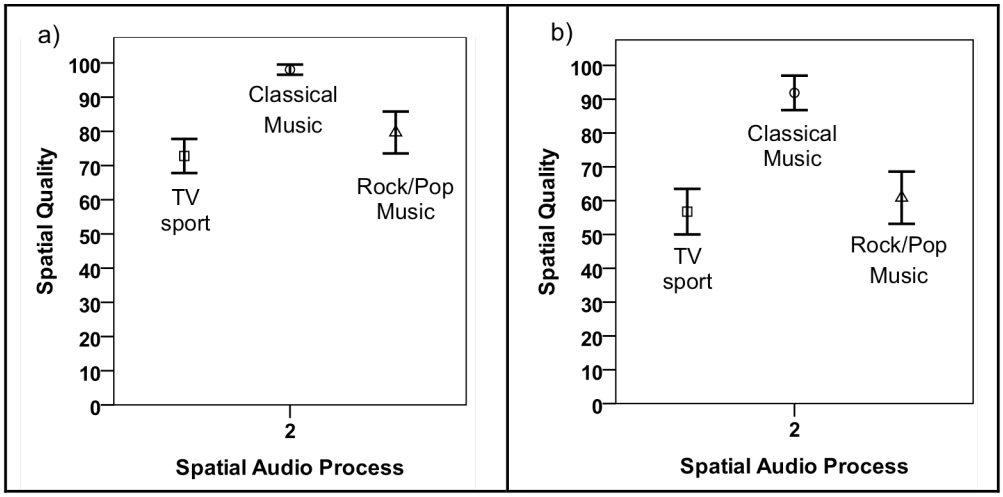


Figure 5. The mean quality degradation resulting from a down-mixing process (SAP 2) is greater for TV/sport and rock/pop music programme items than for classical music, at both (a) LP1 and (b) LP2.

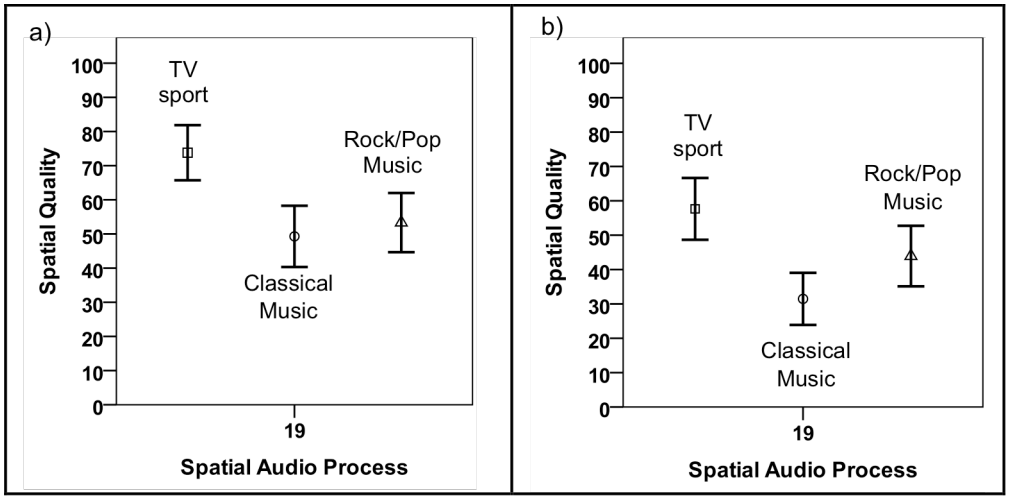


Figure 6. The mean quality degradation resulting from a channel-swapping process (SAP 17) is greater for classical music and rock/pop programme items than for TV/sport, at both (a) LP1 and (b) LP2.

rear channels contained clearly identifiable foreground sources. The effect occurs at both listening positions, as shown in Figure 5. A similar observation was made in a study conducted by Zieliński *et al* [36].

Other aspects of audio content may also have been factors. For example, when the channel order of programme item 1 was changed randomly, a lesser impairment resulted than when the same process was applied to item 2 or item 3. This can be explained by the fact that most of the channels in programme item 1 contain audience applause whose location in the audio scene is unimportant. Hence the channels can be re-routed at random without significant impairment to the overall spatial quality (nevertheless, a slight impairment was created because channels containing the commentators voices were also re-routed). Conversely, re-routing channels in programme items 2 and 3 destroyed the intended audio image. Again this effect occurs at both listening positions, as shown in Figure 6.

2.6 Influence of listening position

The interaction of listening position with SAP had a significant effect on perceived spatial quality ($p < 0.001$, partial-eta-squared = 0.111). This suggests that certain SAPs impaired spatial quality more when auditioned at one listening position than when auditioned at the other. Therefore, in the development of a spatial quality evaluation system, as with scores for different programme items, SAP scores at LP1 should ideally be considered separately from those at LP2. A one-way ANOVA with listening position as the factor was used to determine which stimuli exhibited this effect ($p < 0.05$), and these are listed in Table VII.

The effect can be explained by the physical location change between LP1 and LP2 altering the audio

information that listeners received. For example, when the rear loudspeakers were misplaced to -90º and 90º respectively (SAP 12), only a small impairment to spatial quality resulted at LP1; this fits with the Minimum Audible Angle theory which predicts the inability of the human auditory system to accurately locate sound sources positioned in an area around each ear (at approximately ±90º) [43]. However, LP2 is closer to the right surround loudspeaker and so the misplacement of the rear loudspeakers was likely to have been much more obvious, making the impairment greater and the SAP score lower. This effect is observed for all three programme item types, as shown in Figure 7.

