Towards the automatic assessment of spatial quality in the reproduced sound environment

Robert Conetta

Submitted for the Degree of Doctor of Philosophy

Institute of Sound Recording
Faculty of Arts and Human Sciences
University of Surrey

2011

This thesis and the work to which it refers are the results of my own efforts. Any ideas, data, images or text resulting from the work of others (whether published or unpublished) are fully identified as such within the work and attributed to their originator in the text, bibliography or in footnotes. This thesis has not been submitted in whole or in part for any other academic degree or professional qualification. I agree that the University has the right to submit my work to the plagiarism detection service TurnitinUK for originality checks. Whether or not drafts have been so-assessed, the University reserves the right to require an electronic version of the final document (as submitted) for assessment as above.

Abstract

The research in this thesis describes the creation and development of a method for the prediction of perceived spatial quality. The QESTRAL[1] model is an objective evaluation model capable of accurately predicting changes to perceived spatial quality. It uses probe signals and a set of objective metrics to measure changes to low-level spatial attributes. A polynomial weighting function derived from regression analysis is used to predict data from listening tests, which employed spatial audio processes (SAPs) proven to stress those low-level attributes.

A listening test method was developed for collecting listener judgements of impairments to spatial quality. This involved the creation of a novel test interface to reduce the biases inherent in other similar audio quality assessment tests. Pilot studies were undertaken which established the suitability of the method.

Two large scale listening tests were conducted using 31 Tonmeister students from the Institute of Sound Recording (IoSR), University of Surrey. These tests evaluated 48 different SAPs, typically encountered in consumer sound reproduction equipment, when applied to 6 types of programme material. The tests were conducted at two listening positions to determine how perceived spatial quality was changed.

Analysis of the data collected from these listening tests showed that the SAPs created a diverse range of judgements that spanned the range of the spatial quality test scale and that listening position, programme material type and listener each had a statistically significant influence upon perceived spatial quality. These factors were incorporated into a database of 308 responses used to calibrate the model.

The model was calibrated using partial least-squares regression using target specifications similar to those of audio quality models created by other researchers. This resulted in five objective metrics being selected for use in the model. A method of post correction using an exponential equation was used to reduce non-linearity in the predicted results, thought to be caused by the inability of some metrics to scrutinise the highest quality SAPs. The resulting model had a correlation (r) of 0.89 and an error (RMSE) of 11.06% and performs similarly to models developed by other researchers. Statistical analysis also indicated that the model would generalise to a larger population of listeners.

1 Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener

Contents

List of figures

List of tables

Glossary of terms

Beta values – describe the importance of an independent variable (objective metric) in a regression model. If the magnitude of Beta value is high, the variable has high importance in the model. The polarity of the Beta value indicates the independent variables relationship to the dependent variable.

Correlation (r) – The correlation coefficient is the measure that is used to represent the strength of a linear relationship between two variables. In the context of this project, the two variables are the measured (dependent variable) scores obtained from listening tests and predicted scores obtained from the regression model. It is calculated using the following equation:

$$Correlation(R) = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2}\sqrt{\Sigma(Y_i - \overline{Y})^2}}$$

where X_i denotes the mean subjective score for each stimulus, Y_i the predicted scores and N represents the total number of stimuli. X and Y represent the average value of measured (dependent variable) scores and predicted scores respectively.

Entropy – can be defined as the lack of order or predictability or the degree of disorder or randomness.

F-B (Foreground – Background) – describes a recording in which the front channels reproduce predominant foreground audio content, whereas rear channels contain only background audio content (ambient, reverberant sounds, unclear, "foggy"). Many 5-channel classical music recordings use this spatial mixing style or scene type.

F-F (Foreground – Foreground) – describes a recording in which both front and rear channels contain predominant foreground audio content (mainly close and clearly perceived audio sources). Many 5-channel pop music recordings use this spatial mixing style or scene type.

Inter-aural cross-correlation (IACC) – is based upon the normalised cross-correlation (NCC) function and is a measure of similarity between two signals (x and y) over a period of time, t1-t2 with an offset, τ. IACC measures the similarity of two binaural signals recorded using a binaural simulator, and is calculated as the maximum absolute value of the NCC function.

$$NCC(\tau) = \frac{\int_{t1}^{t2} x(t)y(t+\tau)dt}{\left[\int_{t1}^{t2} x^2(t)dt \int_{t1}^{t2} y^2(t)dt\right]^{\frac{1}{2}}}$$

$$IACC = \left|NCC(\tau)\right| \max$$

$$\text{For } -1ms < \tau < +1ms$$

Karhunen-Lòeve Transform (KLT) – an extension of principal component analysis (PCA), is a linear transform which can be used to statistically analyse the co-variance between audio channels in a multichannel recording. This is achieved by transposing the audio channels into eigen-channels each containing co-varying audio. The eigen-channels are ordered hierarchically; the first being the most statistically important and containing the largest portion of co-varying audio. The statistical contribution each makes to the original audio is indicated by its co-variance value, for example if all audio channels of a 5-channel recording are correlated this will be transposed to a single eigen-channel with a co-variance value of 1, alternatively if the channels are completely uncorrelated it will be transposed to five eigen-channels with a co-variance value of 0. In broadcast applications these eigen-channels are transmitted with several coefficients so that the receiver can then rebuild the audio accurately.

Lateral Fraction (LF) – is a measure of spatial impression and is defined as the ratio of early sound energy arriving laterally over sound energy arriving from all directions.

Loading plot – is a radial plot of loading vectors associated with two principle components (PCs). A loading vector is considered to be the bridge between the variable space and principle component space. A loading plot illustrates the importance that each independent variable (objective metric) in the regression model contributes to each PC. The further the independent variable is from the centre of the plot the greater its importance.

Principal Component Analysis (PCA) – is a multivariate technique for identifying the linear components of a set of variables.

Root Mean Square Error (RMSE) – is a measure of how much the measured (dependent) and predicted scores differ. It is expressed in the same units as the dependent series.

Regression line – is the line of best-fit drawn on a scatter-plot illustrating the measured (dependent variable) vs. predicted scores of a regression model.

Spectral centroid (fc) – is the center of gravity of the frequency spectrum, and has been used as a correlate of brightness of musical instruments.

Spectral rolloff – is the point on frequency spectra at which 95% of the total energy achieved, it can be considered as a representation of upper cut-off frequency of the signal and hence a measure of the bandwidth of audio signal (assuming that the lower cut-off frequency is constant).

Target line – is the line drawn on a scatter-plot illustrating the measured (dependent variable) vs. predicted scores of a regression model which represents the ideal relationship (ie. Y = X).

Variance plot – is the histogram of variances associated with PCs obtained from PCA or PLS regression analysis.

Acknowledgements

The completion of this research project would not have been possible without the help of a number of different people. Thank you to: Francis Rumsey and Slawek Zielinski for their thoughtful and patient supervision and encouragement; Martin Dewhirst for his friendship, ideas and excellence with Matlab; Philip Jackson, Soren Bech and David Meares for challenging my conclusions and for their guidance; my fellow students Kathy, Chris, Laurent, Will, Sunish, Paolo, Ryan, Joey, Daisuke, Duncan and Raf for their empathy; Russell and David for allowing me to pick their brains; Eddie, Alan and Bill for technical support; Mum, Dad, Katie, James, Nan and Grandad, Sam, Scott and all my family and friends, for all their support and motivation during my studies; Bridget Shield for being understanding and supportive.

I would also particularly like to thank Tim Brookes, without whose supervision and encouragement, the completion of this thesis would not have been possible.

Chapter 1 – Introduction

Letowski [1989] proposed that sound quality evaluation could be divided into two distinct domains of perception, the timbral domain and the spatial domain. In this paradigm Letowski suggests that timbral quality concerns the perception of the spectral characteristics of the sound whereas spatial quality concerns the perception of what he terms spaciousness or the spatial characteristics. Recent research [Rumsey *et al*, 2005] has shown that spatial quality accounts for as much as 30% of overall audio quality.

With the continuing advancement of audio technology the desire exists to create or reproduce increasingly real and immersive soundfields or listening experiences [Rumsey, 2001][Soulodre *et al*, 2003b][Davis, 2003]. Manufacturers and service providers in both the entertainment and Information and Communication Technology (ICT) industries are now attempting to deliver spatially enhanced multi-channel audio scenes. This can be observed in the function of the consumer products available in the modern market place; for example, surround sound 'home-cinema' systems, DVD Video and Audio appliances, and gaming consoles [Rumsey, 2001][Soulodre *et al*, 2003b]. Mobile devices such as MP3 players, mobile phones and personal digital assistants (PDAs) are also now becoming increasingly more important in modern life, and have the potential to deliver binaurally enhanced spatially immersive environments to the user via a pair of earphones/headphones [Rumsey, 2002]. Broadcasters such as the British Broadcasting Corporation (BBC) and British Sky Broadcasting (BSkyB) also now have the capability to deliver spatially enhanced multi-channel audio scenes in the form of matrixed 5.1 surround sound via their high definition (HD) television broadcasts [BBC, 2009][BSkyB Ltd, 2009]. The potential of these new technologies and developments motivates a requirement from a technological point of view for audio of a high spatial quality to reach the end user.

In many of these developments the delivery format and rendering (reproduction) format are separate. This aids versatility allowing the content to be delivered in a format that suits the transmission technology (e.g. HD broadcast, DVD) whilst remaining potentially re-playable over many different reproduction formats or audio systems. In practice this means that the audio content can be delivered using a wide variety of different formats and also reproduced over a wide variety of different audio systems. There are, for example, a wide variety of multichannel audio coding schemes used throughout the audio industry which seek efficiency by reducing the amount of data occupied by audio content in a delivery system. These multichannel audio codecs have been shown to have a detrimental effect on the perceived spatial quality when reproduced using an audio system [Marins *et al*, 2008]. This is particularly apparent in the most band-limited delivery conditions such as online streaming or basic rendering devices such as mobile phones and MP3 players, where storage space is

at a premium. Also with various upmixing and downmixing techniques [ITU-R BS.775, 1992-1994][Zielinski *et al*, 2003b] used widely in the industry and the potentially unlimited number of non-standard changes made by the consumer or system developers to loudspeaker locations of numerous reproduction formats, the possible resulting degradations to spatial quality are many. These could include changes in source-related attributes such as perceived location, width, distance and stability; and changes in environment-related attributes such as envelopment and spaciousness [Rumsey, 2002]. Therefore with the above in mind it is clear that a method for assessing perceived spatial quality would be useful in the future as a research and development tool.

Although a possible assessment method could take the form of formal subjective tests, the time and monetary costs of maintaining a listening panel and running listening tests are substantial [Bech and Zacharov, 2006], making this solution not ideal and not always practical. Another possibility is to develop an objective evaluation system which, while not completely replacing subjective testing, could at least be used to provide an initial approximation of perceptual scores. .There is a current model for evaluating perceived sound quality which was created by the International Telecommunication Union (ITU), known as PEAQ [ITU-R BS.1387, 2001], however it does not account for the contribution of spatial quality to the overall user experience, concentrating instead on impairments to timbral quality such as audio coding distortions, noise and bandwidth reductions. Therefore a model capable of objectively evaluating spatial quality would potentially make a valid contribution to this existing ITU standard and also prove valuable for product and service development.

1.1 The QESTRAL project

The QESTRAL (Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener) project utilises the skills of a multidisciplinary collaboration between the Institute of Sound Recording (IoSR), and the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey, with support and expertise from two industrial partners, Bang & Olufsen, Denmark and BBC Research and Development, UK. The project is funded by an Engineering and Physical Sciences Research Council (EPSRC) grant (EP/D041244/1) [Rumsey *et al*, 2005c].

The aim of the project is to develop an artificial listener or objective evaluation model capable of predicting perceived spatial quality. Similarly to PEAQ, the model will employ an intrusive method of evaluation based upon measured comparisons between the soundfield reproduced by a reference system and a version of the reference system impaired by a spatial audio process (SAP) (e.g. downmixing, multichannel audio codec, loudspeaker misplacements etc). The QESTRAL model architecture is illustrated in figure 1.1.

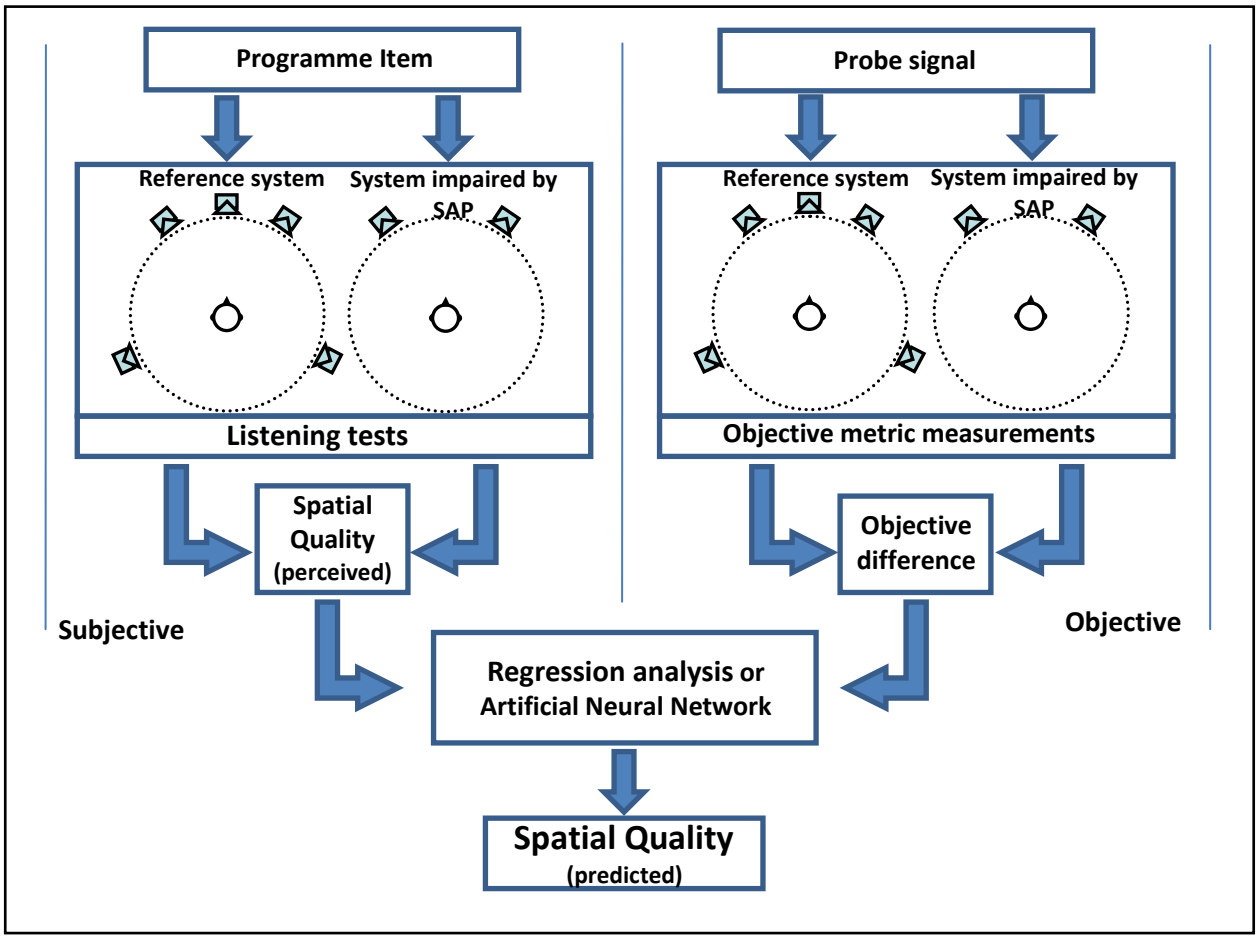


Fig 1.1 QESTRAL model architecture.

