



Audio Engineering Society Convention Paper

Presented at the 125th Convention
2008 October 2–5 San Francisco, CA, USA

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

QESTRAL (Part 2): Calibrating the QESTRAL model using listening test data

R. Conetta¹, F. Rumsey¹, S. Zielinski¹, P.J.B. Jackson², M. Dewhirst², S. Bech³, D. Meares⁴ and S. George¹

¹ Institute of Sound Recording (IoSR), University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom

² Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom

³ Bang & Olufsen a/s, Peter Bangs Vej 15, 7600 Struer, Denmark

⁴ DJM Consultancy, Sussex, United Kingdom, on behalf of BBC Research

Correspondence should be addressed to Robert Conetta (r.conetta@surrey.ac.uk)

ABSTRACT

The QESTRAL model is a perceptual model that aims to predict changes to spatial quality of service between a reference system and an impaired version of the reference system. To achieve this, the model required calibration using perceptual data from human listeners. This paper describes the development, implementation and outcomes of a series of listening experiments designed to investigate the spatial quality impairment of 40 processes. Assessments were made using a multi-stimulus test paradigm with a label-free scale, where only the scale polarity is indicated. The tests were performed at two listening positions, using experienced listeners.

Results from these calibration experiments are presented. A preliminary study on the process of selecting of stimuli is also discussed.

1. INTRODUCTION

The QESTRAL model is a perceptual model that aims to predict changes to spatial quality of service (SQoS) between the soundfield reproduced by a reference system and that of an impaired version of

the reference system. To achieve this, the model required calibration using perceptual data from human listeners. This paper describes the development, implementation and outcomes of a series of listening experiments designed to investigate changes in spatial quality.

Similarly to ‘basic audio quality’ (BAQ), which is defined as the attribute accounting for ‘any and all differences between the reference and impaired items’ in an audio system [ITU-R BS.1534, 2001], ‘spatial quality’ is defined here as the attribute that describes any and all differences only between the spatial characteristics of the stimuli (timbral characteristics of sound are omitted). Hence a judgement of spatial quality can be considered as a global assessment of the perceived impairment to quality of changes to a collection of lower level spatial attributes (such as source location, envelopment, source width, source distance, spaciousness etc), when compared with a reference.

As judgements of spatial quality are made on a quality scale, a hedonic component is included. This is similar to BAQ in that it requires the listener to make a judgement about the degree of acceptability or annoyance of the spatial impairments concerned, as well as about the magnitude of the perceived changes in underlying attributes.

In a listening experiment designed to calibrate a perceptual model it is desirable to select stimuli which stress the entire range of conditions whose quality might need to be predicted by that model. The reliable calibration of a spatial quality model also requires that all of the lower level spatial attributes contributing to the overall judgement of spatial quality are adequately stressed by the stimuli in question. A preliminary experiment was designed to ensure that this requirement was fulfilled and that an optimal selection of programme items and processes was achieved. The stimuli selected were then used in a spatial quality experiment which is the main feature of this paper.

All listening tests were performed on an ITU-R BS.775-1 [1994] conformant 5-channel loudspeaker

array, in an ITU-R BS.1116-1 [1997] conformant listening room at two listening positions, using experienced listeners from the Institute of Sound Recording (IoSR). A 5-channel system was used in these experiments because the QESTRAL prototype perceptual model for predicting spatial quality is based on the assumption of a 5-channel reference reproduction against which impaired versions of stimuli are compared.

2. SELECTION OF PROGRAMME ITEMS AND PROCESSES

For this experiment three 5-channel program items were chosen. These were chosen as benchmark examples representing three typical genres: TV Sport, Classical Music and Pop Music. Each exhibited high envelopment and distinctive source locations. Descriptions of these items are provided in table 2.1.

A large number of processes were selected to be applied to each programme item to create a range of spatial quality impairments. Examples of these processes include downmixes from 5-channels to a smaller number of channels, low bit-rate audio coding, loudspeaker misplacements, and channel routing errors.

Informal listening assessments by the authors suggested that the selected processes created a wide range of impairments to spatial quality. However discussions during these informal assessments suggested that the impairments could theoretically be limited to only one or two of the lower level spatial attributes. To address this, a means of determining which of these attributes had been stressed and by how much was required.

No.	Genre Type	Scene Type	Description
1	TV Sport	F-F	Wimbledon. Commentators and clapping. Commentators panned mid-way between L, C and R. Audience clapping in 360°.
2	Classical Music	F-B	Music. Wide continuous front stage including localisable instrument groups. Ambient surrounds with reverb from front stage.
3	Pop Music	F-F	Music. Wide continuous front stage, including guitars, bass and drums. Main vocal in C. Harmony vocals, guitars and drum cymbals in Ls and Rs.

Table 2.1 Description of program items used in experiments

To solve the problem a short listening experiment was designed. In this experiment two experienced listeners from the IoSR assessed each stimulus against an unprocessed reference for changes to a selection of 8 lower level spatial attributes.

1. Audio scene coverage angle
2. Individual source width
3. Ensemble width
4. Envelopment
5. Spaciousness
6. Distance
7. Depth
8. Individual source location

The attributes were selected based upon Rumsey's scene-based paradigm [Rumsey, 2002] and discussions amongst QESTRAL group members.

Judgements were recorded using 4 assessment levels (1. no changes, 2. slight changes, 3. moderate changes and 4. large changes). All stimuli were loudness equalised by ear in informal assessments, this corresponded to a comfortable playback level of approximately 80dBA (L_{EQ}).

2.1. Results and conclusion

Results from an initial assessment indicated that the lower level attributes under investigation had been stressed by the stimuli. However a number of the processes were judged to have created 'no changes' to the attributes, suggesting that if they were used in an experiment, assessments of spatial quality would predominantly lie at the top of an assessment scale.

Results from a second assessment which trialed some new processes, specifically designed to create larger changes indicated that a better selection could be achieved (Fig A1 (Appendix A)). An optimal set of 40 processes (creating a total of 120 stimuli) was then chosen, which stressed all the attributes to differing degrees across a wide spread of the assessment levels (Fig A2 (Appendix A)).

Descriptions of these processes are given in table B1 (Appendix B).

3. SUBJECTIVE ASSESSMENT OF SPATIAL QUALITY

Using this selection of processes a series of listening tests were designed to collect data on impairments to

spatial quality. The tests were undertaken at two listening positions. In order to avoid listener fatigue the 40 processes were blocked into 4 sessions, each including 10 processes (Tables C1-4 (Appendix C)), resulting in 8 tests per listener. The presentation order of the stimuli within each session was randomised. Listeners assessed the 10 processes as well as 3 hidden anchors with all 3 programme items, creating a total of 48 stimuli assessments per session. One session consisted of the test and one repeat and lasted approximately 30 minutes. Before commencing each session listeners completed a familiarisation using the test interface. This enabled them to hear and practice the assessment of each stimulus featured in the session. Fourteen experienced listeners from the IoSR took part in the tests, each listener completed the sessions in order (Fig D1 (Appendix D)). The instructions given to each listener are shown in Appendix E. A diagram of the loudspeaker layout used in the listening tests is illustrated in figure 3.1 (page 4). (NB. Not shown in the diagram is an additional array loudspeaker system used for process 28 and an acoustically transparent curtain, used to hide the loudspeaker positions from the listener).