Table VII SAPs producing significantly different scores for different programme items in listening test 1 (refer to Table XI for descriptions)

Programme item	Spatial audio process
1	1, 2, 12, 13, 19, 20, 22, 25, 29, 30, 31, 33, 34, 38, 39, 30, 44, 47
2	1, 2, 5, 12, 13, 19, 22, 24, 26, 29, 31, 42, 44, 47
3	2, 3, 12, 13, 17, 24, 29, 31, 33, 34, 35, 36, 38, 39, 40, 44, 47

3 LISTENING TEST 2 RESULTS & DISCUSSION

Listening test 2 compared SAP-degraded audio at central and off-centre listening positions to centrally-auditioned unprocessed reference stimuli. The SAPs employed in this test are indicated in Table XI and were applied to programme items 4–6. The following sections investigate the degree of perceived degradation and the factors affecting it. As with the test 1 analysis, the intention is: (i) to identify any data relating to unreliable listeners or a lack of inter-listener consensus, since these data would be unsuitable for inclusion in the database which will be used in the development of the quality evaluation system; and (ii) where there is consensus among reliable listeners, to learn more about the relationships between SAP, programme item, listening position and quality.

3.1 Data screening

Prior to analysis each listener's responses were assessed in the same manner as listening test 1, so that the most reliable data could be selected for investigation. The complete data sets of thirteen of a total of 68 listeners were removed.

3.2 Analysis of Variance

After screening, the distributions of the SAP scores were assessed for normality using the Kolmogorov-Smirnov test. This showed 65 % of the data to be normally distributed, meaning that parametric testing could be employed. A univariate ANOVA was conducted, with the independent variables included as fixed factors, to investigate the main effects of the independent test variables (SAP, listening position, programme item, session and listener), and their first-order interactions, on perceived spatial quality (dependent variable) ($r^2 = 0.861$). The results for the variables of interest are presented in Table VIII. Session was again found to have no significant effect.

Table VIII ANOVA: significance and effect size of independent variables and interactions in listening test 2

Independent variable	Significance (p)	Partial-eta-squared	F
SAP	<0.001	0.682	466.9
Listener × SAP	<0.001	0.433	12.26
Programme item × SAP	<0.001	0.128	16.02
Listening Position	<0.001	0.085	444.0

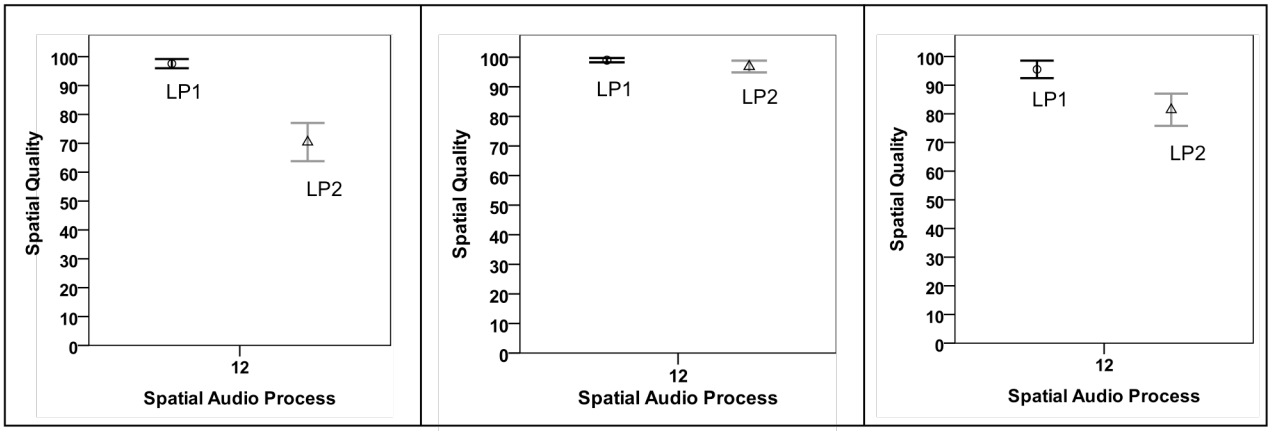


Figure 7. The mean quality degradation resulting from a loudspeaker mis-positioning (SAP 12) is greater at LP2 than at LP1, for programme items (a) 1, (b) 2 and (c) 3.