The model will be a computational model which renders a probe signal(s) for both the reference system (in figure 1.1 3/2-stereo) and the SAP version (in figure 1.1 2-channel stereo downmix) in a virtual environment. Physical characteristics of both rendered soundfields are extracted by measurement, from the listening position, using a set of specially designed objective metrics. The measurements taken from the reference soundfield and SAP soundfield are then compared, and using a regression model or an artificial neural network (ANN), calibrated from the results of listening tests, a prediction of the perceived spatial quality calculated.

As discussed in the previous section audio content delivery formats and reproduction formats can be independent, meaning that the content can be replayed over a number of different audio systems. The range of different audio reproduction formats is wide, from portable handheld devices such as mobile phones to wavefield synthesis and each will potentially alter the spatial quality of the original content from that intended by the broadcaster or mix engineer. Furthermore an additional change in the quality can result if the loudspeakers in these systems are not arranged correctly. However identifying the changes to spatial quality created by different audio systems and loudspeaker arrangements would not be possible solely by measuring the electrical signal in each channel. So to allow the QESTRAL model to be reproduction format independent its measurements of the

reproduced soundfield are based on binaural and other microphone-derived signals received at the listening position.

The QESTRAL model is designed to employ probe signals. The probe signals must be suitable for scrutiny of aspects of the spatial scene. Probe signals have been shown to work successfully in similar applications [Mason, 2006][ITU-R BS.1387, 2001] and have the added advantage that their content can be controlled to allow changes created by the SAP to be detected and measured precisely.

One of the QESTRAL model's unique functions, developed by Dewhirst and based upon a previous publication [Dewhirst *et al*, 2005], allows it to potentially evaluate the reproduced soundfield at a number of different listening positions across the listening area. This could be a useful tool for audio system designers and researchers wishing to determine the extent to which the spatial quality created by the SAP changes across the listening area or for determining optimum listening positions within a system.

1.2 The development of the QESTRAL model

The development of the QESTRAL model will utilise a multidisciplinary collaboration between different researchers and engineers. The contributions of this author to the development of the model are described in this thesis. Where relevant, contributions by other project members will also be discussed.

1.2.1 Work packages required to develop the QESTRAL model

The work required to create the QESTRAL model can be divided into a number of work packages; these are described in table 1.1. The work packages contributed to by this author were:

1. ***Design and implementation of listening tests to evaluate the effect of a wide range of SAPs on the spatial quality, at two listening positions, of a selection of 5-channel programme items***

   The QESTRAL model will be calibrated using detailed statistical analysis of subjective scores collected from listening tests. To make the model generalisable, extensive listening tests will be undertaken to collect data on a wide range of SAPs. A selection of 5-channel audio recordings (or programme items) representing different types of typical multichannel material (including excerpts of music, TV broadcasts and Film scenes) will be used as reference recordings to which the SAPs will be applied. To calibrate the model so that it can be used to predict spatial quality at two different listening positions the tests will be carried out at a listening position in the centre of the loudspeaker array and at a listening position one metre to the right of the central position.

2. ***Development of objective metrics to measure the spatial attributes of reproduced sound from binaural/microphone signals situated at the listening position***

The QESTRAL model will be built using objective metrics that are capable of measuring changes, created by the SAPs, to the spatial characteristics of the programme items. The objective metrics will be developed based upon metrics developed by other researchers or members of the QESTRAL project team.

5. ***Calibration of the QESTRAL model – predicting the subjective scores using the objective model***

In the calibration of the QESTRAL model, the objective metrics will be fitted to the subjective scores through appropriate weighting and combination of metrics for greatest error and lowest error, using statistical regression. Although it is possible to create perceptual models using artificial neural networks, the author has access to and greater experience with regression analysis, so this will be employed.

Package	Description	Work presented or published
1	Design and implementation of listening tests to evaluate the effect of a wide range of SAPs on the spatial quality, at two listening positions, of a selection of 5-channel programme items	This thesis, Conetta *et al* (2008)
2	Development of objective metrics to measure the spatial attributes of reproduced sound from binaural/microphone signals situated at the listening position	This thesis, Jackson *et al* (2008)
3	Development of probe signals to simulate the generic characteristics of programme items	Jackson *et al* (2008)
4	Development and creation of the objective model architecture – including the modelling of SAPs, simulation of listening environment, implementation of probe signals, and coding of objective metrics	Dewhirst (2008), Jackson *et al* (2008)
5	Calibration of the QESTRAL model – predicting the subjective scores using the objective model	This thesis (NB. Early calibrations can be found in Dewhirst *et al*, 2008 and Conetta *et al*, 2008)

Table 1.1 Work packages for the development of the QESTRAL model.

1.2.2 The specific aims of this research project and the organisation of this thesis

In fulfilling work packages 1, 2 and 5, the main aim of this research project is to establish a method by which spatial quality can be predicted. This aim can be broken down into several smaller aims:

(i) define spatial quality for this research project,

(ii) define suitable performance criteria for the QESTRAL model,

(iii) identify a suitable method for the development of the QESTRAL model,

(iv) identify a suitable test environment (i.e. reference reproduction system),

(v) identify appropriate objective metrics for spatial quality,

(vi) design a listening test method to obtain the required subjective data,

(vii) collate subjective data,

(viii) calibrate the QESTRAL model for the prediction of spatial quality.

An introduction to each chapter is given below with the specific aims of each described.

Chapter 2 satisfies aim (i) by identifying what is meant by spatial quality and defines it specifically for the reproduced sound environment. This definition will be used throughout this research project and will form the context under which the aims of this thesis will be achieved. This chapter also reviews current objective models for sound quality, in order to identify novel areas for investigation and to answer aim (ii), determine acceptable performance criteria for the QESTRAL model. The specific aims are:

- to define spatial quality for the reproduced sound environment,
- to identify current objective models for sound quality,
- to identify novel areas for investigation,
- to determine acceptable performance criteria for the QESTRAL model.

Chapter 3 identifies a suitable method for the development of the QESTRAL model, describing an appropriate research procedure that can be used to create the model. This answers aim (iii). Following this an overview of regression analysis techniques is given to identify the most suitable for the calibration of the QESTRAL model. To fulfil aim (ii) performance criteria are established for the calibrated QESTRAL model. Finally a review of different audio reproduction systems is given, which identifies the most appropriate system for this research project and answers aim (iv). The aims are:

- to determine a suitable method for QESTRAL model development,
- to identify the most appropriate regression analysis technique to calibrate the QESTRAL model,
- to identify calibration target specifications,
- to determine the most appropriate audio system for this research project (with which to calibrate the QESTRAL model).

Chapter 4 answers aim (v) by identifying and reviewing objective metrics currently used to measure individual spatial attributes in reproduced sound and in existing spatial quality models. The aim of this review is to identify suitable metrics that could be employed to measure changes to spatial quality that are created by the SAPs. These metrics could then be employed in the QESTRAL model to predict spatial quality. The aim is:

- to identify suitable metrics that could be used in the QESTRAL model to predict spatial quality.

Chapter 5, to answer aim (vi), identifies a suitable listening test method for evaluating a wide range of SAPs that impair the perception of spatial quality. This begins with an overview of existing international standards for the subjective assessment of audio quality, to determine their suitability. The aim is:

- to identify a likely candidate listening test method for characterisation of a wide range of SAP that impair the perception of spatial quality.

Chapter 6 discusses the development, implementation and results of four short listening tests, undertaken as pilot studies, prior to conducting a large scale listening test, for the purpose of confirming the suitability of the listening test method for the evaluation of perceived spatial quality and satisfying aim (vi). The aims are:

- to establish that the chosen method for subjectively assessing spatial quality is reliable and robust,
- to assess the difficulty of the task required of the listening test subjects at two listening positions using a wide range of different SAPs,
- to identify and investigate variables in the experiments that influence perceived spatial quality, and determine their relevance for calibrating of the QESTRAL model,
- to resolve any other issues that are identified as important for the development of the QESTRAL model.

Chapter 7 describes the implementation and results of two large scale listening tests to answer aim (vii). These tests will investigate the influence of a large number of SAPs, on a range of 5-channel programme items at two listening positions. The subjective scores collected from these experiments will be used to calibrate the QESTRAL model. The aims are:

- to determine the effects of a wide range of SAPs on perceived spatial quality at two listening positions,
- to establish how the collected subjective data should be treated for calibrating the QESTRAL model;
    - Determine which test variables should be included separately in the subjective database during the calibration process.
    - Identify the most reliable subjective data for the calibration.

Chapter 8 describes the calibration of the QESTRAL model for the prediction of spatial quality, answering aim (viii). A number of probe signals and relevant objective metrics will be introduced. The process of calibrating the model using regression analysis and the prediction results will be discussed. The context and limitations of the model will also be identified and discussed. The aims are:

- to establish if probe signals and objective metrics developed by the QESTRAL project team can be used to build a system that, after calibration against the listening test data from chapter 7, meets the target specifications proposed in chapter 3,
- to determine if the calibrated QESTRAL model is generalisable and performs within target specifications for the prediction of spatial quality for each of the test variables.

Chapter 9 collates the conclusions from each chapter, providing an overview of the achievements and the contributions of this research project to knowledge and suggestions for further work.

1.3 Summary and conclusions

The research in this thesis is motivated by the increasing importance of spatial audio and the lack of a perceptually-representative objective measure. Many manufacturers and service providers in both the entertainment and ICT industries are beginning to deliver spatially enhanced multi-channel audio scenes. This can be observed in various consumer products, mobile devices and the spatially enhanced multi-channel audio scenes delivered by the BBC and BSkyB via their high definition (HD) television broadcasts. The potential of these new technologies and developments motivates a requirement from a technological point of view for audio of a high spatial quality to reach the end user. In many of these developments the delivery format and rendering (reproduction) format are separate. This aids versatility but creates a wide range of potential impairments to the perceived spatial quality, created for example by multichannel audio codecs, upmixing and downmixing algorithms, and non-standard changes made by the consumer or system developers. Although there is currently a model for evaluating perceived sound quality, this concentrates on impairments to timbral quality and does not account for the contribution of spatial quality to the overall user experience. Therefore a model capable of spatial quality evaluation would potentially make a valid contribution to this existing ITU standard and may also be valuable for product and service development.

The QESTRAL project aims to provide a model capable of predicting perceived spatial quality. The model will be a computational model which renders a probe signal(s) for both the reference system and the SAP version in a virtual environment. Physical characteristics of both rendered soundfields will be extracted by measurement, from the listening position, using a set of specially designed objective metrics. The measurements taken from the reference soundfield and SAP soundfield will then be compared and, using a regression model, calibrated from the results of listening tests, a prediction of the perceived spatial quality calculated. This author's contribution to the QESTRAL project is to establish a method by which spatial quality can be predicted and include (i) defining spatial quality for this research, (ii) defining suitable performance criteria for the QESTRAL model, (iii) identifying a suitable method for the development of the QESTRAL model, (iv) identifying a suitable test environment (i.e. reference reproduction system), (v) identifying appropriate

objective metrics for spatial quality, (vi) designing a listening test method to obtain the required subjective data, (vii) collating subjective data, (viii) calibrating the QESTRAL model for the prediction of spatial quality. The remainder of this thesis documents these contributions.

Chapter 2 – Sound quality and spatial quality in the reproduced sound environment

In Chapter 1 the goals of this research project were identified and explained. This chapter concentrates on investigating and defining the term spatial quality for this research project. An introduction to sound quality and spatial quality is provided. This is followed by an overview of the spatial attributes present in the reproduced sound environment, after which a research definition for spatial quality is established. Current objective models for sound quality are reviewed in order to identify novel areas for investigation and also to determine acceptable performance criteria that the QESTRAL model should achieve.

2.1 Sound quality in the reproduced sound environment

An Oxford dictionary definition describes quality as "the standard of something as measured against other things of a similar kind" [Oxford University Press, 2010]. Gabrielsson and Lindström [1985] suggested that a judgement of quality in the reproduced sound environment is based on a judgement of two things, technical sound quality and perceptual (subjective) sound quality. A similar interpretation has been echoed in the work of Letowski [1989], who recognised a difference between sound quality and sound character, suggesting that sound character is a purely descriptive term, free from emotional response, similar to fidelity, whereas sound quality contains a hedonic judgement. Hence a judgement of sound quality could be described as a mixture of both sensory (non-hedonic) and affective (hedonic) judgements.