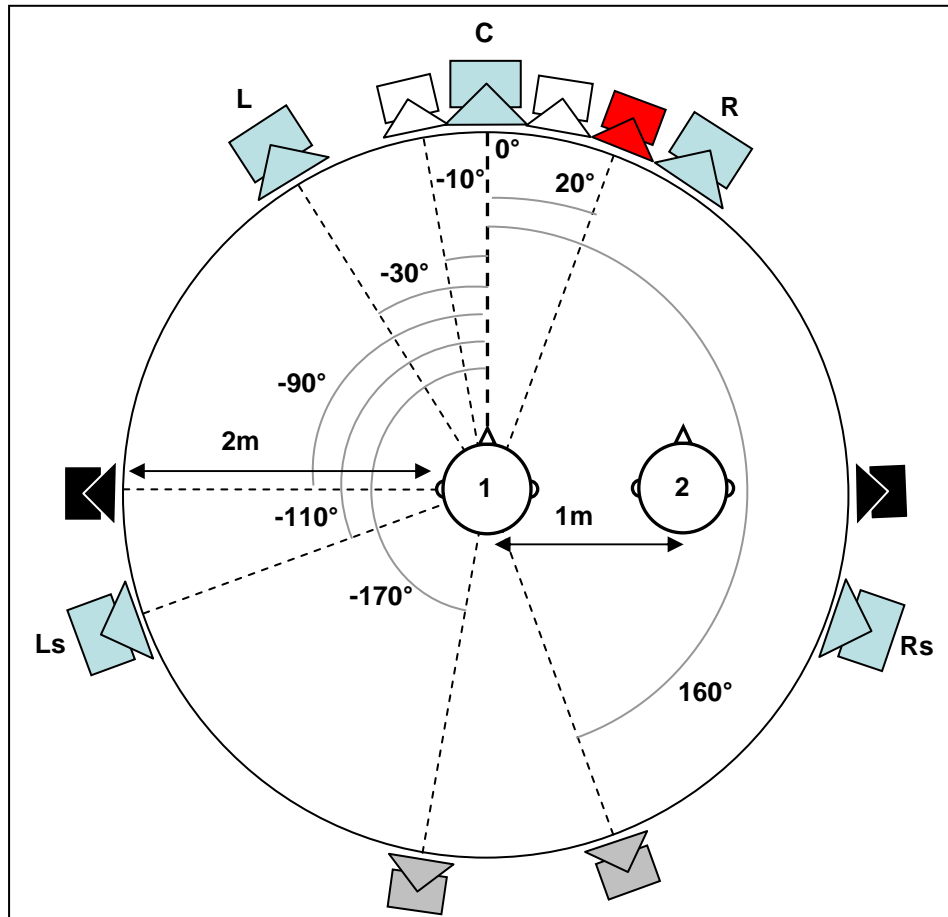


Fig 3.1 Schematic illustrating the listening positions and loudspeaker positions employed during the experiment. Loudspeakers labelled L, C, R, Ls and Rs indicate the ITU-R BS.775 [1994] 5-channel array used as the reference system. Other loudspeaker positions indicate those employed for processes 10-13 (see Table B1).

3.1. Experiment paradigm and Graphical User Interface (GUI)

A multi-stimulus test paradigm similar to MUSHRA (ITU-R BS.1534) [2001] was employed for the experiments. The paradigm used a label free 100 point scale with only the scale polarity indicated. Listeners assessed 8 stimuli per page including 5 processes and 3 hidden anchor processes. They were asked to give the top score (100) for recordings whose spatial quality was identical to that of the reference recordings and to judge any changes to spatial quality as impairments (hence the arrow with the label “Worse”). The GUI is illustrated in figure 3.2.

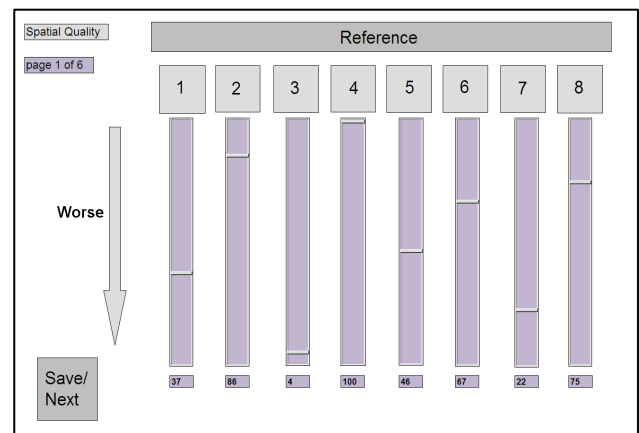


Fig 3.2 Screenshot of GUI.

3.2. Anchor recordings

As mentioned 3 anchor recordings were included in the experiments. These were selected based upon the results of informal listening undertaken by the first author of this paper. The listeners were not informed of the inclusion of these anchors; however they were featured on every page to encourage listeners to utilise the full range of the scale and to reduce the risk of assessment scale biases, such as contraction bias [Zielinski, 2008]. Descriptions of the anchor recordings are given in table 3.1.

Anchor	Anchor description
High	Hidden reference.
Middle	Audio codec (80kbs).
Low	Mono downmix reproduced asymmetrically by the rear left loudspeaker only

Table 3.1 Description of anchor recordings.

4. RESULTS AND DISCUSSION

To analyse the results the data from all 8 sessions was compiled into one data set.

4.1. Post-screening of listeners

An assessment of each listener’s consistency in their scores between test repeats (intra-listener consistency) and the correlation of their scores with the rest of the listener group (inter-listener correlation) was undertaken for each listening position and test. The intra-listener consistency scores ranged between 5% and 20% error, however the majority were centred at 10%, which is similar to the listener error noticed in other tests of a similar nature [e.g. Rumsey, 1998]. Inter-correlation scores revealed that the listeners used the test scale in a similar manner. Hence it was deemed un-necessary to screen any of the listeners

4.2. A Inspection of data distributions

To observe the distribution of the listener scores for each process, histograms for every process condition, both listening position and programme item, were plotted. Some examples are given here. Figure 4.1 illustrates a stimulus with a statistically normal distribution.