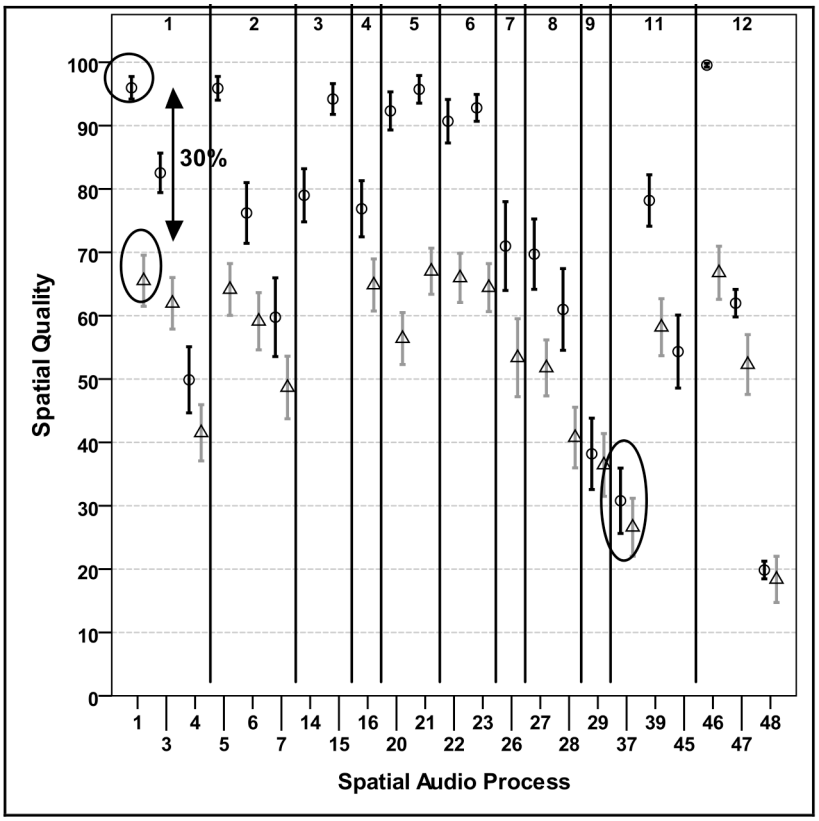


Figure 8. Mean spatial quality ratings for each SAP in listening test 2, averaged across programme item type and listener; note the compression of the rating range at LP2 (triangles) compared to that at LP1 (circles).

3.3 Influence of spatial audio process & listening position

As with listening test 1, SAP had the largest effect on spatial quality ($p < 0.001$, partial-eta-squared = 0.682) in listening test 2. Figure 8 shows means and 95 % confidence intervals for all SAPs (including the anchors) for both LP1 and LP2, averaged across all programme items and listeners. As with the test 1 results, the mean scores and confidence intervals for the evaluated spatial audio processes cover the entire range of the test scale and in all but a few cases have 95 % confidence intervals narrower than 10 points (10 %) of the scale.

Trends in terms of which groups exhibited small, moderate and severe quality impairments are the same as those observed in listening test 1: again groups 1–8 predominantly showed small to moderate quality impairments, but with some SAPs in group 1 (SAP 4: 1.0 downmix), group 9 (SAP 29: 1.0 downmix in all channels) and group 11 (SAP 37: 1.0 downmix + 500 HPF on all channels) reducing quality severely; groups 2, 3, 4 and 8 reduced quality by small to moderate amounts; groups 5, 6 and 7 exhibited only small impairments; and the anchors (group 12) were all scored in their intended locations.

Separating the scores for LP1 (circles) and LP2 (triangles) illustrates how spatial quality was further impaired when listening off-centre. Similar overall scoring trends are observable in the LP1 and LP2 data. However, the range of the scores for LP2 is compressed into the lower part of the test scale. The difference in perceived quality between LP1 and LP2 for the highest quality SAPs is as much as 30 % (e.g. SAP 1, circled), whereas the difference between LP1 and LP2 scores for the lowest-rated SAPs is less than 5 %, and is statistically not significant (e.g. SAP 37, also circled). This smaller difference for the lowest-rated SAPs suggests that the impairment to spatial quality resulting from these processes is already so severe that a shift in the listening position is unable to produce any further degradation.

3.4 Influence of listener

As in listening test 1, listeners' scores reveal a difference in opinion and a lack of consensus for certain stimuli ($p < 0.001$, partial-eta-squared = 0.433). This was investigated further (as in listening test 1) to determine that 19 % of stimuli should be treated as having unreliable average scores. Table IX summarises the results of this analysis.

3.5 Influence of programme item

The interaction of programme item type with process is again shown to have a significant effect on perceived spatial quality ($p < 0.001$, partial-eta-squared = 0.128). A one-way ANOVA using programme item as the factor was used to determine which stimuli exhibited this effect.

The list of SAPs where this test was found to be statistically significant ($p < 0.05$) is given in Table X.

Table IX Stimuli producing unreliable average scores in listening test 2 (refer to Table XI for descriptions)

Listening position	Programme item	Spatial audio process
1	4	4, 6, 7, 27, 28, 29, 45
	5	27, 29, 45
	6	4, 26, 27, 28, 29, 37
2	4	4, 28
	5	4, 16, 20, 28
	6	26, 29, 37

Table X SAPs producing significantly different scores for different programme items in listening test 2 (refer to Table XI for descriptions)

Listening position	Spatial audio process
1	1, 3, 6, 14, 15, 22, 26, 39
2	6, 20, 26, 39

4 SUMMARY & CONCLUSIONS

SAPs commonly encountered in consumer audio reproduction systems are known to generate a range of impairments to spatial quality. By way of two listening tests, this paper investigated the degree of degradation of the spatial quality of six 5-channel audio recordings, resulting from 48 such SAPs, and the influences of listening position and source material on that degradation, and built a quality-annotated database of processed and unprocessed programme items.

Choice of SAP has a large effect on degradation degree. SAPs producing large changes to inter-channel relationships (down-mix and virtual surround algorithms and the introduction of high levels of crosstalk) can reduce quality severely (quality scores significantly <50 %), as can combinations of multiple SAPs. Conversely, inter-channel level and phase misalignment, and channel removal, seem able to degrade quality only slightly (no score significantly below 75 %). Other SAPs (lossy coding, moved or missing loudspeakers and spectral filtering) fall between these two extremes (no score significantly below 50 %). Future development of a spatial audio quality evaluation system must therefore take into account the effects of a wide range of SAPs.