Technical or physical sound quality describes quality in terms of audio measurements such as the signal-to-noise ratio or distortion level, is judged against industry accepted quantitative levels of quality, and can be considered as non-hedonic or objective. Which is perhaps why this term relates more closely to fidelity; Gabrielsson and Lindström [1985] describe fidelity as the similarity between two sounds. By comparison, a judgement of perceptual (subjective) sound quality is influenced by liking or preference for one sound over another. This is a hedonic judgement of the sound quality and is likely to be context dependent, both in terms of a listener's overall taste and in terms of their learned or desired expectations of quality in a particular application. For example Rumsey *et al* [2005a] showed how experienced and naive listeners have different opinions of sound quality. Professor Jonathan Berger of Stanford University, California, reported that in a comparison of different audio delivery formats from low bit-rate encoded mp3 to compact disc, his recent music students showed a preference for mp3 [Dougherty, 2009]. In another example, Toole and Olive [1984] showed through a

comparison of blind versus sighted listening tests, that some listeners were biased in their opinion of loudspeaker sound quality by their expectation, based upon the loudspeaker's visual appearance.

There is agreement amongst a number of researchers that a judgement of sound quality involves the assessment of a number of attributes [Gabrielsson & Lindström, 1985][Letowski, 1989][Blauert & Jekosch, 1997]. Based upon this idea, Letowski proposed a hierarchical paradigm for these attributes. The MuRAL (Multidimensional auditoRy Assessment Language) (see Fig 2.1) is a hierarchical system which determines the relative importance of the different attributes considered by the listener when evaluating sound quality. Attributes which share the same circle of the system are treated as independent and complimentary whereas attributes closer to the centre are hierarchically more important. The lower level (closer to the edge) attributes are seen as purely sensory descriptive assessments, while the higher level attributes (toward the centre) are more likely to include affective assessments.

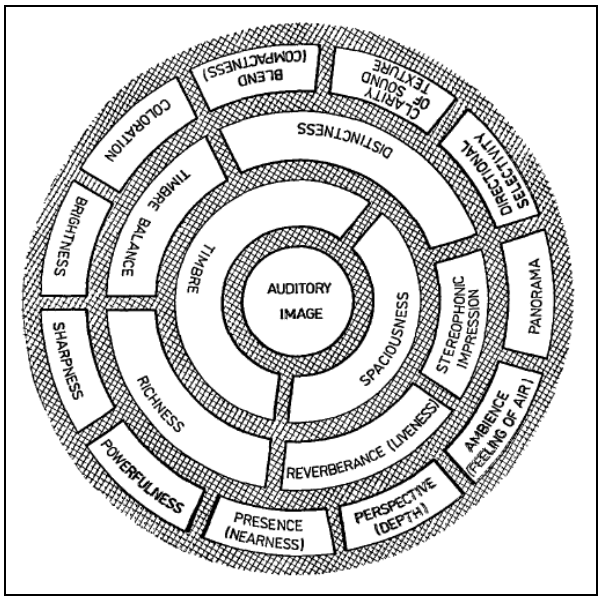


Fig 2.1 MuRAL Hierarchical system for parametric assessment of sound quality [Letowski, 1989].

Supporting this idea the International Telecommunication Union (ITU) describes 'basic audio quality' (BAQ), an attribute used to assess audio quality in its standards BS.1116-1 [1997] and BS.1534 [2001], as the global attribute used to judge any and all differences between the reference and stimulus under test.

2.1.1 A separate evaluation of spatial quality

An assessment of the spatial audio scene is considered as a part of the global evaluation of sound quality [Nakayama *et al*, 1971][Gabrielsson & Lindström, 1985][Letowski, 1989]. In fact Letowksi believes that the two main attributes of sound quality are timbre and spaciousness. He proposed that

listeners perceived two different factors when evaluating sound quality, suggesting that there were two domains to sound quality evaluation; timbral quality and spatial quality (Fig 2.2).

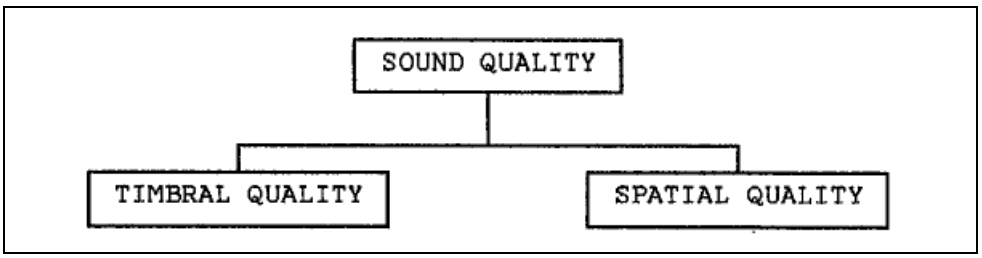


Fig 2.2 Letowski's domains of sound quality [Letowski, 1989].

This suggests that the two domains could be separated and assessed individually. Similarly ITU-R BS.1116-1 and BS.1534 indicate that, as well as BAQ, the test methods can be used to assess frontal image quality and surround spatial quality independently as additional attributes of multichannel surround sound quality (see Table 2.1).

Attribute	Definition
Frontal Image Quality	"This attribute is related to the localization of the frontal sound sources. It includes stereophonic image quality and losses of definition".
Impression of Surround Quality	"This attribute is related to spatial impression, ambience, or special directional surround effects".

Table 2.1 ITU multichannel surround sound quality attributes [BS.1116-1, 1997][BS.1534, 2001].

To further support this idea, Rumsey *et al* [2005] and Zielinski *et al* [2005b] employed a similar method to independently evaluate, the frontal spatial fidelity and surround spatial fidelity of a number of different 5-channel recordings. This suggests that it is possible to collect subjective data on spatial quality for the calibration of the QESTRAL model. However, Zielinksi *et al* [2005b] observed that when the programme material were downmixed, an audio process primarily considered to change the spatial fidelity, the listeners also perceived a change in the timbral fidelity, in addition to the change in spatial fidelity. Similarly when the programme material were bandwidth limited, the listeners perceived a change in the spatial fidelity in addition to the change in timbral fidelity. These observations indicate that the audio processes created some crossover (overlap) between the two domains. Letowski also suggests that when the two domains vary simultaneously our ability to evaluate the sound quality is limited. Therefore, it might be possible for listeners to become confused if a spatial audio process (SAP) causes a change in the quality across both domains. In a severe case (e.g. a very low bit-rate multichannel audio codec) this might result in the listener's opinion of the spatial quality being influenced by the perceived timbral quality. Unfortunately, in the context of this research project, it will not be possible to completely separate these two domains. So it will be important to establish the potential influence of changes to timbral quality, created by different SAPs, on a listener's opinion of spatial quality.

2.1.2 Sound quality: summary and conclusions

Quality is a comparative judgement whereby the standard of something is compared to other things like it. In reproduced sound, an opinion of sound quality is established through a combination of both sensory and affective judgements. This could alternatively be described as an assessment of the technical sound quality and the perceptual (subjective) sound quality.

Letowski proposed that listeners perceive and assess two different domains of sound quality. He identified these as timbral quality and spatial quality. The ITU also suggest that their audio quality assessment methods can be employed for the assessment of these two domains. Rumsey *et al* and Zielinski *et al* have employed this idea to investigate timbral and spatial fidelity as separate attributes. This suggests that it would be possible to collect subjective data on spatial quality for the calibration of the QESTRAL model. However Zielinski *et al* noted that the audio processes they used degraded both timbral fidelity and spatial fidelity. Letowski indicates that we have limited ability to evaluate quality when different domains vary simultaneously. Therefore it might be possible for listeners to become confused if a spatial audio process (SAP) causes a change in the quality across both domains, and their opinion of the spatial quality influenced by the perceived timbral quality. An investigation will be undertaken to determine the influence this might have on the evaluation of spatial quality in this project.

2.2 Defining spatial quality for this research project

Letowski [1989] suggested that spatial quality is also a global assessment made up of a number of lower level attributes (see Fig 2.1). These lower level attributes are the spatial attributes that characterise the spatial audio scene in the reproduced sound environment. To fully understand spatial quality an investigation of spatial attributes present in the reproduced sound environment is required. Letowksi identified a few of these in his MuRAL, however recent elicitation experiments have established that a large number of different spatial attributes are perceivable within the reproduced sound environment.

2.2.1 Elicitation experiments

Several elicitation experiments have been undertaken by different researchers to identify the spatial attributes we perceive in the reproduced sound environment. A number of different methods have been developed for eliciting descriptive responses from perceptual information however a discussion on these falls outside the scope of this thesis.

A series of experiments conducted by Berg and Rumsey [1999a, 1999b, 2000a, 2000b, 2001, 2003 and 2006] employed repertory grid technique (RGT) to identify spatial attributes from the verbal responses from listeners. They used a number of different audio excerpts from audio recordings made using different recording techniques for replay over four different audio systems: mono, 2-channel

stereo, 4-channel surround and 3/2 stereo [ITU-R BS.775-1, 1992-1994]. The tests were quite extensive as they employed a large number of listeners. Similarly Zacharov and Koivunmiemi [2001a, 2001b and 2001c] employed quantitative descriptive analysis (QDA) to elicit attributes reproduced by three different reproduction systems: 2-channel stereo, 3/2 stereo, and an 8-channel periphonic Ambisonic system. They used a selection of different audio events (e.g. a passing train, male voice) and conducted the tests in three different acoustic environments (i.e. anechoic, BS.1116 standard listening room [ITU-R BS.1116-1, 1997] and a reverberation chamber). Gaustavino and Katz [2004] also conducted a number of elicitation experiments discovering that the attributes elicited were similar to those found by Berg and Rumsey, and Zacharov and Koivunmiemi, further supporting these studies. Choisel and Wickelmaier [2005, 2006a, 2006b], inspired by Berg and Rumsey, used RGT to determine what attributes were perceived in four different audio systems (i.e. 1.0 mono, 2-channel stereo, wide 2-channel stereo and 3/2 stereo) using downmixed and upmixed 5-channel music recordings including pop and classical music. From each of these studies different terminologies arose for similar attributes. Berg and Rumsey [2006] provided an interpretation of their work with Zacharov and Koivuniemi. A comparison of the attributes elicited in these three studies is given in table 2.2.

Berg and Rumsey [2006]	Zacharov and Koivuniemi [2001]	Choisel and Wickelmaier [2005]
Localisation	Sense of direction	-
Width	Broadness	Width
Envelopment	Broadness	Envelopment
Distance or depth	Distance to events, sense of depth	Distance
Room perception	Sense of space	Spaciousness
Naturalness and presence	Naturalness	Clarity and naturalness

Table 2.2 Comparison of spatial attributes elicited by several researchers.

2.2.2 Rumsey's perceptual hierarchy paradigm

Rumsey [2002] developed a novel scene-based paradigm which expands upon the elicitation experiments discussed above, defining meanings for each of the attributes and organising them into a perceptual hierarchy. The paradigm uses a macro- and micro-attribute system, the micro-attributes describing individual scene elements and the macro-attributes describing groupings of the individual scene elements or environmental attributes, this allows the scene to be described globally as an environment in which groups of sources, or individual sources can be identified.

2.2.2.1 Width

Width can be considered as both a micro and macro-attribute of the audio scene because it relates to the dimensions of the sources and the environment. There are four types of width attribute (Fig 2.3).

Individual Source Width: Describes the width of a sound source individually. It is often believed that this is negatively correlated to locatedness as described by Blauert [2001] (Fig 2.4).

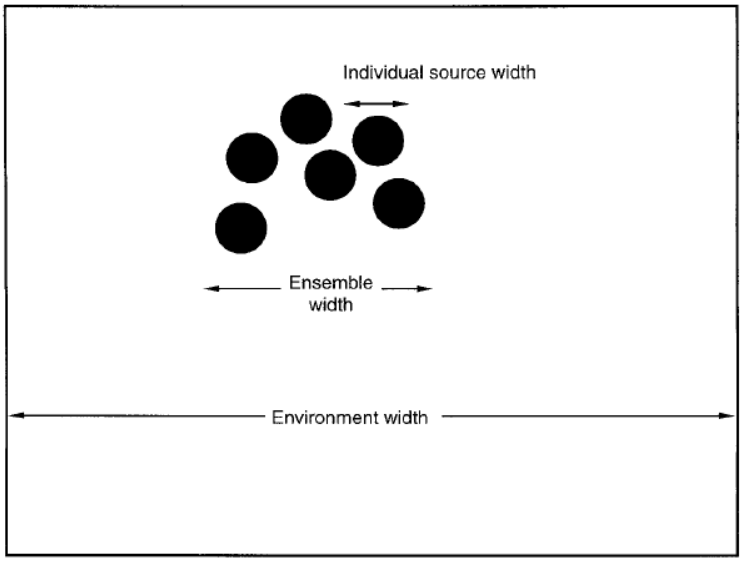


Fig 2.3 Examples of width attributes found in an audio scene [Rumsey, 2002].

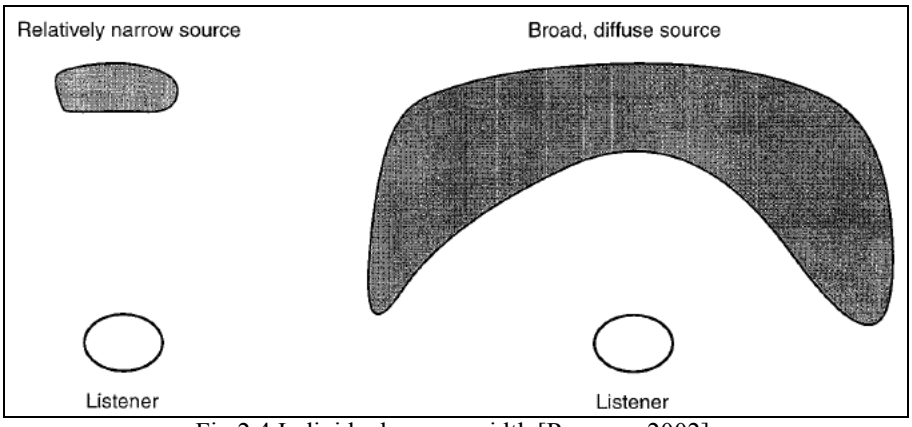


Fig 2.4 Individual source width [Rumsey, 2002].

Ensemble Width: The perceived width of a group of sources which share a common cognitive label.

Environment Width: The width of the background stream or reverberant energy within the scene. It describes the difference between a wide space and a narrow space.