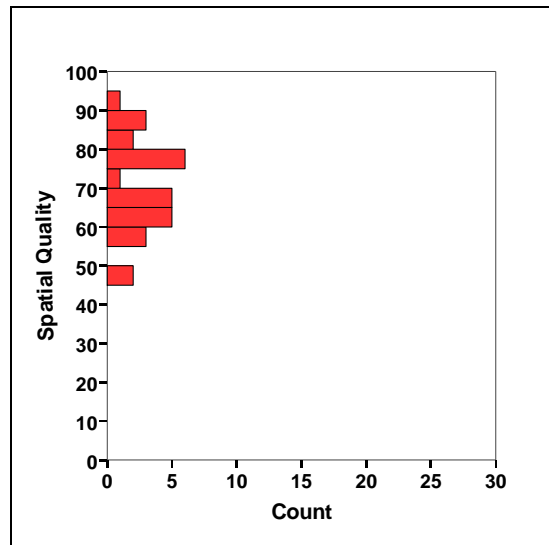


Fig 4.1 Data distribution: Process 2 for programme item 1 listening position 1.

A number of stimuli were revealed as having both wide and statistically multi-modal distributions. A wide distribution of scores can occur when the listeners disagree on the score that a stimulus should be given. This often happens when a stimulus is difficult to evaluate in the task. Multi-modal distributions occur when the listeners fall into two or more groups with differing opinions of where to scale a stimulus. An example of a condition displaying both effects is shown in figure 4.2.

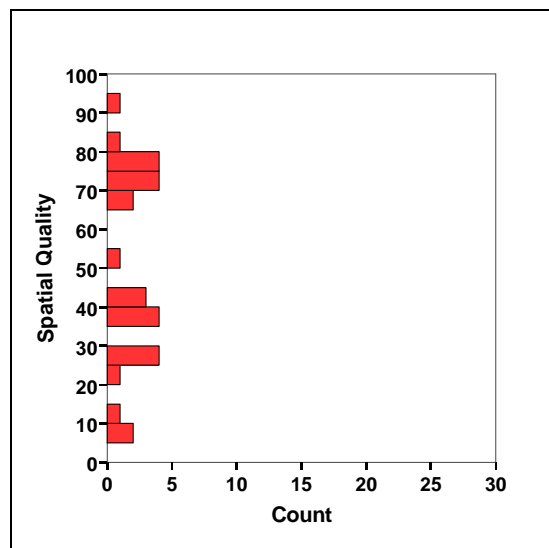


Fig 4.2 Data distribution: Process 10 for programme item 2 listening position 2.

At listening position 2, process 10 for programme item 2 shows multi-modal and also a very flat and wide distribution. In this case there are two (or more) obvious clusters of scores.

Process 10 (Loudspeaker Mis-placement 1) moves the channels L and R from -30° and 30° respectively to -10° and 10° squashing the front image. The distribution of the scores suggests that for some listeners the spatial quality impairment resulting from this process is not too large, when listening from an off-centre listening position (1m to the right of the centre) relative to the reference recording. However for others it is.

Listening Position	Programme Item	Processes with wide or multi-modal data distributions
1	1	17, 23, 28, 34
	2	3, 7, 10, 15, 17, 20, 23, 25
	3	17, 20, 28, 40
2	1	17, 18, 32
	2	3, 10, 15, 16, 17, 20, 25, 32
	3	8, 23, 25, 40

Table 4.1 Stimuli which exhibit wide or multi-modal data distributions.

Table 4.1 lists other stimuli found to exhibit wide or multi-modal distributions. As stated above these distributions indicate that there was no consensus between the listeners in terms of their assessment of spatial quality. The mean scores from these cases are therefore considered to be ambiguous, and should be excluded from the database used in the calibration of the QESTRAL model.

These findings suggest that listening position, programme item type and listener may have a significant effect on perceived spatial quality.

4.3. ANOVA

A univariate ANOVA was conducted to investigate the main effects and 1st order interactions of the experimental factors on spatial quality (Fig 4.3). Process, listening position (LP), programme item (ProgItem), session and listener were included in the model as fixed factors. The structure of the ANOVA model is shown in equation 4.1.

$$Y_{A,B,X,\Delta,E} = \pi + \alpha_A + \beta_B + \chi_X + \delta_\Delta + \varepsilon_E + \phi_{A,B} + \varphi_{A,X} + \gamma_{A,\Delta} + \eta_{A,E} + \iota_{B,X} + \kappa_{B,\Delta} + \lambda_{B,E} + \mu_{X,\Delta} + \nu_{X,E} + o_{\Delta,E} + \omega_{A,B,X,\Delta,E} \quad (\text{eq. 4.1})$$

Where:

π = overall mean,

α_A = process effect,

β_B = listening position effect,

χ_X = programme item effect,

δ_Δ = session effect,

ε_E = listener effect,

$\phi_{A,B}$ = interaction of listening position with process,

$\varphi_{A,X}$ = interaction of programme item with process,

$\gamma_{A,\Delta}$ = interaction of session with process,

$\eta_{A,E}$ = interaction of listener with process,

$\iota_{B,X}$ = interaction of programme item with listening position,

$\kappa_{B,\Delta}$ = interaction of listening position with session,

$\lambda_{B,E}$ = interaction of listener with listening position,

$\mu_{X,\Delta}$ = interaction of programme item with session,

$\nu_{X,E}$ = interaction of listener with programme item,

$o_{\Delta,E}$ = interaction of listener with session,

and $\omega_{A,B,X,\Delta,E}$ = the error.

Tests of Between-Subjects Effects						
Dependent Variable: Spatial Quality						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	10667847.5 ^a	828	12883.874	114.139	.000	.905
Intercept	27424241.5	1	27424241.47	242953.1	.000	.961
Process	8987630.088	42	213991.193	1895.762	.000	.889
LP	9156.158	1	9156.158	81.115	.000	.008
ProgItem	30590.375	2	15295.188	135.501	.000	.027
Session	733.868	3	244.623	2.167	.090	.001
Listener	217014.541	13	16693.426	147.888	.000	.162
Process * LP	146676.614	42	3492.300	30.939	.000	.116
Process * ProgItem	326174.616	84	3883.031	34.400	.000	.226
Process * Session	3544.741	6	590.790	5.234	.000	.003
Process * Listener	741961.872	546	1358.905	12.039	.000	.398
LP * ProgItem	3026.274	2	1513.137	13.405	.000	.003
LP * Session	2726.328	3	908.776	8.051	.000	.002
LP * Listener	12198.683	13	938.360	8.313	.000	.011
ProgItem * Session	732.020	6	122.003	1.081	.371	.001
ProgItem * Listener	26548.902	26	1021.112	9.046	.000	.023
Session * Listener	11071.014	39	283.872	2.515	.000	.010
Error	1120095.639	9923	112.879			
Total	47555925.0	10752				
Corrected Total	11787943.2	10751				

a. R Squared = .905 (Adjusted R Squared = .897)

Fig 4.3 Univariate ANOVA output.