The effect of the interaction between listener and SAP can also be large (although, in this study, less than that of SAP alone). Although the majority (86 %) of the collected data show inter-listener consensus, it appears that for some stimuli there is disagreement between listeners with regard to the degree of degradation present. Means of data relating to such stimuli cannot be treated as reliable and so should not feed into the development of a future spatial audio quality evaluation system. For the majority of stimuli evaluated, however, there is agreement between listeners and so score averages relating to the bulk of the data collected can be used.

There can also be a noticeable effect from the interaction between SAP and programme item. This effect is observable for some SAPs more than others. SAPs that alter the playback positions of one or more channels (e.g. down-mixing algorithms, repositioned loudspeakers, channel-order changes) seem particularly susceptible to this interaction, which in many cases can be accounted for by variations in spatial scene type from item to item (e.g. whether or not the surround channels contain distinct sound sources). The size and frequency of this interaction effect means that, in the development of a spatial quality evaluation system, SAP scores obtained from one programme item should ideally be considered separately from those obtained from another. If a single particularly revealing programme item is sought for SAP quality testing then an item having foreground sources in every channel should be chosen.

Listening position is important in two respects. Firstly, interaction effects are observable: listening position can affect the degree of perceived quality degradation resulting from a SAP. This is particularly evident when a primary effect of a SAP is to alter the output or position of a loudspeaker that is closer to the listener in an off-centre listening position. Therefore, as with programme items, in follow-on work SAP scores obtained at one listening position should ideally be considered separately from those obtained at another. Secondly, combining off-centre listening with another SAP can reduce quality by as much as 30 % compared to auditioning that SAP centrally, but the additional deleterious effects of off-centre listening lessen (to insignificance) when a severely degrading SAP is used.

Taken together, these findings, and the quality-annotated database, can guide the development of a regression model of perceived overall spatial audio quality, incorporating previously developed spatially relevant feature-extraction algorithms. A quality evaluation system based on such a model will have the potential to provide an indication of likely perceived audio quality where human assessment would be impractical or impossible. The development of such a model will be documented in a follow-up paper.

5 ACKNOWLEDGEMENTS

This research was completed as a part of the QESTRAL Project (Engineering and Physical Sciences Research Council EP/D041244/1), a collaboration between the University of Surrey (UK), Bang & Olufsen (Denmark) and BBC Research and Development (UK).

6 REFERENCES

[1] Rumsey, F. (2001) “Spatial Audio”, Focal Press 2001.

[2] Soulodre, G.A., Lavoie, M.C., Norcross, S.G. (2003) “Objective Measures of Listener Envelopment in

Multi-channel Surround Systems". *J. Audio Eng. Soc.*, Vol.51 (9), pp. 826–840.

[3] Davis, M. (2003) "History of Spatial Coding" *J. Audio Eng. Soc.*, Vol.51 No.6, pp. 554–569.

[4] Rumsey, F. (2011) "Spatial Audio—Eighty years after Blumlein", *J. Audio Eng. Soc.*, Vol.59 (1/2), pp. 57–65.

[5] Rumsey, F. (2011) "Audio for Games", *J. Audio Eng. Soc.*, Vol.59 (5), pp. 341–345.

[6] Rumsey, F. (2002) "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a Scene-Based Paradigm". *J. Audio Eng. Soc.*, Vol.50 (9), pp. 651–666.

[7] Belloch, J.A., Ferrer, M., Gonzalez, A., Martinez-Zaldivar, F.J. & Vidal, A.M. (2013) "Headphone-Based Virtual Spatialization of Sound with a GPU Accelerator", *J. Audio Eng. Soc.*, Vol.61 (7/8), pp. 546–561.

[8] BBC (2009) "Surround Sound" http://www.bbc.co.uk/bbchd/what_is_hd.shtml [Accessed 11/08/10].

[9] BSkyB Ltd (2009) "Experience more with Sky+HD" http://packages.sky.com/hd/?DCMP=ILC-SkyCOM_HD [Accessed 11/08/10].

[10] Marins, P., Rumsey, F. & Zieliński, S. (2008) "Unravelling the relationship between basic audio quality and fidelity attributes in low bit-rate multi-channel audio codecs" presented at the *Audio Engineering Society 124th Convention*, May 17–20, Amsterdam, Netherlands, Preprint 7335.

[11] ITU-R BS.775-1 (1992-1994) "Multi-channel stereophonic sound system with and without accompanying picture" International Telecommunication Union recommendation.

[12] Zieliński, S., Rumsey, F., Bech, S. (2003) "Comparison of Quality Degradation Effects Caused by Limitation of Bandwidth and by Down-mix Algorithms in Consumer Multichannel Audio Delivery Systems" *presented at Audio Engineering Society 114th Convention*, 22–25 March, Amsterdam, The Netherlands, Paper 5802.

[13] Rumsey. F. (2013) "Spatial audio processing", *J. Audio Eng. Soc.*, Vol.61 (6), pp. 474–478.

[14] Bech, S. & Zacharov, N. (2006) "Perceptual audio evaluation: theory, method and application". John Wiley and Sons Ltd., West Sussex, England.

[15] ITU-R BS.1387 (2001) "Method for objective measurements of perceived audio quality"

International Telecommunication Union recommendation.

[16] Liebetrau, J., Sporer, T., Kämpf, S. & Schneider, S. (2010) "Standardization of PEAQ-MC: Extension of ITU-R BS.1387-1 to multichannel audio" presented at the Audio Engineering Society 40th International Conference: Spatial Audio, October 8–10, Tokyo, Japan.