Scene Width: This is the global (macro) attribute which enables a description of the entire scene including the reverberant energy (i.e. both foreground and background streams).

2.2.2.2 Depth and distance

Rumsey suggests evaluating depth and distance separately. Distance describes the perceived distance between the listener and the source. Whereas depth describes the distance between the front and back of a source, an ensemble of sources or the auditory environment (see Fig 2.5). Similarly to width, depth and distance relate to the dimensions of the sources and the environment, and so both can be considered as both micro and macro-attributes of the audio scene.

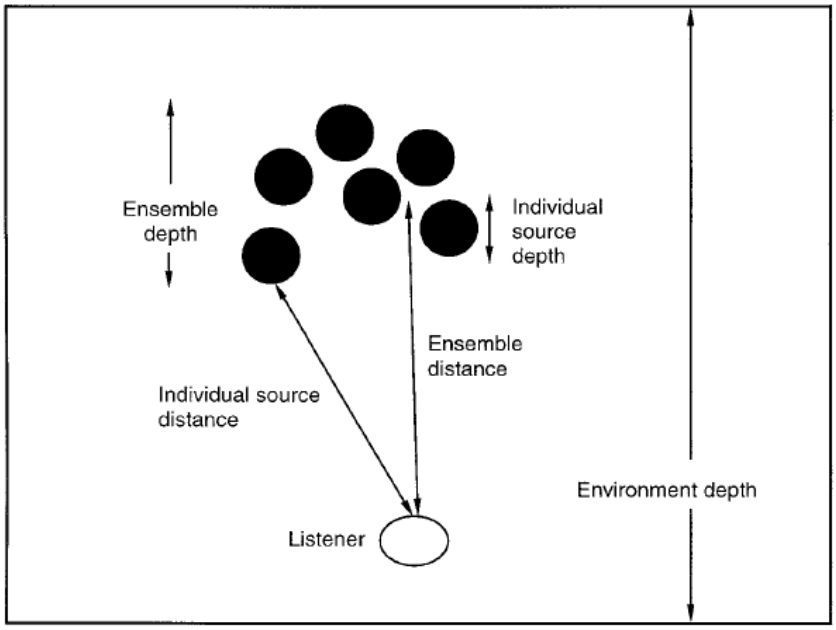


Fig 2.5 Examples of depth and distance attributes found in an audio scene [Rumsey, 2002].

Individual Source Distance: The perceived distance of a source from the listener.

Ensemble Distance: The perceived distance of the middle of an ensemble from the listener.

Individual Source Depth: The perceived depth of an individual source.

Ensemble Depth: The perceived depth of a group of sources.

Environment Depth: The perceived depth of the (reflective) source environment.

Scene Depth: The global depth of the entire audio scene, including the environment.

2.2.2.3 Envelopment

Envelopment is considered to be an environmental attribute and has historically been open to varied interpretation. The attribute envelopment was initially identified in concert hall acoustics research as listener envelopment (LEV) [Beranek, 1996]. In this context it is concerned with the enveloping sensation created by late arriving lateral reflections (e.g. after 50ms). However in the context of reproduced sound this definition may not be suitable, particularly when there is little or no reverberant content in the recording. With multichannel audio systems it is possible to produce additional types of envelopment whereby a sense of immersion can arise from one or more dry sources. Rumsey suggests that envelopment has three definitions.

Individual Source Envelopment: The sense of being enveloped by a single source.

Ensemble Source Envelopment: The sense of being enveloped by a group of sources.

Environmental Envelopment: The sense of being enveloped by the reverberant audio environment (background stream).

2.2.2.4 Presence

As spaciousness is similar to, and covered in the paradigm by, the definitions of environment width and depth, a new attribute was defined called presence. It is similar to environmental envelopment and similar to what Griesinger [1997] has called spatial impression.

Presence: the sense of being inside an (enclosed) space or scene; the feeling of being present in the audio space rather than absent.

2.2.2.5 Miscellaneous spatial attributes

Rumsey also considers additional attributes in the paradigm. These attributes do not belong to any attribute group.

Scene left-right skew: Degree to which a spatial audio scene is skewed to the left or to the right from a stated reference position.

Scene front-back skew: Degree to which a spatial audio scene is skewed to the front or back from a stated reference position.

Source Stability: Degree to which individual sources remain stable with respect to time (assuming nominally stationary sources)

Scene Stability: Degree to which entire scene remains stable in space with respect to time.

Source Focus: Degree to which individual sources can be precisely located in space (this may be closely related to Individual Source Width).

Scene Width Homogeneity: Evenness of distribution of scene elements compared with a reference scene.

2.2.3 Spatial quality: summary and conclusions

Elicitation experiments conducted by several different researchers have identified that we perceive a number of attributes in spatial audio scenes related both to the individual scene elements themselves and to the reproduced environment. Despite a difference in terminology there are clear similarities allowing a generic set to be established, and using this Rumsey developed a hierarchical paradigm expanding and defining each of these terms.

The ITU describes BAQ as the attribute accounting for ‘any and all differences between the reference and impaired items’ in a recording. In this context spatial quality can be defined for the reproduced sound environment in this research project as the attribute that describes any and all differences between the reference and impaired items, but only in the spatial characteristics of the recording. Hence in this respect an evaluation of spatial quality can be considered as a higher level assessment of the lower level spatial attributes, such as those identified in this section, the evaluation being drawn from both hedonic and non-hedonic judgements of the lower level attributes. It is hoped

that listeners will be capable of assessing spatial quality consistently. However this will be need to be confirmed via pilot studies, before a large scale collection of subjective responses can be undertaken.

2.3 Review of current sound quality models

A number of objective models for predicting sound quality have been created by different researchers. These models are reviewed here in order to identify novel areas for investigation and to determine acceptable performance criteria for the calibrated QESTRAL model.

2.3.1 Method for objective measurements of perceived audio quality (PEAQ) (ITU-R BS.1387)

ITU-R BS.1387 – 'Method for objective measurements of perceived audio quality (PEAQ)' – is the adopted standard for the objective assessment of perceived audio quality. The ITU recognised that an objective model with the ability to estimate perceived audio quality would be useful as a design tool for modern digital systems in broadcast applications, particularly in light of modern bit-rate reduction schemes. It was agreed that traditional objective measurements of audio quality such as Signal-to-Noise Ratio (SNR) and Total Harmonic Distortion (THD) were not reliable representations of perceived audio quality and furthermore were not sophisticated enough to scrutinise non-linear and non-stable changes to audio quality such as those produced by modern low bit-rate audio codecs. PEAQ is based upon six independently developed models:

Disturbance Index (DIX) [Thiede and Kabot, 1996]

Noise-to-Mask Ratio (NMR) [Brandenburg, 1987]

Objective Audio Signal Evaluation (OASE) [Sporer, 1997]

Perceptual Audio Quality Measure (PAQM) [Beerends and Stemerdink, 1992]

PERCEVAL [Paillard *et al*, 1992]

Perceptual Objective Measure (POM) [Colomes *et al*, 1995]

PEAQ uses an intrusive approach whereby audio quality changes are evaluated by comparisons between a reference audio system and an impaired version of the reference system (device under test (DUT)). PEAQ uses a selection of natural test signals (speech or music) and synthetic test signals to scrutinise the DUT. The different models have a correlation (r) of between 0.67 – 0.86 with the subjective data used to calibrate them. However PEAQ is only designed to consider timbral changes to BAQ in monophonic audio systems and, although it can be used to assess the BAQ of 2-channel stereo systems, it does not take account of the spatial characteristics and therefore is not capable of measuring changes to spatial quality in multichannel audio systems.

2.3.2 Quality Advisor (QA)

Zielinski *et al* [2004][2005a] developed a form of parametric model for predicting the BAQ of a multichannel audio system. The Quality Advisor (QA) was designed as a decision making tool for broadcast engineers and codec designers. It was created by a combination of two previously developed models; 'Predictor A' which was designed to predict the change to BAQ resulting from bandwidth limitation and 'Predictor B' which was designed to predict changes created by downmixes. For simplicity the QA used a look up table of subjective data collected from listening tests [Zielinski *et al*, 2003a, 2003b] to advise the user of the resulting change in quality. The user is required to input a number of criteria describing the source material and required data reduction. The QA's output provides the user with a number of methods for reducing the data rate while maintaining high quality.

The QA was calibrated to a high standard with a correlation (r) of 0.93 and a root mean square error (RMSE) of 9% to the subjective data. However, as the authors acknowledged, the scope of the QA is limited as it is restricted to only providing the user with solutions based upon the audio processes investigated (a selection of bandwidth limitations and downmixes). The authors also recognised that a better model could be produced by employing metrics which measure the physical characteristics of the audio material.

2.3.3 Model created by Choi *et al*

Choi *et al* [2008] (an earlier version of the model was also discussed in Choi *et al*, 2007) proposed a multichannel addition to the PEAQ standard. The model used ten model output variables (MOV) from PEAQ with three additional spatial metrics; Interaural Level Difference (ILD) distortion, interaural time difference (ITD) distortion and interaural cross-correlation coefficient (IACC) distortion, to predict degradations to BAQ created by multichannel audio codecs.

As with PEAQ their model uses an intrusive method of prediction. It has three sequential parts. The first stage synthesises binaural signals from the reference system and DUT. The second stage is a peripheral ear model also used in PEAQ which converts the binaural signals to neural signals. In the third stage the metrics are used to predict the subjective scores. This is achieved through an artificial neural network or linear estimator.

Their model was calibrated using listening tests investigating the effect of low bit rate multichannel audio codecs on BAQ as opposed to spatial quality. A validation of the model was also calculated using a different group of listeners from the same database. The model showed good correlation with the subjective database. Using the artificial neural network a correlation (r) of 0.85 was achieved with an RMSE of 5.09%. While using the linear estimator a correlation (r) of 0.79 with an RMSE of 5.44% was achieved.

Although this model does provide a form of spatial audio addition to the PEAQ standard it is limited as it has only been calibrated for the evaluation of multichannel audio codecs and would

therefore require re-calibration before it could be used to assess other types of SAPs (e.g. downmixing or loudspeaker misplacements).

2.3.4 Models created by George *et al*

George [2009] (and [George *et al*, 2006a/b]) developed objective evaluation models for the prediction of frontal spatial fidelity, surround spatial fidelity and the timbral fidelity of multichannel audio systems. These models use an intrusive method of prediction comparing an impaired audio system against a reference audio system. The models were calibrated using data collected by Zielinski *et al* [2003a, 2003b], and validated using data collected by George [2009], and were designed to have a target specification correlation (r) of 0.9 between the subjective and predicted scores and RMSE (Root Mean Square Error) of 10%. This target specification was based upon the performance of PEAQ [ITU-R BS.1387, 2001] and PESQ [1996] and the reported listener error from the listening tests [Zielinski *et al*, 2005b].

To produce the fidelity models subjective data was collected from tests employing a similar test method to MUSHRA [ITU-R BS.1534, 2001], using several different items of 5-channel programme material processed using bandwidth limitation or by downmixing. The objective data was collected using a selection of 22 metrics (these will be discussed in section 4.2.2) to measure the physical characteristics of both unprocessed and processed programme material. Using regression analysis the objective metrics were fitted to the subjective data to meet the target specifications. The results of the calibration and validation calculations are shown in table 2.3

	Calibration		Validation	
Model	Correlation (r)	RMSE (%)	Correlation (r)	RMSE (%)
Frontal spatial fidelity	0.91	9.33	0.88	15.45
Surround spatial fidelity	0.95	8.87	0.87	14.19
Timbral fidelity	0.95	7.72	0.92	8.37

Table 2.3 Performance summary of quality models developed by George [2009].

These models performed very well, however as with Choi *et al's* model, George *et al's* models are limited as they were calibrated for the evaluation of programme material processed using only bandwidth limitation and downmixing.

2.3.5 Sound quality models: summary and conclusions

PEAQ is the current standard for objectively measuring perceived audio quality. It was created from a number of different models created by several researchers using an intrusive approach whereby changes to BAQ are evaluated by comparisons between a reference audio system and a DUT. A selection of natural test signals (speech or music) and synthetic test signals were employed to scrutinise the DUT. However PEAQ is only designed to consider timbral changes to BAQ in monophonic audio systems and is not capable of measuring changes to spatial quality in multichannel

audio systems. A recent model developed by Choi *et al* proposed an expansion of the PEAQ model to multichannel audio systems. The model showed good correlation with the subjective database. Using the artificial neural network a correlation (r) of 0.85 was achieved with an RMSE of 5.09%. While using the linear estimator a correlation (r) of 0.79 with an RMSE of 5.44% was achieved. However as discussed above this model is only capable of assessing programme material processed by multichannel audio codecs. George *et al* produced models that considered spatial and timbral quality separately. Similarly to PEAQ an intrusive approach was employed, however metrics extracted characteristics from the programme material employed in the listening tests instead of test signals. These models performed very well (see Table 2.3) however again they are limited to the evaluation of programme material processed using bandwidth limitation and downmixing.

George specified performance criteria for the development of his models. The target specifications for the models were for them to achieve a correlation (r) equal to or greater than 0.9 and RMSE of less than 10%. This was based upon the performance of PEAQ and PESQ and achieving an RMSE (%) similar or better than the reported listener error from the listening tests..Similar criteria will be considered for the QESTRAL model and will be discussed in chapter 3.

The models created by Choi *et al* and George *et al* can both be considered as models that incorporate spatial quality to some degree and both showed good performance, however they were both calibrated using a limited selection of audio process types (multichannel audio coding, bandwidth limitation and downmixes). Although the degradation to spatial quality created by these processes could be considered as of high importance for research and product development engineers, there are other potential degradations to spatial quality which are of similar importance. These could include degradations created unintentionally by the consumer such as the misplacement of loudspeakers from their intended positions, or connecting the loudspeakers to the incorrect output of the distribution amplifier. Other degradations could also include broadcasting errors such as the inter-channel level misalignment or phase reversal or even combinations of all of the above. Therefore the QESTRAL model will be designed to measure a greater range of SAPs such as those mentioned here.

2.4 Summary and conclusions

Chapter 2 concentrated on investigating and defining the term spatial quality for this research project. An introduction to sound quality and spatial quality was provided. This was followed by an overview of the spatial attributes present in the reproduced sound environment, after which a research definition for spatial quality was established. Current objective models for sound quality were reviewed in order to identify novel areas for investigation and also to determine acceptable performance criteria that the QESTRAL model should achieve.