The factor Process had a significant and the largest effect on spatial quality. Session was not significant. 1st and 2nd order interactions reveal that listening position, programme item and listener all had a significant effect on spatial quality. To illustrate the most important experimental factors or interactions figure 4.4 depicts main effects and interactions with an effect size greater than 0.1. These are discussed in the proceeding sections.

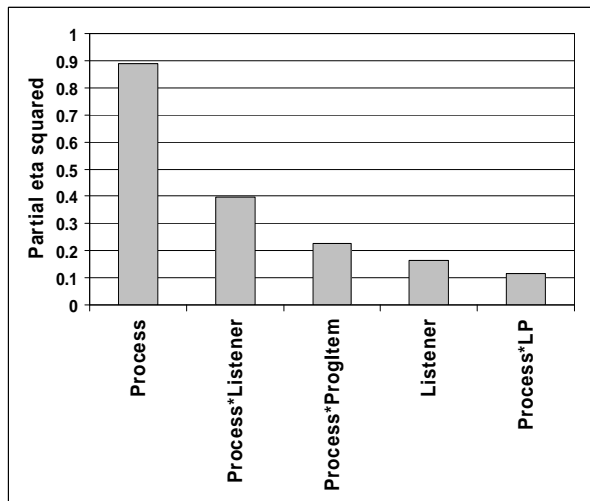


Fig 4.4 Main effects and 1st order interactions with an effect size greater than 0.1.

4.4. The effect of Process on spatial quality

As a summary, figure 4.5 shows means and 95% confidence intervals for all processes and anchors, averaged across both programme item and listening position. However this method of observation is oversimplified and hides the influence of listening position, programme item type and listener revealed by the ANOVA analysis.

Figure 4.5 allows easy comparisons between the processes investigated in the experiment. The results have been divided into groups (Table 4.2).

Group	Process type
1	Down-mixing from 5 CH
2	Audio coding
3	Loudspeaker mis-placement
4	Channel routing errors
5	Inter-channel level mis-alignment
6	Inter-channel out-of-phase errors
7	Missing channels
8	Filtering
9	Inter-channel crosstalk
10	Virtual surround algorithms
11	Combinations of 1-10
12	Anchor recordings

Table 4.2 – Process groups.

Primarily figure 4.5 shows that the mean scores cover the entire range of the test scale, and the 95% confidence intervals are narrower than 10 points (10%) of the scale.

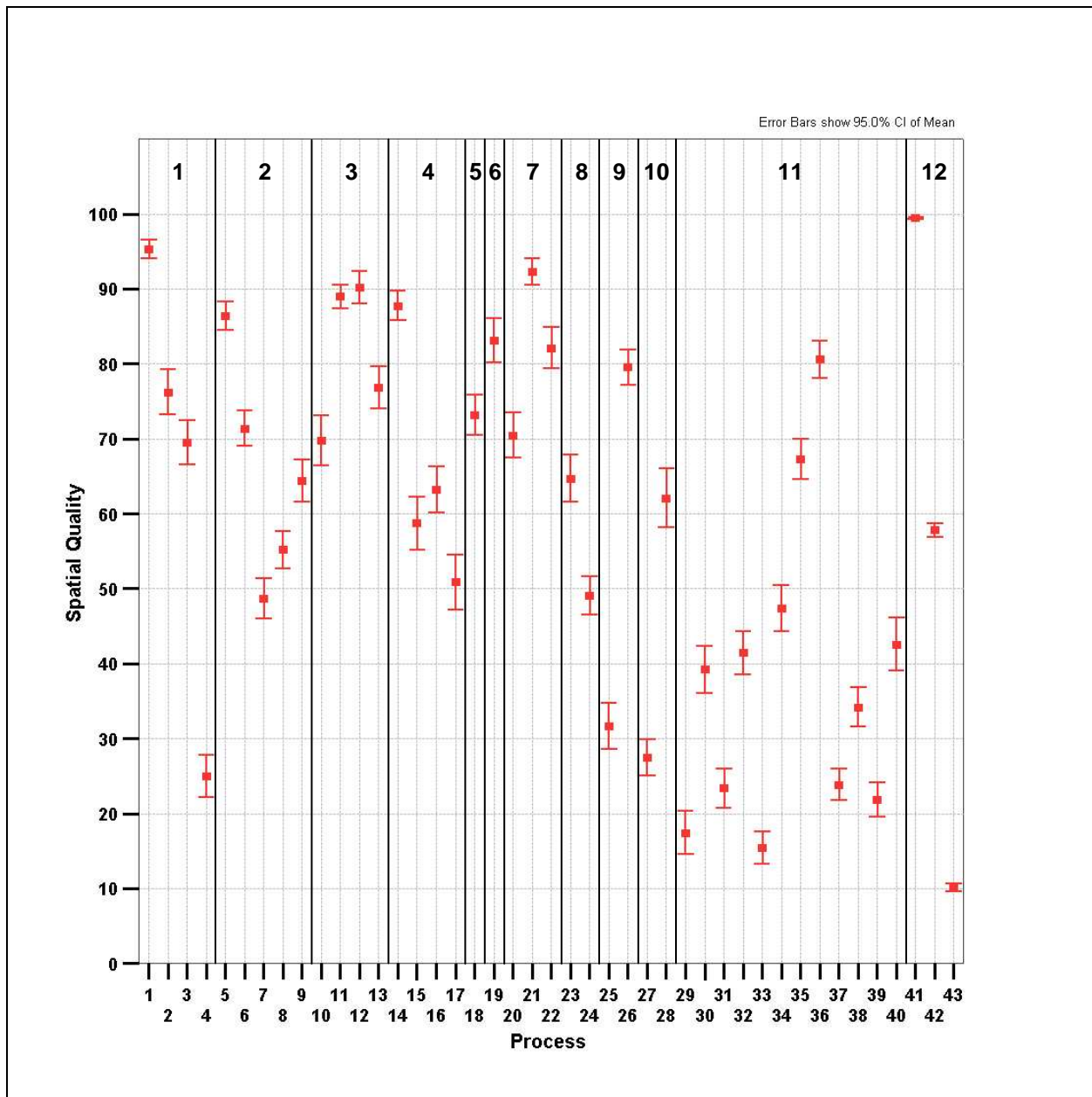


Fig 4.5 Means and 95% confidence intervals averaged across program item and listening position.

To briefly summarise the results in figure 4.5:- Observing first the anchor recordings (group 12); the high anchor (process 41) was scored at the top of the scale, the mid anchor (process 42) was scored around the centre and the low anchor (process 43) at the

bottom. The 3/1 downmix (process 1) created the least impairment of all processes. The largest impairments were created by combinations of processes (group 11). Groups 1-10 predominantly created less severe impairments. For example; 3.0

and 2.0 downmixes (processes 2 and 3). The majority of loudspeaker mis-placement (group 3) and missing channel (group 7) processes did not create large impairments. Only the lowest bit rate audio codecs created substantial impairments in group 2. Swapping L and R channels (process 14) created a greater impairment than other channel routing errors (group 4).