[17] Choi, I., Shinn-Cunningham, B.G., Chon, S. B, & Sung, K. (2008) "Objective Measurement of

Perceived Auditory Quality in Multi-channel Audio Compression Coding Systems" *J. Audio Eng. Soc.*, Vol. 56 (1/2), pp. 3–17.

[18] George, S. (2009) "Objective models for predicting selected multi-channel audio quality attributes" PhD Thesis, Institute of Sound Recording, University of Surrey.

[19] Seo, J-H., Choi, I., Chon, S. B., & Sung, K-M (2010) "Improved prediction of multichannel audio quality by the use of envelope ITD of high frequency sounds" *presented at the Audio Engineering Society 38th International Conference: Sound Quality Evaluation*, June 13–15, Piteå, Sweden.

[20] Dewhirst, M. (2008) "Modelling perceived spatial attributes of reproduced sound" PhD Thesis, Institute of Sound Recording, University of Surrey. http://epubs.surrey.ac.uk/2081/

[21] Jackson, P.J.B, Dewhirst, M., Conetta, R., Rumsey, F., Zieliński, S., Bech, S., Meares D. & George, S (2008). "QESTRAL (Part 3): System and metrics for spatial quality prediction" *presented at the Audio Engineering Society 125th Convention*, Oct 2–5, San Francisco, USA, Preprint 7597.

[22] Conetta, R. (2011) "Towards the automatic assessment of spatial quality in the reproduced sound environment" PhD Thesis, Institute of Sound Recording, University of Surrey. http://epubs.surrey.ac.uk/39628/

[23] Letowski, T. (1989) "Sound Quality Assessment: Cardinal Concepts" presented at the 87th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, Vol.37 pp.1062, Preprint 2825.

[24] Berg, J. (2009) "The contrasting and conflicting definitions of envelopment" *presented at 126th Audio Engineering Society convention*, 7–10 May, Munich, Germany. Paper 7808.

[25] Berg, J., Rumsey, F. (2006) "Identification of Quality Attributes of Spatial Audio by Repertory Grid Technique" *J. Audio Eng. Soc.*, Vol.54 (5), pp. 365–379.

[26] Choisel S. & Wickelmaier, F. (2005) "Extraction of Auditory Features and Elicitation of Attributes for the Assessment of Multichannel Reproduced Sound" *presented at the Audio Engineering Society 118th Convention*, May 28–31, Barcelona, Spain, Preprint 6369.

[27] Koivuniemi, K., Zacharov N., (2001) "Unravelling the Perception of Spatial Sound Reproduction: Language Development, Verbal Protocol Analysis and Listener Training" *presented at the Audio Engineering Society 111th Convention*, Nov 30–Dec 3, New York, USA, Preprint 5424.

[28] Rumsey, F., Zieliński, S., Kassier, R., and Bech, S. (2005). Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences. *J. Acoust. Soc. Am.*, 117(6), pp. 3832–3840.

[29] Zacharov N., Koivuniemi, K. (2001a) "Unravelling the perception of spatial sound reproduction: Techniques and experimental design" *presented at the Audio Engineering Society 19th International Conference*, June 21–24, Schloss Elmau, Germany, Paper number 1929.

[30] Zacharov, N., Koivuniemi, K. (2001b) "Unravelling the Perception of Spatial Sound Reproduction: Analysis & External Preference Mapping" *presented at the Audio Engineering Society 111th Convention*, September, New York, USA, Preprint 5423.

[31] Clark, H.A.M., Dutton, G.F. & Vanderlyn, P.B. (1958) "The 'Stereosonic' Recording and Reproducing System" *J. Audio Eng. Soc.*, Vol. 6 (2), pp. 102–117.

[32] Eargle, J. (1986) "An Analysis of Some Off-Axis Stereo Localization Problems" *presented at 81st Audio Engineering Society convention*, Los Angeles, California, 12–16 November. Paper 2390.

[33] Merchel, S. & Groth, S. (2010) "Adaptively Adjusting the Stereophonic Sweet Spot to the Listener's Position" *J. Audio Eng. Soc.*, Vol. 58 (10), pp. 809–817.

[34] ITU-R BS.1116-1 (1997) "Methods for the subjective assessment of small impairments in audio systems including multi-channel sound systems" International Telecommunication Union recommendation.

[35] Bang & Olufsen (2011) "Beolab 3 specifications" http://www.bang-olufsen.com/en/sound/loudspeakers/beolab-3 [Accessed 06/09/14].

[36] Zieliński, S., Rumsey, F., Bech, S. & Kassier, R. (2003c) "Effects of down-mix algorithms on quality of surround sound" *J. Audio Eng. Soc.*, Vol. 51 (9), pp.780–798.

[37] Zieliński, S., Rumsey, F. & Bech, S. (2003b) "Effects of Bandwidth Limitation on Audio Quality in Consumer Multichannel Audiovisual Delivery Systems" *J. Audio Eng. Soc.*, Vol. 51 (6), pp.475–501.

[38] Zieliński, S., Rumsey, F., Bech, S. & Kassier, R. (2005) "Comparison of Basic Audio Quality and Timbral and Spatial Fidelity Changes Caused by Limitation of Bandwidth and by Down-mix Algorithms in 5.1 Surround Audio Systems" *J. Audio Eng. Soc.*, Vol. 53 (3), pp.174–192.

[39] ITU-R BS.1534 (2001) "Method for the subjective assessment of intermediate audio quality" International Telecommunication Union recommendation.