A quality judgement is a comparative judgement whereby the standard of something is compared to other things like it. In reproduced sound, an opinion of sound quality is established

through a combination of both sensory and affective judgements. Letowski proposed that listeners perceive and assess two different domains of sound quality, identifying them as timbral quality and spatial quality. He suggested that spatial quality is a global assessment made up of a number of lower level attributes (see Fig 2.1). Elicitation experiments conducted by several different researchers have identified that we perceive a number of different spatial attributes in spatial audio scenes related both to the individual scene elements themselves and the reproduced environment. Rumsey developed a hierarchical paradigm expanding and defining each of these terms. Based upon these studies and the attribute BAQ defined by the ITU, a definition for spatial quality was established for this research project as the attribute that describes any and all differences between the reference and impaired items, but only in the spatial characteristics of the recording. Hence in this respect an evaluation of spatial quality can be considered as a higher level assessment of the lower level spatial attributes, such as those identified in section 2.2.

When investigating frontal spatial fidelity and surround spatial fidelity, Zielinski *et al* found that the audio processes they investigated degraded both timbral fidelity and spatial fidelity. Letowski also suggests that we have limited ability to evaluate quality when different domains vary simultaneously. Therefore it might be possible for listeners to become confused if a SAP causes a change in the quality across both domains, and their opinion of the spatial quality may be influenced by the perceived timbral quality. In the context of this research project, it will not be possible to completely separate these two domains. So it will be important to establish the potential influence of changes to timbral quality, created by different SAPs, on a listener's opinion of spatial quality (see section 6.4).

A selection of sound quality models were reviewed however only the recent models created by Choi *et al* and George can be considered as incorporating spatial quality to some degree. Both of these showed good performance, however they were both calibrated using a limited selection of audio process types (multichannel audio coding, bandwidth limitation and downmixes). The QESTRAL model will be designed to measure a greater range of SAPs.

George specified performance criteria for the development of his models. This was based upon the performance of PEAQ and PESQ and the reported listener error from the listening tests. The target specifications were for the models to achieve a correlation (r) equal or greater than 0.9 and RMSE of less than 10%. Similar specifications will be employed for the calibration of the QESTRAL model.

Chapter 3 – Methods for the development of the QESTRAL model

In chapter 2 a working definition for spatial quality was established and current objective models for sound quality were reviewed in order to identify novel areas for investigation and guidelines for acceptable performance criteria that the QESTRAL model should achieve.

Chapter 3 discusses topics relating to how the QESTRAL model will be created. Firstly an appropriate method for the development of the QESTRAL model is established. This section describes an appropriate research procedure that could be used to create the model. Following this a discussion of the most appropriate method of regression analysis for calibrating the model is presented. This leads into a discussion of suitable target specifications for its performance. Finally a discussion on reproduction systems is provided, from which the most appropriate system to use as a reference system in the QESTRAL model is chosen.

3.1 QESTRAL model development method

As discussed in section 2.1 an assessment of sound quality is considered as a global judgement of a number of lower level attributes. Bech [1999] indicates a framework that could be employed for the development of a perceptual model for the objective evaluation of sound quality. This can be divided into two approaches. George [2009] describes these as direct and indirect prediction. A direct prediction is where the model is calibrated using subjective data collected on a global assessment of sound quality and objective metrics selected to measure the global and/or lower level attributes that comprise it. For an indirect prediction subjective data is collected on the lower level attributes of sound quality independently and objective metrics are selected to measure each one. The model is calibrated by mapping the predicted low level attributes to the global attribute using multivariate analysis. Both methods are illustrated in figures 3.1 and 3.2.

A definition for spatial quality was established in section 2.2: spatial quality is the attribute that describes any and all differences between the reference and impaired items, but only in the spatial characteristics of the recording. Hence in this respect an evaluation of spatial quality can be considered as a higher level assessment of the lower level spatial attributes (e.g. the attributes identified in section 2.2). Although a number of studies have identified various attributes of the spatial audio scene and a perceptual hierarchy has been proposed [Rumsey, 2002], the suggested contribution that each lower level spatial attribute has to sound quality or spatial quality has not been quantified. Achieving this would require a substantial amount of time and research, which would not be possible during this research project and so it is for this reason that a direct prediction method will be employed for the

development of the QESTRAL model. This approach has been used successfully by other researchers such as Zielinski *et al* [2003a, 2003b, 2005b] and George [2009] to predict frontal spatial fidelity and surround spatial fidelity (as discussed in section 2.3.4).

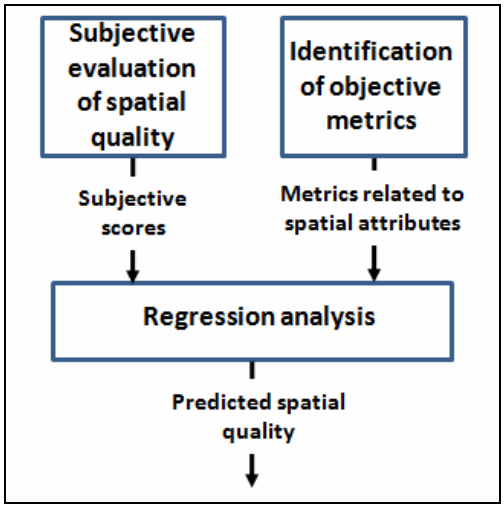


Fig 3.1 Direct prediction development procedure.

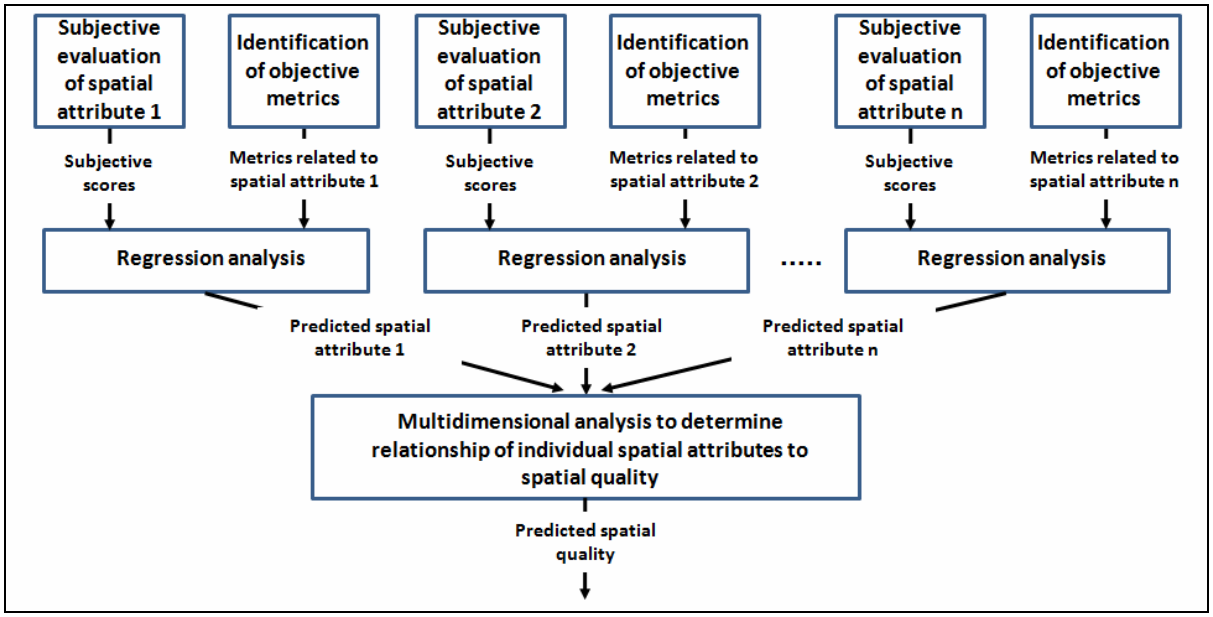


Fig 3.2 Indirect prediction development procedure.

A potential risk of using the direct approach is that the experimenter mistakenly limits the generalisability of the model by only collecting data on some of the component attributes that contribute to the global attribute (e.g. if the stimuli tested do not exhibit traits of all of the lower level spatial attributes). Therefore to develop a generalisable model using a direct prediction method the subjective data used to calibrate the model should be collected from a set of stimuli that exhibits a range of changes to all of the lower level attributes. In the case of this project this means that the SAPs investigated should stress the lower level spatial attributes (see section 2.2).

3.2 Calibrating the QESTRAL model using linear regression analysis

It is possible to calibrate objective evaluation models using Artificial Neural Networks (ANN) as shown by Choi *et al* [2008]. Nevertheless, due to greater availability and experience, the QESTRAL model will be calibrated using linear regression analysis. Regression analysis is a method by which the relationship between a set of variables can be explained [Draper *et al*, 1981]. These variables can be divided into two groups; independent variables (objective metrics) and dependent variables (subjective scores collected from listening tests). Multiple Linear Regression (MLR) is a regression analysis method that attempts to establish a linear relationship between a number of independent variables and the dependent variable. The calibration of the QESTRAL model for the objective evaluation of spatial quality has not been attempted before, therefore to achieve the best performing model the use of a large number of metrics will be investigated to identify the best combination for the prediction of spatial quality. However if a large number of metrics are used problems with multicolinearity can occur. Multicolinearity between the metrics indicates that they predict similar components of the dependent variable, which Field [2005] indicates can limit their achievable prediction of the dependent variable. Principal Component Regression (PCR) is a form of MLR which attempts to deal with the problem of multicolinearity between metrics, by using principal component analysis (PCA) to group co-varying metrics into orthogonal groups called principal components (PCs). The PCs are used as new independent variables to predict the dependent variable. However to use this method successfully knowledge about the dependent variable is required in order to manually identify the optimal selection of metrics. Therefore for this research project another type of regression analysis is more appropriate.

3.2.1 Partial least squares regression

Similarly to PCR, in PLS regression comparable information or components is/are identified within the metrics relevant for the prediction of the dependent variable (spatial quality) and grouped into latent variables (or principal components). However the grouping of metrics into latent variables is also determined for the highest predictive power of the dependent variable, rather than only co-variance in the metrics. Metrics that do not fit into the latent variables are discarded. The contribution of the metrics within each latent variable is still free to vary so that an optimal weighting can be identified. This approach effectively deals with multicolinearity while also allowing the optimal selection of metrics to predict the dependent variable to be determined [Esbensen, 2002]. Therefore PLS regression was chosen as the preferred method of regression modelling, because it is suitable for calibrating models using a large selection of metrics [Abdi, 2007] and also gives the investigator freedom to experiment with different metric combinations.

The software employed to run the PLS regression and calibrate the model will be Camo's The Unscrambler version 9.8. The Unscrambler provides an intuitive graphical output which is particularly

useful for calibrating a regression model. In a forced entry method of calibration all metrics are included and considered, and the most suitable combinations are determined through a series of iterations. Particularly important in the absence of a separate data set to validate the model, The Unscrambler also allows the ability of the model to predict a new data set to be forecast using a leave-one-out full cross-validation [Esbensen, 2002].

3.3 QESTRAL model target specifications

In section 2.3 the performance of several quality models, PEAQ and those developed by Choi *et al* [2008] and George [2009]) was discussed. Based upon this discussion a number of target specifications for the performance of the QESTRAL model are defined.

The maximum correlation achieved by PEAQ, between the predicted and subjective data, was 0.86. Therefore this value is chosen as the minimum correlation for the QESTRAL model. This will also make it competitive with the models created by both Choi *et al* and George. It is desirable to achieve this for both the calibration and cross-validation of the model. The second criterion is that the model should have a root mean square error (RMSE) (a measure of the error between the predicted and subjective data) similar or better than the average intra-listener error observed in the subjective data collected from the listening tests. George employed this idea to set a threshold for the RMSE (%) of his models, based on the principle that the model should not be less reliable than the listeners. The exact value will be chosen after the listener performance in the listening tests has been analysed.

It is most important that the model accurately predicts the subjective data collected during this research, but it is also desirable that the model will generalise and be capable of accurately predicting databases of subjective scores collected from the evaluation of different types of SAPs using different listeners. So some additional constraints for the model will be included to help achieve this.

Although PLS regression was designed to accept a degree of multicolinearity between independent variables, if the QESTRAL model is to be generalised, the metrics selected for the final model should exhibit low multicolinearity. Low multicolinearity would indicate that each metric measured something unique in the changes to the spatial characteristics created by the SAPs. Multicolinearity can be measured from the variance inflation factor (VIF) of each metric used in the model [Field, 2005]. The value of the VIF indicates whether there is strong linear relationship (correlation) between the independent variables used in the model (e.g. A high VIF indicating that multicolinerality exists between them). Based upon the work of other statisticians Field recommends a number of different thresholds which suggest that a VIF greater than 5 (and certainly greater than 10) reveals that an independent variable has high multicolinearity to the other variables in the model, while the closer the mean VIF is to 1 the lower the multicolineairty. Hence it is proposed that in the interest of achieving low multicolinearity in the QESTRAL model the metrics used in the model should exhibit an average VIF close to 1.

If a model uses a large number metrics, the number of degrees of freedom available in the regression calculation, to determine the most suitable weighting for each metric, is reduced [Field, 2005]. There is high chance in this case that the model will be over-fitted and predict the error in the data rather than the trend. An over-fitted model is not reliable because it is context dependent and therefore only suitable for predicting the data it was calibrated with. It follows that a model with fewer metrics is more robust because there is a larger number of degrees of freedom to determine the most suitable coefficient for each metric in the model. Therefore it is desirable that the QESTRAL model uses the minimum number of metrics and principal components (PCs) to achieve the target specifications. In addition to helping the model generalise this will also mean that using the model to predict spatial quality will be simple and straightforward.

Although the constraints discussed above will be considered during the QESTRAL development process, the generalisability of the model will also be checked statistically using tests suggested by Field [2005].

3.4 Spatial audio reproduction systems – selecting a system for this study

As this research was conducted during a finite period where only a limited amount of experimental work was possible, it was necessary to select just one audio system with which to calibrate the QESTRAL model. In this context the most suitable system was determined by its ability to reproduce the psychoacoustic cues for spatial attributes and by it's commercial popularity. A brief discussion is provided of the abilities of the most popular consumer audio systems to reproduce spatial attributes.