4.5. The effect of listener on spatial quality (Process*Listener)

The interaction of listener with process has the second largest effect on perceived spatial quality. This was first identified in the histograms in section 4.2 (Table 4.1) and suggests that there is no consensus between listeners for certain stimuli.

It might be necessary in these cases to investigate a method of listener segmentation, such as those already identified in table 4.1 and particularly in cases where a multi-modal distribution is observed. It may be possible to determine the reasons for the differences in opinions. The influence of listener could then be considered in the calibration of the QESTRAL model.

4.6. The influence of program item type on spatial quality (Process*ProgItem)

The interaction of programme item type with process was shown to have a significant effect on perceived spatial quality. This suggests that for the different programme items certain processes created a greater or lesser impairment. The scores for processes that demonstrate this are shown in figure 4.6 (page 10), as means and 95% confidence intervals averaged across listening position.

To highlight this two examples are given:-

1) For process 2 (3.0 downmix) a far smaller impairment was perceived of programme item 2 (classical) than of items 1 and 3. This is likely to be because the rear channels of program item 2 contain only reverberant information from the front image and downmixing them into the front channels in this instance was not perceived as overly degrading. This is different to programme items 1 and 3 whose rear channels contain clearly identifiable foreground sources.

2) With process 17 (channel routing error 4) the channel order had been randomised. This process was

perceived as creating a lesser impairment of programme item 1 than either 2 or 3. This could be because the majority of the channels in programme item 1 contain audience applause. This information is decorrelated and could be re-routed to different channels without significant impairment to the image (NB. the perceived impairment was possibly created by the re-routing of the channels which contain the commentators). However in the cases of programme items 2 and 3 re-routing the channels destroys the intended image.

The evidence above suggests that programme item type should be considered in the calibration of the QESTRAL model.

4.7. The influence of listening position on spatial quality (Process*LP)

The interaction of listening position with process was shown to have an effect on perceived spatial quality. This suggests that between the two listening positions certain processes created a greater or lesser impairment in perceived spatial quality. The processes that demonstrate this are shown in figure 4.7 (page 11), as means and 95% confidence intervals averaged across program item.

To highlight this two examples are given:-

1) In process 21 (channel missing 2) the rear left loudspeaker (Ls) is missing. From listening position 2 (1m to the right of centre) this was perceived as less of an impairment than from listening position 1 (central position). This could be because the increased distance from Ls at listening position 2, its removal is masked.

2) In the case of process 12 (loudspeaker misplacement 3) the rear loudspeakers have been misplaced to -90° and 90° respectively. From listening position 1, possibly due to the cone of confusion, this did not create a perceivably large impairment, as shown by the high score. Whereas from listening position 2 which is substantially closer to the rear right loudspeakers, the misplacement is much more obvious and therefore the impairment becomes apparent and is therefore scored lower.

The evidence above suggests that listening position should be considered in the calibration of the QESTRAL model.

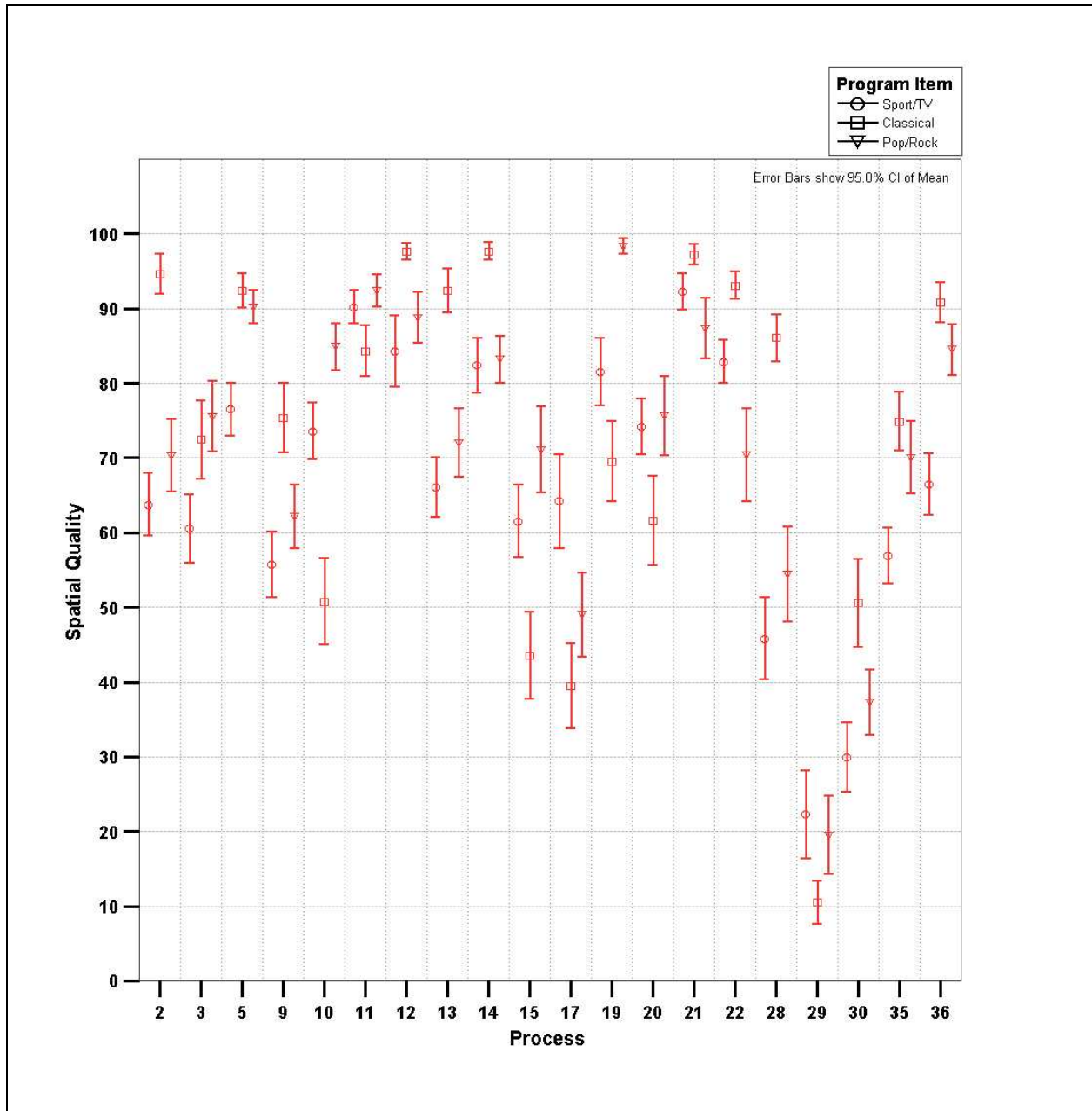


Fig 4.6 Processes whose scores vary depending upon program item displayed as means and 95% confidence intervals averaged across listening position.