[40] Zieliński, S., Rumsey, F. & Bech, S. (2008) "On Some Biases Encountered in Modern Audio Quality Listening Tests—A Review" *J. Audio Eng. Soc.*, Vol.56 (6), pp. 427–451.

[41] Rumsey, F. (1998) "Subjective Assessment of the Spatial Attributes of Reproduced Sound" *presented at the Audio Engineering Society15th International Conference: Audio, Acoustics & Small Spaces*, Oct 31–Nov 2, Copenhagen, Denmark.

[42] Field, A. (2005) "Discovering Statistics Using SPSS" 2nd Edition, SAGE Publications Ltd, UK.

[43] Moore, B.C.J. (2003) "An introduction to the psychology of hearing" 5th edition, Academic Press, UK.

7 APPENDIX

The instructions given to each listener before commencing listening test 1 and 2 are presented, followed by Table XI which lists SAP descriptions and groupings.

Listener Instructions

Thank you for participating in this experiment.

Please read the instructions below.

Description of subject task and scale for spatial quality score

You are asked to compare a number of spatial sound recordings, which have been processed or degraded in various ways, with an unprocessed original reference recording. You are asked to rate the spatial quality of the processed items. A spatial quality scale is a hybrid scale that is primarily a fidelity evaluation (one measuring the degree of similarity to the reference). However it also enables you to give an opinion about the extent to which any differences are inappropriate, unpleasant or annoying. In other words, which affect your opinion of the quality of the spatial reproduction compared with the reference. So, for example, if you can hear a change in the spatial reproduction compared with the reference but it doesn't make much difference to your overall opinion about the spatial quality, you should rate it towards the top of the scale. On the other hand, if the spatial change is very pronounced and you consider it to be annoying, unpleasant or inappropriate, you should probably rate it towards the bottom of the scale. In the middle should go items that have clearly noticeable changes in the spatial reproduction and that are only moderately annoying, unpleasant or inappropriate. It is up to you how you interpret these terms but the aim is to come up with an overall evaluation of your opinion of the spatial quality of the processed items compared with the reference. It comes down to a judgement about how acceptable the impairments of the test items are when you know what the original recording (the reference) should sound like.

In order to avoid any potential biasing effects of verbal labels with particular meanings at intervals on the scale, the scale you will use simply has a magnitude and an overall direction labelled 'worse'. Any item rated at the top of the scale should be considered as identical to the reference. Try to use the whole scale, rating the worst items in the test at the bottom of the scale and the best ones at the top. Try to ignore any changes in quality that are not spatial, unless they directly affect spatial attributes.

The following are examples of changes in spatial attributes that you may hear and may incorporate in your overall evaluation (in no particular order of importance, and not meant to exclude any others you may hear):

Changes in location

Changes in rotation or skew of the spatial scene

Changes in width

Changes in focus, precision of location or diffuseness

Changes in stability or movement

Changes in distance or depth

Changes in envelopment (the degree to which you feel immersed by sound)

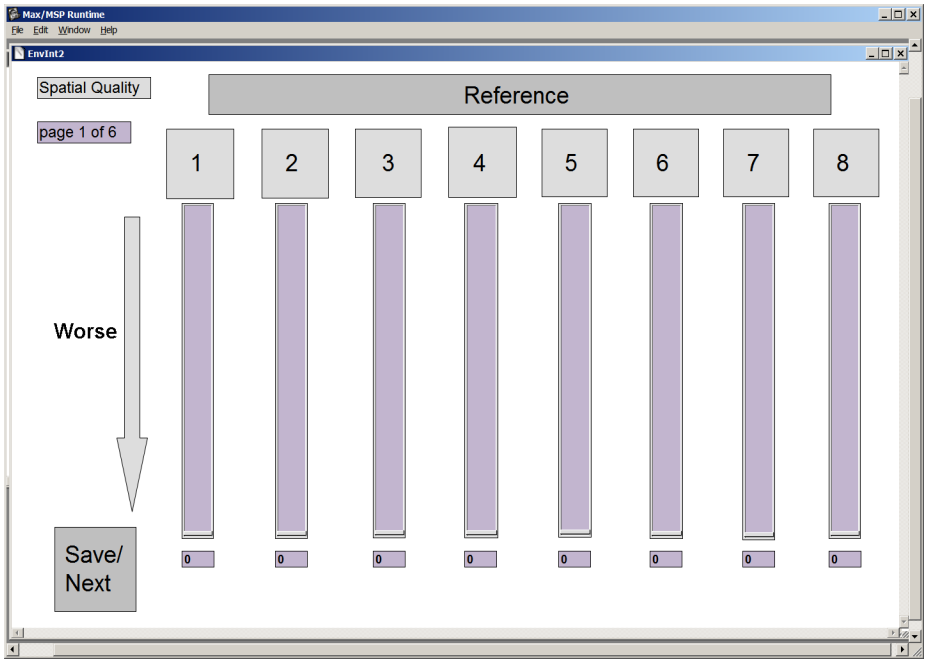
Changes in continuity (appearance of 'holes' or gaps in the spatial scene)

Changes in perceived spaciousness (the perceived size of the background spatial scene, usually implied by reverberation, reflections or other diffuse cues)

Other unnatural or unpleasant spatial effects (e.g. spatial effects of phasiness)

User Interface

Each page contains 8 test recordings to be evaluated for **spatial quality** against a reference recording.



This experiment consists of 12 pages split over two parts, 'a' and 'b'.