3.4.1 Monophonic (1.0)

The first audio systems which took the form of gramophones and phonographs (invented by Thomas Edison) in the late 19th and early 20th century were monophonic. Having only one channel they are only capable of reproducing limited spatial cues for depth and distance based upon the human auditory system's perception of reverberation [Rumsey, 2001]. However it is accepted that they can also reproduce spatial cues associated with single source location from the loudspeaker's localised position. Monophonic systems are still in use today, for example many small/portable radio sets are monophonic. However they have largely been superseded by 2-channel stereophony and other more spatially advanced systems.

3.4.2 2-channel stereophony (stereo)

2-channel stereo describes an audio system where two loudspeakers are positioned in front of the listener usually with the loudspeakers positioned at a subtended angle of 60° [ITU-R BS.775-1, 1992-1994] (Fig 3.3).

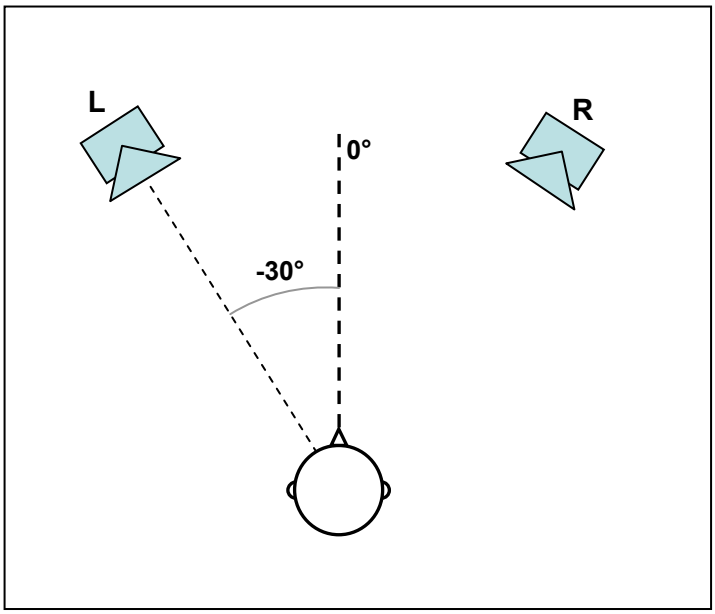


Fig 3.3 2-channel stereophony loudspeaker configuration [ITU-R BS.775-1, 1992-1994].

2-channel stereo was developed in the thirties [Blumulein, 1958] and commercialised in the late fifties and early sixties and has become commonplace in the home since the late sixties. The majority of music releases and radio and TV broadcasts are delivered to be replayed using this format.

A 2-channel stereo system is capable of reproducing cues for individual scene elements such as localisation, width, depth and distance. However it is only capable of reproducing these cues in front of the listener. Nevertheless this system has been shown to be capable of reproducing the sensation of relatively high envelopment [Conetta, 2007][George, 2009].

3.4.3 3/2 stereo

3/2 stereo, also known as 5.1 surround if a low frequency effect (LFE) channel is included in the system, has become familiar and is very popular in both professional and consumer circles. It is currently the standardized surround sound loudspeaker layout for consumer applications such as home cinema and DVD [Rumsey, 2001] and is the format for which the programme material delivered by broadcasters in their HD broadcasts is intended. It is also popular for audio-only applications. The setup provides three loudspeakers in front of the listener and two behind. The arrangement of loudspeakers for this system is defined in ITU-R BS.775-1 [1992-1994] (Fig. 3.4).

This layout (Fig 3.4) was designed for use in home cinema applications allowing 2-channel stereo (L and R) with an additional centre channel (C) in the front section (which is most often used

for dialogue), and two channels (Ls and Rs) for supporting ambience or effects content in the rear section, behind the listener. As discussed in section 2.2, 3/2 stereo is capable of reproducing a large number of different spatial attributes.

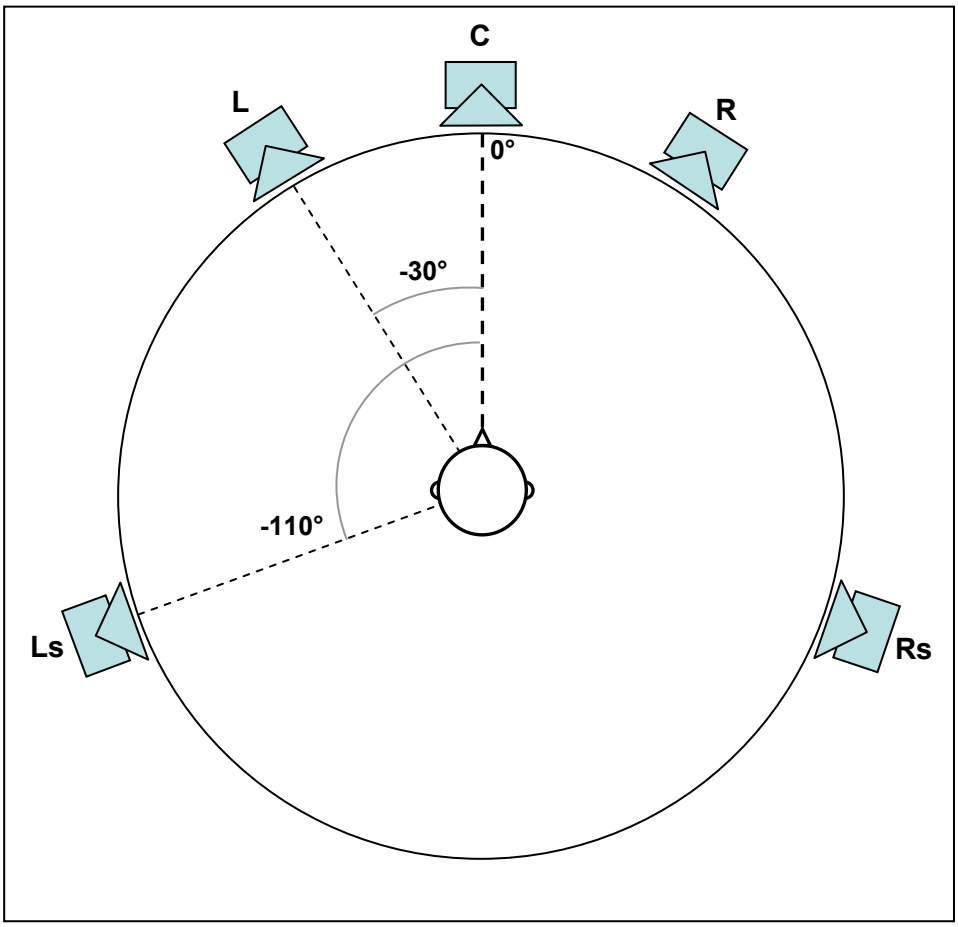


Fig 3.4 3/2 stereo loudspeaker configuration [ITU-R BS.775, 1994].

Similarly to 2-channel stereophony it is possible to reproduce cues for individual scene elements such as localisation, width, depth and distance in front of the listener. The addition of the rear loudspeakers makes this also possible behind and to the sides of the listener. However due to the distances between the loudspeakers the cues are by comparison much less stable [Martin *et al*, 1999][Rumsey, 2001]. Hiyama *et al* [2002] reported that the position of the loudspeakers is optimal for the reproduction of a diffuse soundfield. This is important for the reproduction of cues for environmental attributes such as envelopment and spaciousness. Morimoto [1997] supports this argument, indicating that rear loudspeakers can be used to enhance listener envelopment.

Similarly to 2-channel stereophony, 3/2 stereo is simply a loudspeaker layout format and it is thus the responsibility of the service provider to decide how best to deliver their content. There are two ways of delivering content over this system [Rumsey, 2001]. The first is via matrixing where the audio is delivered in a data compressed form (e.g. as two channels instead of five) and then recovered before replay by a decoder such as Dolby Surround or Dolby Prologic. This approach is often employed by

broadcasters. The second is where the audio is delivered in its original 5-channel form for immediate replay and does not require decoding. This is seen on multichannel audio dedicated media such as DVD Audio releases.

3.4.4 Other reproduction systems

There are various other surround sound or spatial audio systems such as Ambisonics, 7.1, 10.2 [Rumsey, 2001] and Wave Field Synthesis (WFS) [De Vries, 2007]. Ambisonics and WFS for example are both very sophisticated systems with the ability to reproduce very realistic spatial scenes and have both been the subject of much research. Nevertheless currently these systems have yet to achieve commercial success similar to the three systems discussed, and therefore are not included in this discussion.

3.4.5 Spatial audio reproduction systems: summary and conclusions

The considerations for the selection of a suitable audio system were the ability of the system to reproduce spatial attributes and its commercial popularity. It can be seen that as the sophistication of the system increases the ability to reproduce spatial attributes increases. However it was important to make this study ecological and therefore the system chosen also had to be representative of those in widespread use.

BS.775 3/2 stereo is capable of reproducing the highest number of spatial attributes of the systems reviewed. It is also currently the only surround sound system in widespread use, with a large number of service providers producing content for it. It is also capable of replaying mono and 2-channel stereo material and so allows these systems to be investigated simultaneously. Therefore this system is the most suitable choice for calibrating the QESTRAL model.

3.5 Summary and conclusions

Chapter 3 discussed topics relating to how the QESTRAL model will be created. Firstly an appropriate method for the development of the QESTRAL model was established. Following this a discussion of the most appropriate method of regression analysis for calculating the model was presented and a discussion of suitable target specifications for its performance. Finally a discussion on reproduction systems was provided, from which the most appropriate system to use as a reference system in the QESTRAL model was chosen.

A direct prediction method, as defined by Bech, will be employed for the development of the QESTRAL model. In a direct prediction method the subjective data is collected on a global assessment of audio quality and objectives metrics are selected to measure the global and/or lower level attributes that comprise it. A potential risk of using the direct approach is that the experimenter mistakenly limits the validity of the model by only collecting data on some of the component attributes that contribute to

the global attribute (e.g. if the stimuli tested do not exhibit traits of all of the lower level spatial attributes). In the case of this project this means that the SAPs investigated should stress the lower level spatial attributes (see section 2.2). A method of determining that this is achieved will be developed.

The QESTRAL model will be calibrated using partial least squares (PLS) regression. This method of regression analysis was chosen because it is adept at calibrating models using a large selection of independent variables and gives the investigator freedom to experiment with the use of different metrics. The QESTRAL model will be calibrated to meet the following target specifications (Table 3.1)

Criteria	Target specification
Correlation (r)	≥ 0.86
Root Mean Square Error (RMSE) (%)	≈ average intra-listener error
Variance Inflation Factor (VIF)	Mean VIF ≈ 1

Table 3.1 QESTRAL model target specifications.

It is also desirable to calibrate the QESTRAL model so that it may perform well in the prediction of the perceived change to spatial quality created by SAPs not investigated in this project. Therefore the QESTRAL model will be calibrated using the minimum amount of metrics and principle components required to meet the target specifications. The generalisability of the model will also be checked statistically using a number of statistical tests suggested by Field.

The considerations for the selection of a suitable reference audio system were the ability of the system to reproduce spatial attributes and its widespread use. After a study of current commercial reproduction systems it was decided that 3/2 stereo was the most suitable system for this research. This system is also capable of replaying mono and 2-channel stereo material and so allows these systems to be investigated simultaneously.

Chapter 4 – Review of objective metrics that could be used in the QESTRAL model

This chapter reviews objective metrics currently used to measure individual spatial attributes in reproduced sound and existing spatial quality models. The aim of this review is to identify suitable metrics that could be employed to measure changes to spatial quality that are created by the SAPs. These metrics could then be employed in the QESTRAL model to predict spatial quality.

4.1 Metrics for individual spatial attributes of reproduced sound

The research definition for spatial quality given in section 2.2 describes it as a global evaluation of changes to a number of lower level spatial attributes affected by a SAP, when compared to an unprocessed reference recording. A number of metrics have been developed, by different researchers, to measure changes to individual spatial attributes. This section investigates a selection of relevant metrics that could be used to measure changes to the lower level spatial attributes created by the SAPs.

4.1.1 Metrics used by Choisel and Wickelmaier

Choisel and Wickelmaier [2006a] describe the correlation of a number of different metrics, designed to measure both timbral and spatial characteristics, to perceptual attributes elicited from listeners in their experiments [Choisel and Wickelmaier, 2005][Choisel and Wickelmaier, 2006b] (discussed in section 2.2.1). Firstly using a panel of listeners they identified and quantified the presence of eight spatial and timbral attributes within a selection of audio recordings. They then measured these recordings using metrics derived from room acoustics (see Table 4.1) and using linear regression examined the correlation (see Table 4.2) between a particular metric and the subjective responses to the recordings. This allowed them to identify the suitability of their metrics for measuring individual spatial attributes.

The spatial metrics based upon the measurement of interaural cross-correlation (IACC) and lateral fraction (LF) showed good correlation with perceived width, envelopment, spaciousness and distance.

The spectral metrics spectral centroid (f_c) and sharpness (S) correlated poorly with the elicited attributes. In particular they did not correlate highly with the timbral attributes brightness and clarity in their tests, although the researchers had hoped they would. This study is important because it establishes the relationship of selection of different metrics to different spatial attributes, in particular

metrics based upon IACC and LF correlated well with the perception of a number spatial attributes, such as envelopment, width and spaciousness.

Metric	Type	Description
IACC	Spatial	IACC calculated from binaural recordings of the stimuli.
$IACC_f$	Spatial	Half-wave rectification of IACC using a third-order Butterworth low pass filter with a 1-kHz cutoff frequency.
LFT	Spatial	Total lateral reflection. The ratio of early sound energy arriving laterally over sound energy arriving from all directions Barron and Marshall [1981].
$IACC_{sim}$	Spatial	Identical to $IACC_f$ but calculated directly from the the loudspeaker signals in a simulated soundfield.
LF_{sim}	Spatial	Identical to LFT but calculated directly from the the loudspeaker signals in a simulated soundfield.
f_c	Timbral	The spectral centroid calculated from 1/3 octave band spectra of the binaural recordings.
S	Timbral	Sharpness [Zwicker and Fastl, 1999] calculated from the binaural recordings using Br¨uel & Kjær's PULSE Sound Quality software.