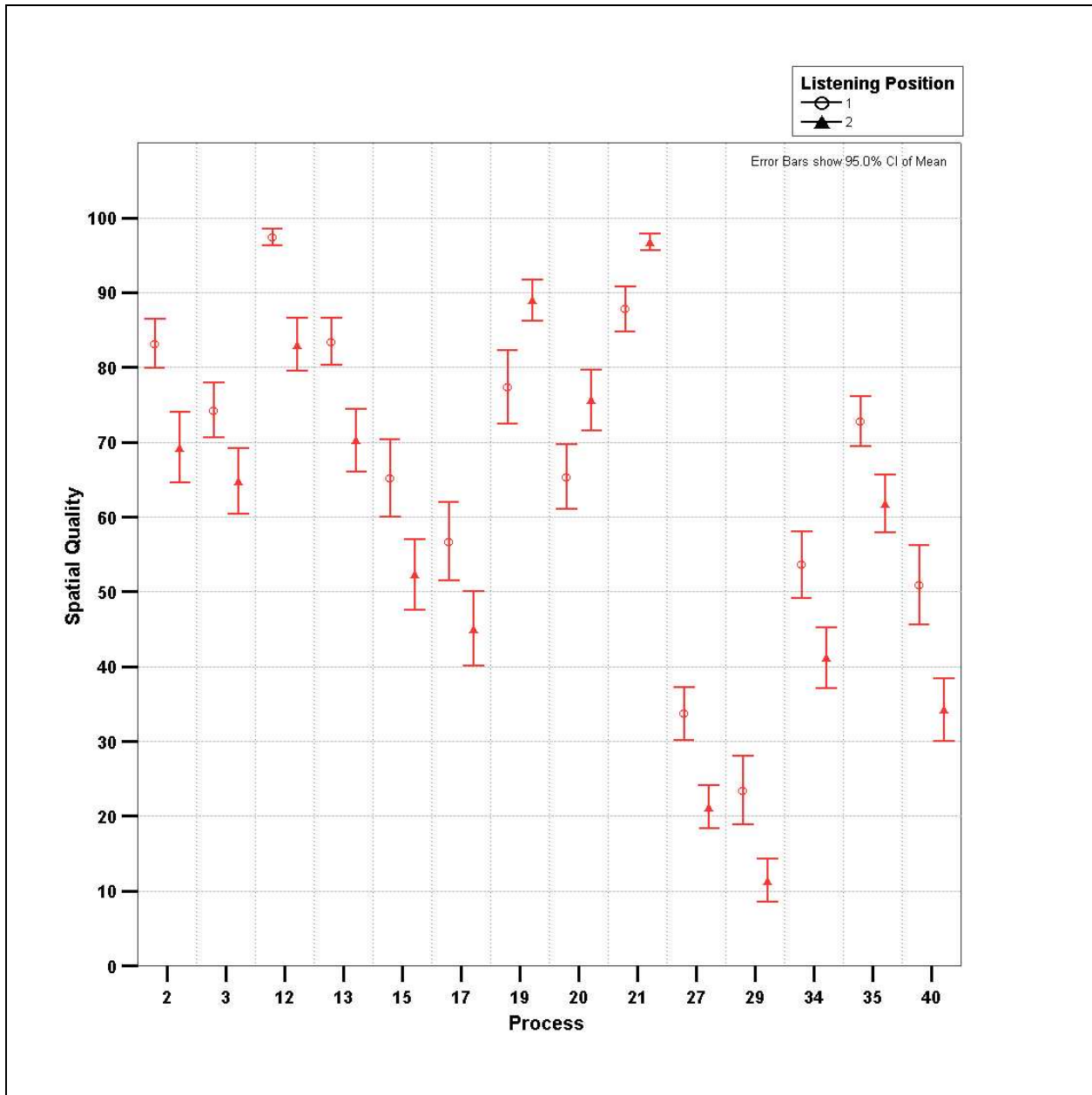


Fig 4.7 Processes whose scores vary depending upon listening position displayed as means and 95% confidence intervals averaged across program item.

5. CONCLUSIONS

This paper has described the development and implementation of a listening experiment investigating the perception of impairments to spatial quality. Results from these experiments have been presented and discussed in relation to the calibration of the QESTRAL model (a perceptual model that aims to predict changes to spatial quality of service (SQoS) between the soundfield reproduced by a reference system and that of an impaired version of the reference system).

Evaluation of the results indicates that programme item type and listening position have a significant effect on perceived spatial quality. This suggests that these factors should be considered in the calibration of the QESTRAL model.

The wide and multi-modal data distributions observed in section 4.2 (Table 4.1) suggested that listeners found it very difficult to assess some of the processes. The mean scores from these cases are considered to be ambiguous, and should be excluded from the database used in the calibration of the QESTRAL model. It is not yet clear whether a unitary concept of spatial quality can be defined and understood sufficiently well to enable a group of experienced listeners to make reliable and consistent judgements. It may also be necessary to consider the subdivision of listeners into groups representing different populations in the calibration of the QESTRAL model. However further work is required to determine how this can be done,

6. ACKNOWLEDGEMENTS

This research was completed as a part of QESTRAL Project (Engineering and Physical Sciences Research Council EP/D041244/1) in collaboration with University of Surrey, UK, Bang & Olufsen, Denmark and BBC Research, UK.
www.surrey.ac.uk/soundrec/QESTRAL

7. REFERENCES

ITU-Recommendation BS.775-1 (1992-1994) *Multichannel stereophonic sound system with and without accompanying picture*. International Telecommunication Union recommendation, Geneva.

ITU-Recommendation BS.1116-1 (1997) *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. International Telecommunication Union recommendation, Geneva.

ITU-Recommendation BS.1534 (2001) *Method for the subjective assessment of intermediate audio quality*. International Telecommunication Union recommendation, Geneva.

Rumsey, F. (1998) Subjective Assessment of the Spatial Attributes of Reproduced Sound. In *Proceedings of the AES 15th International Conference: Audio, Acoustics & Small Space.*, 31 Oct – 2 Nov 1998. Copenhagen, Denmark.

Rumsey, F. (2002) Spatial quality evaluation for reproduced sound: Terminology, meaning, and a Scene-Based Paradigm. *J. Audio Eng. Soc*, Vol.50 No.9, pp. 651-666.