When you come to the end of each part you will be prompted to save your responses. Please enter your initials followed by the test id (eg. RCa and RCb).

Once you are happy with your responses click the save/next button to continue to the next page (NB. You'll we need to move each fader at least once (even if intend to return it to zero) before you can proceed to the next page).

Familiarisation

Before commencing the experiment you are required to complete a familiarisation session. This aims to familiarise you with the entire stimuli set that you will encounter in this study. Please think about how you would scale (rate) the spatial quality for each.

Table XI SAPs assessed at LP1 and LP 2 in listening test 1 and listening test 2. All test items use 5 reproduction channels except where the description states otherwise (e.g. downmixing, channel removal)

SAP group	No.	Description	Listening test 1		Listening test 2	
			LP1	LP2	LP1	LP2
1	1	3/1 downmix: L = L, R = R, C = C, S = 0.7071*Ls + 0.7071*Rs.	✓	✓	✓	✓
	2	3.0 downmix: L = L + 0.7071*Ls, R = R + 0.7071*Rs, C = C.	✓	✓	×	×
	3	2.0 downmix: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.7071*Rs.	✓	✓	✓	✓
	4	1.0 downmix: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	✓	✓	✓	✓
2	5	Audio codec @ 160 kbs	✓	✓	✓	✓
	6	Audio codec @ 64 kbs	✓	✓	✓	✓
	7	Audio codec @ 64 kbs	✓	✓	✓	✓
	8	2 stage cascade (80 kbs)	✓	✓	×	×
	9	4 stage cascade (64 kbs)	✓	✓	×	×
3	10	L and R re-positioned at -10º and 10º	✓	✓	×	×
	11	C is skewed; re-positioned at 20º	✓	✓	×	×
	12	Ls and Rs re-positioned at -90º and 90º	✓	✓	×	×
	13	Ls and Rs re-positioned at -170º and 160º	✓	✓	×	×
	14	L and C moved 1m to right and not facing listening position	×	×	✓	×
	15	Ls moved 1m to right and not facing listening position	×	×	✓	×
4	16	L and R swapped	✓	✓	✓	✓
	17	L and R swapped for Ls and Rs	✓	✓	×	×
	18	Channel order rotated	✓	✓	×	×
	19	Channel order randomised	✓	✓	×	×
5	20	L, C and R each attenuated by 6 dB	✓	✓	✓	✓
	21	Ls and Rs each attenuated by 6 dB	×	×	✓	✓
6	22	C phase-inverted	✓	✓	✓	✓
	23	L, C and R phase-inverted	×	×	✓	✓
7	24	R removed	✓	✓	×	×
	25	Ls removed	✓	✓	×	×
	26	C removed	✓	✓	✓	✓
8	27	500 Hz HPF on all channels	✓	✓	✓	✓
	28	3.5 kHz LPF on all channels	✓	✓	✓	✓
9	29	1.0 downmix in all channels	✓	✓	✓	✓
	30	Partly correlated (0.5 bleed in adjacent channel pairs)	✓	✓	×	×
10	31	Line array virtual surround	✓	✓	×	×
	32	2 channel virtual surround	✓	✓	×	×
11	33	Channel order randomised + R, Ls and C removed	✓	✓	×	×
	34	3.0 downmix + R removed	✓	✓	×	×
	35	2.0 downmix + channel order randomised	✓	✓	×	×
	36	2.0 downmix + L and R re-positioned at -10º and 10º	✓	✓	×	×
	37	1.0 downmix + 500 Hz HPF on all channels	✓	✓	✓	✓
	38	L and R re-positioned at -10º and 10º + Ls and Rs re-positioned at -170º and 160º	✓	✓	×	×
	39	Audio codec – 160 kbs + 2.0 downmix	✓	✓	✓	✓
	40	Audio codec – 160 kbs + Ls and Rs re-positioned at -90º and 90º	✓	✓	×	×
	41	Audio codec @ 64 kbs + 1.0 downmix	✓	✓	×	×
	42	Audio codec @ 64 kbs + channel order randomised	✓	✓	×	×
	43	2 channel virtual surround + R removed	✓	✓	×	×
	44	2 channel virtual surround + L and R re-positioned at -10º and 10º	✓	✓	×	×
	45	Audio codec @ 64 kbs + Ls moved 1 m to right and not facing listening position	×	×	✓	✓
12	46	High Anchor: unprocessed reference	✓	✓	✓	✓
	47	Mid Anchor: audio codec (80 kbs)	✓	✓	✓	✓
	48	Low Anchor: mono downmix reproduced asymmetrically by Ls only	✓	✓	✓	✓

THE AUTHORS


Robert Conetta


Tim Brookes


Francis Rumsey


Sławomir Zieliński


Martin Dewhirst


Philip Jackson


Søren Bech


David Meares


Sunish George

Robert Conetta is an Acoustics Engineer at Sandy Brown Associates LLP. Previously he was an Acoustics Consultant at Marshall Day Acoustics and a Research Fellow at the Acoustics Research Centre, London South Bank University. At LSBU he worked with Professor Bridget Shield, Professor Julie Dockrell (IOE) and Professor Trevor Cox (Salford) to investigate the effect of noise and classroom acoustic design on pupil performance on the ISESS project.

Rob studied for his PhD at the Institute of Sound Recording, University of Surrey under the supervision of Professor Francis Rumsey, Dr Sławomir Zieliński and Dr Tim Brookes. He worked as part of a team of researchers, funded and supported by Bang and Olufsen and BBC research, on the QESTRAL (Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener) project. For his contribution to the project, he received University of Surrey's Research Student of the Year Award in 2010.