Table 4.1 Metrics employed by Choisel and Wickelmaier to measure timbral and spatial characteristics.

Metric	Width	Envelopment	Spaciousness	Distance	Brightness	Elevation	Clarity
IACC	0.75	0.67	0.56	0.23	0.42	0.34	0.38
$IACC_f$	0.6	0.71	0.83	0.57	0.46	0.51	0.62
LFT	0.88	0.71	0.90	0.48	0.39	0.23	0.66
$IACC_{sim}$	0.75	0.78	0.74	0.81	0.30	0.32	0.57
LF_{sim}	0.9	0.77	0.93	0.65	0.40	0.28	0.71
f_c	0.2	0.05	0.01	0.09	0.10	0.29	0.00
S	0.09	0.01	0.01	0.03	0.04	0.40	0.01

Table 4.2 Correlation (r) of the metrics employed by Choisel and Wickelmaier to the perceptual attributes elicited in their study.

4.1.2 Automatic localisation models

Localisation is an ability of the human auditory system which allows the listener to establish the location, or position, of a sound event in their environment. It is fundamental to a listener's perception of the spatial scene. To localise a sound event in their environment a listener predominantly uses two auditory cues – the interaural time difference (ITD) for low frequency sounds and the interaural level difference (ILD) [Blauert, 2001]. The models discussed below are based upon these two primary cues.

Pocock [1982] devised a method for sound event localisation using signals collected from a KEMAR dummy head. From these the Interaural Time Difference (ITD) and Interaural Level Difference (ILD) of the sound event, as perceived by a listener, were calculated to estimate its location. The main limitation of this model was that it could only be used under acoustically anechoic conditions.

A decade later, Macpherson [1991] expanded upon what Pocock had achieved, enabling the model to be used in both anechoic and reverberant environments and for the detection of both transient and steady state signals (although these could not be realised simultaneously). Macpherson's model could only be used for frontal horizontal analysis; this was appropriate for use with the two-channel stereophonic systems which were prevalent at the time. Pulkki [1999] developed a similar tool which included the ability to evaluate timbre.

These early models provided simple localisation tools for the evaluation of reproduced stereophonic images. However, the majority of auditory localisation information from a sound event is conveyed in the initial transient phase, usually in the first 2ms of the event, which the early models were not particularly accurate at detecting. Supper [2005] introduced a model that detected auditory onsets. This used a binaural system and fast predictive filtering to evaluate transient information across various critical frequency bands. Expanding upon this research, Supper then developed a localisation (lateralisation) tool that utilised this onset detection method. In this tool the lateral angles are resolved using mapping and duplex theory weighting combinations of ITD and ILD measurements of a binaural signal divided into 24 critical frequency bands using gammatone filter bank. Adding to Supper's work, Dewhirst [2008] made several modifications to improve its performance. These included reducing error in the look-up tables used to calculate the ITD and ILD measurements, adding simulated head movements and altering the way in which the ITD and ILD measurements were combined. Dewhirst's improvements were validated using a formal listening test. The algorithm has a coefficient of determination (R^2) of 0.98 to the listening test results.

4.1.3 Metrics for measuring envelopment and width

A number of studies have proposed metrics for the measurement of envelopment and width. The measurement of envelopment is particularly important as it is believed that much of the enthusiasm for multi-channel audio systems stems from their ability to reproduce this attribute [Soulodre et al, 2002]. As indicated by Choisel and Wickelmaier [2006a] metrics based upon IACC are useful for the measurement of attributes such as perceived envelopment, width and spaciousness in reproduced sound. This idea originated in concert hall acoustics research. Beranek [1996], summarising the work of others, showed that the mean IACC measured at 500Hz, 1000Hz and 2000Hz correlated well with a listener's opinion of two spatial components of a concert hall listening experience, the apparent source width (ASW) and listener envelopment (LEV). It was suggested the IACC measured up to 80ms after the sound event was most correlated with ASW, while the IACC measured after 80ms after the sound event correlated most highly with LEV.

Mason [2002] have shown how metrics based on the measurement IACC correlated well with subjective scores collected on envelopment, apparent source width and depth in the reproduced sound environment. This was also shown by Choisel and Wickelmaier as discussed above.

Based upon a series of experiments which concluded that the perception of LEV in reproduced sound was influenced by the overall playback level and the level and angular distribution of late arriving sound, Soulodre et al [2003] proposed a metric for the measurement of perceived LEV called GS_{perc}. This metric was an improvement upon a metric which they had previously developed called LG (Lateral Gain) and was a combination of a measure of the relative level of late energy and a spatial metric, based upon LF (the authors suggest that the LF could also be represented by IACC).

More recently this author [Conetta *et al*, 2007][Conetta, 2007], Dewhirst [2008] and George [2009] developed separate regression models for the prediction of perceived envelopment in the context of reproduced sound. However in the listening tests used to characterise the perception of envelopment, they used a definition of perceived envelopment more appropriate for reproduced sound, as discussed in section 2.2.2.3.

Over three experiments, this author [Conetta, 2007] employed an 8-channel surround system and created a range of audio scenes exhibiting different levels of envelopment synthetically, using either anechoic (when investigating Direct Envelopment) or highly reverberant (when investigating Indirect Envelopment) mono speech sources. The studies revealed that the perception of envelopment was predominantly influenced by a number of different factors such as soundfield density (i.e. number of mono speech sources), inter-channel correlation, ensemble or scene width, the location or position of the sources, playback level and frequency content. Using this information regression models were created employing metrics based upon IACC, RMS level, Karhünen-Loeve Transform (KLT) and Entropy. The performance of each metric in the models is summarised, in table 4.3, in terms of their standardised Beta coefficients, which allows the relative importance of the metrics in the model to be compared [Field, 2005].

Metric	Description	Direct Envelopment (Experiment 1)	Direct Envelopment (Experiment 2)	Indirect Envelopment (Experiment 3)
IACC0	The mean IACC value calculated across 22 frequency bands (150Hz-10kHz) from both ear signals of a head and torso simulator with a 0˚ head orientation.	-0.383	-0.317	-0.38
IACC0*IACC90	The product of the IACC0 and IACC90 values above. IACC90 is the mean IACC value calculated across 22 frequency bands (150Hz-10kHz) from both ear signals of a head and torso simulator with a 90˚head orientation.	-0.269	-0.256	-0.31
CardKLT	The contribution in percent of the first eigenvector from a Karhunen-Loeve Transform (KLT) decomposition of four cardioid microphones placed at the listening position and facing in the following directions: 0˚, 90˚, 180˚ and 270˚.	-0.306	-0.254	-0.315
EntropyL	Entropy of the left ear signal of a head and torso simulator with a 0˚ head orientation.	0.413	0.336	0.27
TotEnergy	Calculated root mean square of the pressure value measured by a pressure microphone.	0.294	0.254	-
	Correlation (r)	0.96	0.94	0.89
	RMSE (%)	5.94%	8.41%	11.54%

Table 4.3 The performance of the three models created by Conetta to predict perceived envelopment. Including a description of each metric and their Beta coefficients.

The Beta coefficients show that the most important metrics were 'IACC0' and 'EntropyL'. 'IACC0' being important in these models further supports the research of Choisel and Wickelmaier and others, discussed above.

In the results of these studies it was observed that perceived envelopment increased when the number of voice sources used in the audio scenes was increased. Entropy, a measure of the information in a signal, was included to measure the change in soundfield density. The decision to measure the entropy of the left ear signal was arbitrary as it was felt that the density of the soundfield would be equal all around the listener.

The use of a multiplicative metric, 'IACC0*IACC90' was inspired by George at al [2006] where it was employed in the prediction of frontal spatial fidelity (FSF) and surround spatial fidelity (SSF). Based upon results produced by Hands [2004], George hypothesised that interactions between metrics might enhance the prediction power of his models. Hence it was proposed that 'IACC0*IACC90' might have a good correlation with perceived envelopment because it combined an assessment of the IACC along the median plane and frontal plane and therefore provided more information about the correlation of audio scene in 360° around the listening position.

The results of the studies also showed that the perceived envelopment increased when the voice sources in the synthesised audio scenes were uncorrelated. Similarly Blauert [2001] found that a listener's perception of spatial impression was altered by inter-loudspeaker coherence. 'CardKLT' was employed to measure the correlation between the front, rear, left and right segments of the scene using the Karhünen-Loeve Transform (KLT).

Based upon the work of Soulodre *et al* [2003] TotEnergy was employed as a measure of playback level, however it is a relatively crude metric and most often has the lowest importance in the models.

The use of synthetic audio scenes in this study reduces its ecological value. However it has allowed a number of variables which affect the perception of envelopment in the reproduced sound environment to be identified. This has informed the development of metrics, some of which were used in the research already discussed above.

Dewhirst [2008], expanding upon this work, achieved similar results by incorporating metrics based upon Interaural Time Difference (ITD) and Interaural Level Difference (ILD) into his envelopment prediction models.

George [2009] developed a regression model for the prediction of the perceived envelopment arising from a variety of different audio recordings (i.e. mono, 2-channel stereo and 5-channel), from different genres, which had been bandwidth limited, dowmixed or coded using low bit-rate multichannel audio codecs. A total of 71 different metrics, were employed to calibrate the model using PLS regression. The performances of the metrics selected for this model, are described in terms of their standardised Beta coefficients in table 4.4.

Metric	Description	Beta coefficients
R_{raw}	Spectral rolloff of a 1.0 downmix of the audio recording.	0.19
ASD	Area of sound distribution across the listening area, calculated using SAT (spatial analyser tool) [Jiao, 2007].	0.25
I_{OB60}_I_{OB150}	Multiplication of the mean IACC value calculated at 500Hz, 1kHz and 2kHz frequency bands with a 60° and 150° virtual dummy head orientation.	-0.28
KLT_{V1}_I_{OB60}	Multiplication of KLT_{V1}, the contribution in percent of the first eigenvector from a Karhunen-Loeve Transform (KLT) calculated from the audio recording, and the mean IACC value calculated at 500Hz, 1kHz and 2kHz frequency bands with a 60° virtual dummy head orientation.	-0.29
KLT_{V1}_CCA_{log}	Multiplication of KLT_{V1}, and CCA_{log}, the logarithm of the centroid of coverage angle around the listening position calculated using SAT.	0.22
Correlation (r) (Calibration/Validation)		0.91/0.90
RMSE (%) (Calibration/Validation)		8.15%/7.75%

Table 4.4 The performance in calibration and validation of George's model to predict perceived envelopment. Including a description of each metric and their Beta coefficients.

Spectral rolloff (R_{raw}) was significant in the model. This metric was designed to measure the timbral characteristics of the stimuli and, since it had the lowest importance, its significance in the model might seem puzzling. However, it measures the loss of high frequency content in stimuli that have been bandwidth limited or processed using low-bit multichannel audio codecs and George suggests that this high frequency loss may have influenced the listeners' perception of envelopment.

Area of sound distribution (ASD) was designed to measure the extent of the distribution of sound around the listener and could be considered as a measure of ensemble width or scene width.

The multiplication of the mean IACC values calculated at 60° and 150° had high importance in the model. A similar metric was used by this author. However George does not discuss why the interaction of IACC measured at these particular angles was selected during the model calculation process.

'KLT_{V1}_I_{OB60}' and 'KLT_{V1}_CCA_{log}' are both metrics with an interaction with KLT. A metric based upon KLT was also significant in the models created by this author where it was employed to measure the correlation of the audio scene. 'CCA_{log}' is a metric similar to ASD, designed as a measure of ensemble width or scene width.

4.1.4 Spatial attribute metrics: summary and conclusions

In their study Choisel and Wickelmaier investigated the correlation of a selection of metrics to timbral and spatial attributes which listeners had identified and quantified in various audio recordings. In particular this established that metrics based upon the IACC show good correlation with spatial attributes of the reproduced sound environment such as perceived width, envelopment and spaciousness.

Various models of localisation have been developed and these predominantly rely upon measuring the interaural time difference and interaural level difference.

Research conducted in the context of concert hall acoustics and reproduced sound has also shown that metrics based upon the IACC correlate well with perceived envelopment and width.

Soulodre *et al* proposed a metric for the prediction of perceived envelopment, which combined measurements of the relative level and the angular distribution of late energy. This author, Dewhirst and George developed regression models which correlated well with the subjective scores collected from their listening tests. These models used metrics based upon the IACC, KLT, Entropy, ITD and ILD, and also included metrics for scene or ensemble width and timbral characteristics. In these models multiplicative metrics (where two metrics are multiplied together) were also used to good effect.

The metrics discussed in this section will be used as inspiration for the choice of metrics employed to calibrate the QESTRAL model for the objective evaluation of spatial quality.

4.2 Metrics used in spatial sound quality models

To provide insight into how metrics similar to those discussed in the previous section could be used in the QESTRAL model, the metrics employed by Choi *et al* [2008] and George [2009] in their spatial quality models (previously discussed in sections 2.3.3 and 2.3.4) are investigated.

4.2.1 Metrics used in the model created by Choi *et al*

Choi *et al* [2008] proposed a multichannel addition to the PEAQ standard. Their model employed metrics for both timbral and spatial attributes, using ten MOVs from the basic version of PEAQ with three additional spatial metrics, to predict degradations of BAQ created by low bit-rate multichannel audio codecs. The metrics were calculated from binaural signals synthesised from 5.1 recordings. The model showed good correlation with the subjective database. Table 4.5 summarises the performance (in terms of correlation (r)) and gives a basic description of each metric used in the model (NB. These are described further in ITU-R BS.1387 [2001]).

All of the metrics were negatively correlated, indicating that they had an inverse relationship to BAQ. The metrics used to measure the spatial characteristics of the DUT showed the highest correlation, however it is not clear how important these metrics were in the model, because the Beta coefficients for each metric were not published.