Zielinski, S. Rumsey, F. and Bech, S. (2008) On Some Biases Encountered in Modern Audio Quality Listening Tests – A Review. *J. Audio Eng. Soc*, Vol. 56 No. 6, pp.427-451.

8. APPENDICES

8.1. Appendix A

Results from section 2

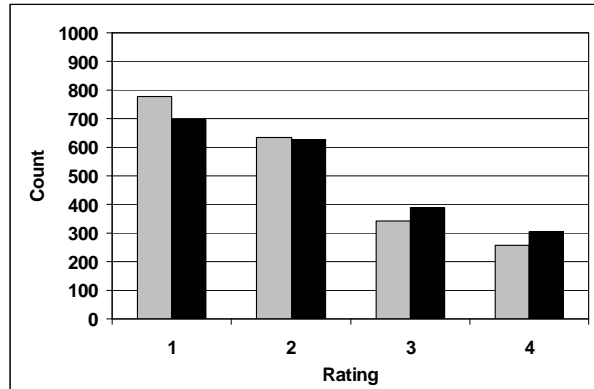


Fig A1. Histograms illustrating an overview of all responses after initial investigation (grey) and after optimization (black) of process selection.

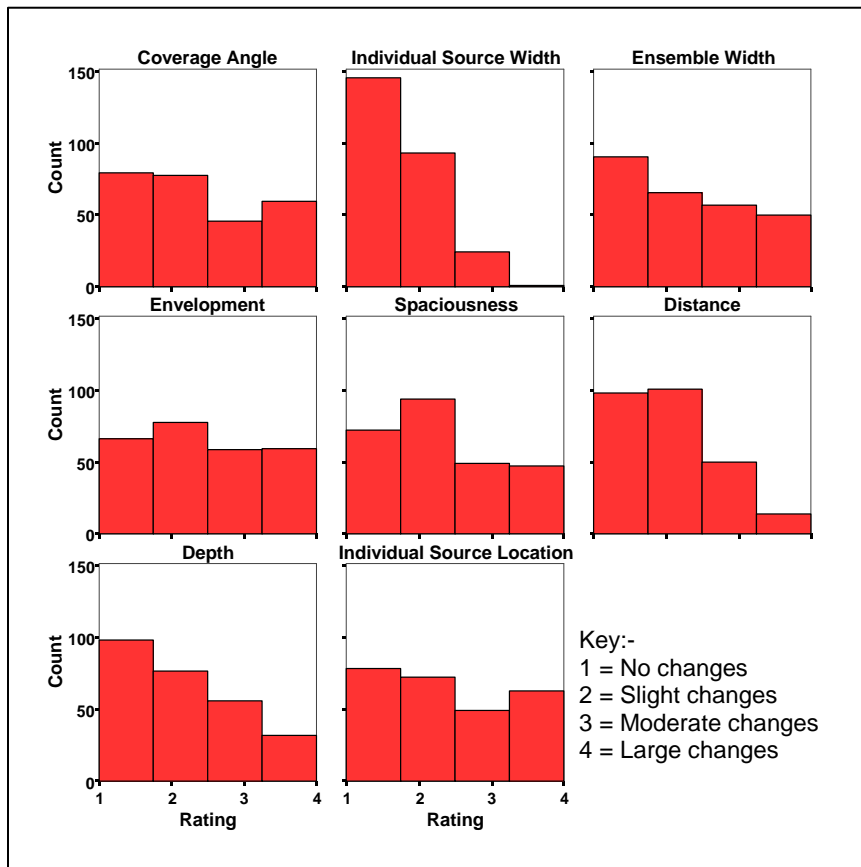


Fig A2. Histograms illustrating an overview of all responses for each spatial attribute.

8.2. Appendix B

Table B1. shows the final list of processes used in the listening tests assessing spatial quality .

No.	Process	Description	Process Type
1	Downmix 1	3/1: L = L, R = R, C = C, S = 0.7071* <i>Ls</i> + 0.7071* <i>Rs</i> .	1
2	Downmix 2	3.0: L = L + 0.7071* <i>Ls</i> , R = R + 0.7071* <i>Rs</i> , C = C.	
3	Downmix 3	2.0: L = L + 0.7071* <i>C</i> + 0.7071* <i>Ls</i> , R = R + 0.7071* <i>C</i> + 0.7071* <i>Rs</i> .	
4	Downmix 4	1.0: C = 0.7071* <i>L</i> + 0.7071* <i>R</i> + C + 0.5* <i>Ls</i> + 0.5* <i>Rs</i> .	
5	Codec A	160kbs	2
6	Codec B	64kbs	
7	Codec C	64kbs	
8	Cascaded codec A	2 stage cascade (80kbs)	
9	Cascaded codec B	4 stage cascade (64kbs)	3
10	Loudspeaker mis-placement 1	L and R re-positioned at -10° and 10°	
11	Loudspeaker mis-placement 2	C is skewed; re-positioned at 20°	
12	Loudspeaker mis-placement 3	<i>Ls</i> and <i>Rs</i> re-positioned at -90° and 90°	
13	Loudspeaker mis-placement 4	<i>Ls</i> and <i>Rs</i> re-positioned at -170° and 160°	4
14	CH routing error 1	L and R swapped	
15	CH routing error 2	L and R swapped for <i>Ls</i> and <i>Rs</i>	
16	CH routing error 3	CH order rotated	
17	CH routing error 4	CH order randomised	5
18	Inter-channel level mis-alignment	L, C and R -6dB quieter than <i>Ls</i> and <i>Rs</i>	
19	Inter-channel out-of-phase	C 180° out-of-phase	6
20	Missing channel 1	R removed	7
21	Missing channel 2	<i>Ls</i> removed	
22	Missing channel 3	C removed	
23	Filtering 1	500Hz HPF on all channels	8
24	Filtering 2	3.5kHz LPF on all channels	
25	Inter-channel crosstalk 1	1.0 downmix in all CH	9
26	Inter-channel crosstalk 2	Partly correlated (0.5 bleed in adjacent channels)	
27	Virtual surround algorithms 1	Line array virtual surround	10
28	Virtual surround algorithms 2	2 CH virtual surround	
29	Combination 1	CH routing error 4 + Missing channel 1, 2 and 3	11
30	Combination 2	Downmix 2 + Missing channel 1	
31	Combination 3	Downmix 3 + CH routing error 4	
32	Combination 4	Downmix 3 + Loudspeaker miss-placement 1	
33	Combination 5	Downmix 4 + Filtering 1	
34	Combination 6	Loudspeaker miss-placement 4 + Loudspeaker miss-placement 1	
35	Combination 7	Codec A + Downmix 3	
36	Combination 8	Codec A + Loudspeaker miss-placement 3	
37	Combination 9	Codec C + Downmix 4	
38	Combination 10	Codec C + CH routing error 4	
39	Combination 11	Virtual surround algorithms 2 + Missing channel 1	
40	Combination 12	Virtual surround algorithms 2 + Loudspeaker miss-placement 1	

Table B1. List of processes used for the spatial quality listening experiments.