•

Tim Brookes received the B.Sc. degree in mathematics and the M.Sc. and D.Phil. degrees in music technology from the University of York, York, U.K., in 1990, 1992, and 1997, respectively. He was employed as a Software Engineer, Recording Engineer and Research Associate before joining, in 1997, the academic staff at the Institute of Sound Recording, University of Surrey, Guildford, U.K., where he is now Senior Lecturer in Audio and Director of Research. His teaching focuses on acoustics and psychoacoustics and his research is in psychoacoustic engineering: measuring, modeling, and exploiting the relationships between the physical characteristics of sound and its perception by human listeners.

•

Francis Rumsey is an independent technical writer and consultant, based in the UK. Until 2009 he was Professor and Director of Research at the Institute of Sound Recording, University of Surrey, specialising in sound quality, psychoacoustics, and spatial audio. He led the QESTRAL project on spatial sound quality evaluation from 2006–9. He is currently chair of the AES Technical Council, Consultant Technical Writer and Editor for the AES Journal. Among his musical activities he is organist and choirmaster of St Mary the Virgin Church in Witney, Oxfordshire.

•

Sławomir Zieliński received M.Sc. and Ph.D. degrees in Telecommunications from the Technical University of Gdańsk, Poland. After graduation in 1992, he worked as a lecturer at the same University for eight years. In 2000 Dr Zieliński joined the University of Surrey, UK, where he initially worked as a research fellow and then as a lecturer at the Department of Music and Sound Recording. Since 2009 he has been working as a teacher at the Technical Schools in Suwałki, Poland.

During the past 20 years Dr Zieliński taught classes in a broad range of topics including electronics, electroacoustics, audio signal processing, sound synthesis, studio recording technology, and more recently information and communications technology. He co-supervised six Ph.D. students. In 2007–2008 he was a member of the AES British Section Committee. He is the author or co-author of more than 70 scientific papers in the area of audio engineering. His current research interests include psychoacoustics and audio quality assessment methodology.

•

Martin Dewhirst received an MMath degree from the University of Manchester Institute of Science and Technology, Manchester, UK and a Ph.D. degree from the Institute of Sound Recording and the Centre for

Vision, Speech and Signal Processing at the University of Surrey, Guildford, UK.

He is a lecturer at the Institute of Sound Recording, University of Surrey, where his teaching focuses on signal processing and sound synthesis. His current research interests include the relationship between audio quality and lower level perceptual attributes and modeling the perceived attributes of reproduced sound using objective measurements. Dr. Dewhirst is an associate member of the Audio Engineering Society.

•

Philip Jackson is Senior Lecturer in speech and audio processing at the Centre for Vision, Speech & Signal Processing (University of Surrey, UK) which he joined as in 2002, following a postdoctoral research fellowship (University of Birmingham, UK), with MA in Engineering (Cambridge University, UK) and PhD in Electronic Engineering (University of Southampton, UK). With Dr Wenwu Wang in CVSSP, he leads the Machine Audition Group (A-lab) of around a dozen research fellows and students. His research in acoustical and spoken-language processing has contributed to various projects (e.g., BALTHASAR, DANSA, Dynamic Faces, QESTRAL, UDRC, POSZ and S3A) in active noise control for aircraft, acoustics of speech production, source separation for automatic speech recognition (ASR), use of articulatory models for ASR, audio-visual processing for speech enhancement and visual speech synthesis, as well as spatial aspects of subjective sound quality evaluation. He has over 100 academic publications in journals, conference proceedings and books. He reviews for journals and conferences including Journal of the Acoustical Society of America, IEEE Transactions on Audio, Speech & Language Processing, IEEE Signal Processing Letters, InterSpeech and ICASSP, and is associate editor for Computer Speech & Language (Elsevier).

•

Søren Bech received a M.Sc. and a Ph.D. from the Department of Acoustic Technology (AT) of the Technical University of Denmark. From 1982–92 he was a research Fellow at AT studying perception and evaluation of reproduced sound in small rooms. In 1992 he joined Bang & Olufsen where he is Head of Research. In 2011 he was appointed Professor in Audio Perception at Aalborg University.

His research interest includes human perception of reproduced sound in small and medium sized rooms. experimental procedures and statistical analysis of data from sensory analysis of audio and video quality. General perception of sound in small rooms is also a major research interest.

•

David Meares is a graduate in electrical engineering from Salford University, Salford, UK. In his 38 years at the BBC, he rose to be head of the studio group. He led a wide range of projects, including acoustic scale modeling, digital television, applications of speech recognition, display technology, surround sound, and compression coding. He represented the BBC in a number of international standards groups and on international collaborative projects. This broad experience suits him ideally for the wide number of tasks he has been doing for International Broadcasting Convention over many years. Since introducing the idea 16 years ago, he has organized the New Technology Campus and has served on the papers committee and at various times on the management committee and the conference committee.

•

Sunish George received the B.Tech degree from Cochin University of Science and Technology, Kerala in 1999, and the M.Tech degree in digital electronics and advanced communication from Manipal Institute of Technology, Karnataka in 2003. After his graduations, he worked in various Indian software companies developing digital signal processing-based applications. He completed his PhD from the Institute of Sound Recording, University of Surrey in July 2009. The focus of his doctoral work was to contribute towards the development of a generic objective model that predicts multichannel audio quality. He is currently working at Harman Becker Automotive Systems GmbH, Germany.