8.3. Appendix C

Tables C1-4 list of processes selected for listening sessions 1-4.

No.	Process	Description	Process Type
3	Downmix 3	$2.0: L = L + 0.7071 * C + 0.7071 * Ls, R = R + 0.7071 * C + 0.7071 * Rs.$	1
4	Downmix 4	$1.0: C = 0.7071 * L + 0.7071 * R + C + 0.5 * Ls + 0.5 * Rs.$	
11	Loudspeaker misplacement 2	C is skewed; re-positioned at 20°	3
12	Loudspeaker misplacement 3	Ls and Rs re-positioned at -90° and 90°	
16	CH routing error 3	CH order rotated	4
17	CH routing error 4	CH order randomised	
20	Missing channel 1	R removed	7
25	Inter-channel crosstalk 1	1.0 downmix in all CH	9
30	Combination 2	Downmix 2 + Missing channel 1	11
33	Combination 5	Downmix 4 + Filtering 1	

Table C1. Processes selected for test 1.

No.	Process	Description	Process Type
2	Downmix 2	$3.0: L = L + 0.7071 * Ls, R = R + 0.7071 * Rs, C = C.$	1
10	Loudspeaker misplacement 1	L and R re-positioned at -10° and 10°	3
13	Loudspeaker misplacement 4	Ls and Rs re-positioned at -170° and 160°	
15	CH routing error 2	L and R swapped for Ls and Rs	4
18	Inter-channel level misalignment	L, C and R -6dB quieter than Ls and Rs	5
21	Missing channel 2	Ls removed	7
22	Missing channel 3	C removed	
31	Combination 3	Downmix 3 + CH routing error 4	11
32	Combination 4	Downmix 3 + Loudspeaker miss-placement 1	
34	Combination 6	Loudspeaker miss-placement 4 + Loudspeaker miss-placement 1	

Table C2. Processes selected for test 2.

No.	Process	Description	Process Type
1	Downmix 1	$3/1: L = L, R = R, C = C, S = 0.7071 * Ls + 0.7071 * Rs.$	1
8	Cascaded codec A	2 stage cascade (80kbs)	2
9	Cascaded codec B	4 stage cascade (64kbs)	
14	CH routing error 1	L and R reversed	4
26	Inter-channel crosstalk 2	Partly correlated (0.5 bleed in adjacent channels)	9
27	Virtual surround algorithms 1	Line array virtual surround	10
28	Virtual surround algorithms 2	2 CH virtual surround	
29	Combination 1	CH routing error 4 + Missing channel 1, 2 and 3	11
39	Combination 11	Virtual surround algorithms 2 + Missing channel 1	
40	Combination 12	Virtual surround algorithms 2 + Loudspeaker miss-placement 1	

Table C3. Processes selected for test 3.

No.	Process	Description	Process Type
5	Codec A	160kbs	2
6	Codec B	64kbs	
7	Codec C	64kbs	
19	Inter-channel out-of-phase	C 180° out-of-phase	6
23	Filtering 1	500Hz HPF on all channels	8
24	Filtering 2	3.5kHz LPF on all channels	
35	Combination 7	Codec A + Downmix 3	11
36	Combination 8	Codec A + Loudspeaker miss-placement 3	
37	Combination 9	Codec C + Downmix 4	
38	Combination 10	Codec C + CH routing error 4	

Table C4. Processes selected for test 4.

8.4. Appendix D

A flowchart illustrating a listeners path through tests 1 and 2 is given in figure D1. This process was repeated for tests 3 and 4.

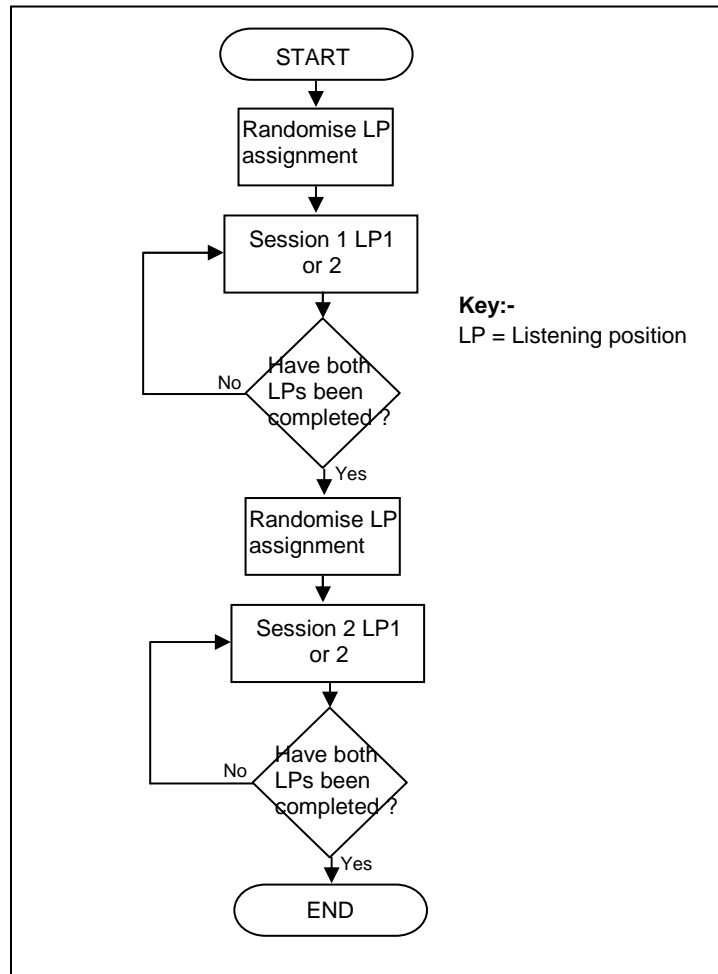


Fig D1. Flowchart illustrating a listener's path through tests 1 and 2 of the Spatial Quality experiment.

8.5. Appendix E

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.

Spatial Quality

page 1 of 6

Reference

1 2 3 4 5 6 7 8

Worse

Save/Next

37 86 4 100 46 67 22 75

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 (e.g. RCa and RCb).

Once you are happy with your responses click the save/next button to continue to the next page (NB. You'll 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.

Questionnaire

After you have completed the experiments there is a short questionnaire.

***Please note that for experimental accuracy it is important that you remain facing forward and refrain from moving your head while rating the stimuli**

****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 consistency and accuracy of your judgements is crucial to the success of the test. Please do not commence the experiment unless you feel confident in the task. Additionally if you are suffering from fatigue during the test please ask the test supervisor for a break.**

*******If you have any questions please ask the test supervisor.**