Although they do not reveal the role that each spatial metric plays in the model, the authors indicate that they employed interaural level difference distortion (ILDD) and interaural time difference distortion (ITDD) to measure changes to perceived source locations, and interaural cross-correlation distortion (IACCD) to measure changes to the apparent source width. However as discussed, this study was limited to the evaluation of multichannel audio codecs, and therefore it is unknown how well these metrics would correlate with subjective scores collected from a study evaluating a wider selection of processes, such as that proposed for this research project. Interestingly in their discussion of the model they hypothesise that metrics calculated from different head rotations might also be

useful in future calibrations of their model. This idea was also employed by this author and George in their envelopment models discussed above, and might also be useful in the QESTRAL model.

	Metric	Description	Approx. Correlation (r)
Timbral	ADB	Averaged distortion block; ratio of total distortion to total number of distorted blocks.	-0.67
	NMRtotB	Logarithm of averaged total noise to masker energy ratio.	-0.51
	NLoudB	Averaged noise loudness.	-0.51
	AModDif1B	Averaged modulation difference.	-0.45
	WModDif1B	Windowed averaged modulation difference.	-0.43
	RDF	Relative fraction of frames with significant noise component.	-0.42
	EHS	Harmonic structure of error.	-0.42
	AModDif2B	Averaged modulation difference with emphasis on modulation changes where reference contains little modulations.	-0.36
	AvgBwRef	Bandwidth of reference signal.	-0.05
	AvgBwTst	Bandwidth of signal under test.	-0.01
Spatial	ILDD	Difference between source directions of signal under test and original signal due to ILD. Computed for high-frequency sounds (above 2500 Hz).	-0.78
	IACCD	Difference between apparent source widths of signal under test and original signal due to IACC difference.	-0.62
	ITDD	Difference between source directions of signal under test and original signal due to ITD. Computed for low-frequency sounds (below 1500 Hz).	-0.61

Table 4.5 Individual correlation (r) with BAQ of the metrics used by Choi *et al*.

4.2.2 Metrics used in the models created by George *et al*

Although developing models for measuring spatial characteristics, George [2009] (and [George *et al*, 2006a/b]) employed metrics for the timbral characteristics of the audio scene. In the results of the listening tests used to characterise the perception of spatial fidelity George's models, Zielinski *et al* [2005b] observed that stimuli which had been bandwidth limited not only degraded the perceived timbral fidelity but also degraded the perceived FSF and SSF. Similarly they noticed that downmixing stimuli degraded the perceived FSF, SSF and also timbral fidelity.

George applied transformation functions to each of the metrics he used to improve their individual correlation to FSF and SSF. He also included multiplicative metrics in his models. Once he had generated a wide selection of metrics (55 in total) George calibrated his models using regression analysis. Ten of the 55 metrics were found to make a statistically significant contribution to the models he developed. Table 4.6 describes the performance of these significant metrics in terms of their standardised Beta coefficients.

Model	Metric	Description	Beta coefficients
FSF	lbb0	Broadband IACC with a 0° virtual dummy head orientation	-0.27554
	COH	Centroid of spectral coherence	0.35164
	I0	Maximum IACC value calculated at 500Hz, 1kHz and 2kHz frequency bands with a 0° virtual dummy head orientation	-0.2225
	I150	Maximum IACC value calculated at 500Hz, 1kHz and 2kHz frequency bands with a 150° virtual dummy head orientation	-0.21139
	I180	Maximum IACC value calculated at 500Hz, 1kHz and 2kHz frequency bands with a 180° virtual dummy head orientation	-0.15083
	BFIbb90	interaction Ibb90 × BFratio Broadband IACC at 0o head position. Back-to-front energy ratio Broadband IACC with a 90° virtual dummy head orientation	-0.16213
SSF	Rrsc	Rescaled average spectral roll-off	0.19951
	COH	Centroid of spectral coherence	0.16721
	I60	Maximum IACC value calculated at 500Hz, 1kHz and 2kHz frequency bands with a 60° virtual dummy head orientation	-0.2635
	I90	Maximum IACC value calculated at 500Hz, 1kHz and 2kHz frequency bands with a 90° virtual dummy head orientation	-0.21927
	I120	Maximum IACC value calculated at 500Hz, 1kHz and 2kHz frequency bands with a 120° virtual dummy head orientation	-0.26795
	I180	Maximum IACC value calculated at 500Hz, 1kHz and 2kHz frequency bands with a 180° virtual dummy head orientation	-0.23674

Table 4.6 The performance of the models created by George *et al* to predict perceived FSF and SSF. Including a description of metrics used in each and their Beta coefficients.

IACC based metrics were the most useful metrics in both models. George employed both broadband (i.e. measurements are taken across the full bandwidth of the signal) and octave band (i.e. similar to the method discussed by Beranek [1996] measurements are taken only at 500Hz, 1kHz, 2kHz) IACC measured with a virtual dummy head rotated to angles at 10° intervals between 0° and 180°. The continued significance of these metrics in these models suggests their importance for a model predicting spatial quality.

Interestingly 'COH' was the most important metric for the prediction of FSF and was also significant for the prediction of SSF. Based upon informal studies George discovered that this metric demonstrated a higher correlation with bandwidth limited stimuli than downmixed stimuli. George suggested that bandwidth limitation, particularly at high frequencies, impaired the perceived distance of the sources or audio scene making them appear more distant. This idea is supported by Moore [2003] who has shown that the perceived distance of an auditory event is related to its frequency content. George expanded upon this further, suggesting that COH was so important in the prediction of FSF because the programme material he used contained predominantly foreground scene sources (NB. Each item of programme material was either F-F or F-B scene type). Therefore, when the listeners were asked to assess the FSF of the bandwidth limited stimuli, a change in the perceived

distance of the sources or the scene was clearly noticeable between the reference and stimulus. However, as previously discussed, these studies were limited to the evaluation of bandwidth limitation and downmix processes, and therefore the importance of COH to a wider study should be considered with caution. Its importance could have been inflated because approximately half of the data used to calibrate the models was collected using bandwidth limited stimuli. George acknowledges that this limits the validity and generalisability of the models and hence in a wider study such as this project, this metric may not be as important. However it is accepted from George's results that metrics to measure the timbral characteristics of the audio scene could be useful for the objective evaluation of spatial quality, particularly if timbral quality is shown to have an influence on a listener's perception of spatial quality (as discussed in section 2.1.1).

4.2.3 Spatial quality model metrics: summary and conclusions

Choi *et al* employed metrics to measure both timbral and spatial characteristics to predict degradations in BAQ imparted by low bit-rate multichannel audio codecs to a selection of 5.1 multichannel recordings. The metrics measuring spatial characteristics (ILDD, ITDD and IACCD) were shown to have the highest independent correlation to the subjective scores, which implies that they are important metrics for the measurement of spatial quality. However, as discussed, this model was limited to the evaluation of multichannel audio codecs, so it was unknown how well these metrics would correlate with subjective scores collected from a study evaluating a wider number of spatial audio processes, as will be the case with this project.

George *et al* created models for the prediction of frontal spatial fidelity (FSF) and surround spatial fidelity (SSF) in which he employed a wide selection of metrics for both spatial characteristics and timbral characteristics. George created his models using an iterative approach to calibration and found IACC based metrics were the most useful metrics in both models. He employed both broadband (i.e. taken across the full bandwidth of the signal) and octave band measurements, and interactions between different head orientation angles. The significance of metrics based upon the measurement of IACC indicates their potential importance for a model predicting spatial quality. George also found that COH, a metric designed to measure changes to timbral quality, made a significant contribution to the prediction of FSF and SSF. However it is believed that this metric's importance was inflated because approximately half of the data used to calibrate the models were collected using bandwidth limited stimuli. Hence in a wider study such as this project, this metric may not be as important. However metrics for the timbral characteristics of the audio scene could be useful for the objective evaluation of spatial quality.

4.3 Summary and conclusions

This chapter reviewed objective metrics currently used to measure individual spatial attributes in reproduced sound and existing spatial quality models. The aim of this review was to identify suitable metrics that could be employed in the QESTRAL model to measure changes to spatial quality that are created by the SAPs.

In their study Choisel and Wickelmaier described the correlation of a number of different metrics designed to measure both timbral and spatial characteristics. In particular this identified that metrics based upon the measurement of IACC show good correlation with spatial attributes in the reproduced sound environment such as perceived width, envelopment and spaciousness.

A number of models for localisation have been developed and these predominantly rely upon measuring the interaural time difference and interaural level difference.

Research conducted in the context of concert hall acoustics and reproduced sound has also shown that metrics based upon the measurement of IACC correlate well with perceived envelopment and width. A number of metrics have been shown to correlate well with perceived changes to envelopment. Soulodre et al proposed a metric which combined measurements of the relative level and the angular distribution of late energy. Conetta, Dewhirst and George used metrics based upon measurements of the IACC, KLT, Entropy, ITD, ILD and also included metrics to measure scene or ensemble width and the timbral characteristics. In these models multiplicative metrics were also used to good effect.

Choi *et al* employed metrics to measure both timbral and spatial characteristics to predict degradations of BAQ created by low bit-rate multichannel audio codecs to a selection of 5.1 multichannel recordings. The metrics ILDD, ITDD and IACCD measuring spatial characteristics were shown to have the highest independent correlation to the subjective scores.

George *et al* created models for the prediction of frontal spatial fidelity (FSF) and surround spatial fidelity (SSF) in which he employed a wide selection of metrics for both spatial characteristics and timbral characteristics. George found that IACC based metrics were the most useful metrics in both models which indicates their potential importance for a model predicting spatial quality. George also found that a metric designed to measure changes to timbral quality made a significant contribution to the prediction of FSF and SSF. However the importance of this metric in situations where bandlimiting is less common might be lower. Nevertheless it suggests that metrics designed to measure the timbral characteristics of the audio scene could be useful for the objective evaluation of spatial quality.

Chapter 5 – Identifying a listening test method for the evaluation of spatial quality

Chapter 4 identified metrics for SAP-induced changes to spatial quality, that could potentially be employed in the QESTRAL model.

This chapter aims to identify a suitable listening test method for evaluating a wide range of SAPs that impair the perception of spatial quality. This begins with an overview of existing international standards for the subjective assessment of audio quality, to determine their suitability. However a number of limitations to these standards are identified which motivates the development of a modified listening test method for assessing spatial quality. The development and design of this method are discussed.

5.1 Listening test standards for audio quality

Formal subjective testing is currently regarded as the most reliable method for the evaluation of audio quality [Zielinski *et al*, 2008]. This research project requires a suitable method for investigating spatial quality, and so existing standards for the subjective assessment of audio quality were studied.

The International Telecommunication Union (ITU) has developed and standardised listening test methods for the evaluation of audio quality that are used extensively in research. These are BS.1116-1 [1997], BS.1534 [2001] and BS.800 [1996]. BS.800 was developed for the analysis of speech quality, which is unrelated to this research and is therefore not discussed. BS.1116-1 and BS.1534 were both developed for the evaluation of full bandwidth audio material.

5.1.1 ITU-R BS.1116-1

ITU-R BS.1116 [1997] was designed for the assessment of small impairments to high quality audio (principally resulting from low bit-rate coding schemes). The test method presents the listener with three stimuli (audio recordings), A, B and C, which they can switch between at will. Stimulus A represents an unprocessed signified reference condition and B and C are randomly assigned to represent an unsignified reference, commonly referred to as the ‘hidden reference’, and a processed recording, often called the ‘test condition’. This method comprises a ‘double-blind, triple-stimulus with hidden reference’ test. During the test listeners are asked to compare stimuli B and C with reference stimulus A and, on a continuous grading scale, rate their sound quality. A graphical user interface (GUI) is often employed to present the test. An example of a typical GUI is shown in figure 5.1.

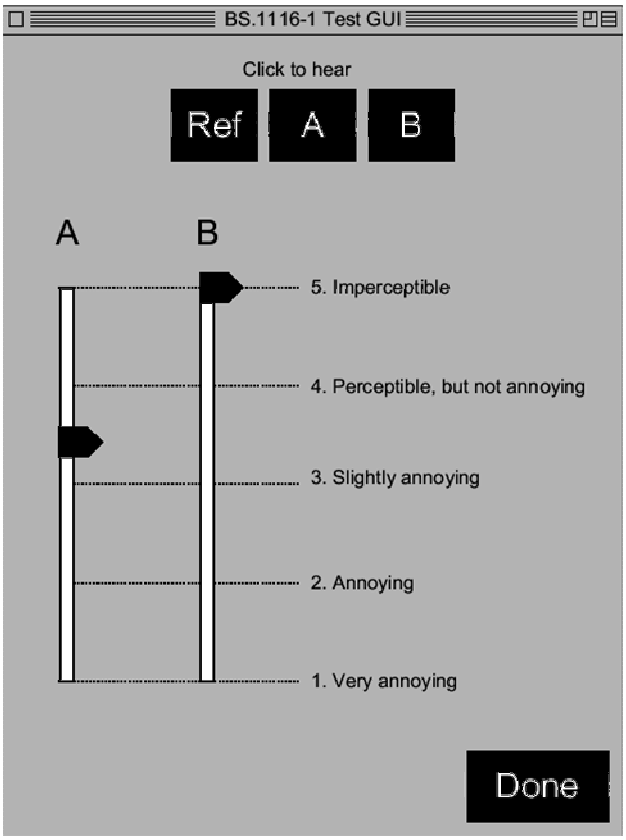


Fig. 5.1 An example of an ITU-R BS.1116-1 GUI [Martin, 2006].

The grading scale is continuous and includes labels to describe or anchor the scale intervals, derived from the ITU-R five-grade impairment scale [ITU-R BS.1284, 1998] (see Table 5.1). Listeners are explicitly asked to give the stimulus they identify as the hidden reference the highest score, corresponding to the scale label 'imperceptible'. They can use the rest of the scale to judge the quality of the test condition.

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

Table 5.1 ITU-R five-grade impairment scale [ITU-R BS.1116-1, 1997].

Most commonly this method is used to evaluate BAQ, a global attribute used to describe any and all differences between the reference and test condition (as described in section 2.1). This is likely to incorporate assessments of both the timbral quality and spatial quality together. However, as mentioned in section 2.1.1, the standard can also be used for the independent assessment of attributes similar to spatial quality such as, in multichannel audio systems, the attributes 'front image quality' and 'impression of surround quality'.

The method recommends that listeners should be selected from a panel of expert listeners with normal hearing and that each listener should be fully trained in the aims of the test prior to the

experiment. A test consists of two parts; a familiarisation stage and a grading stage. The familiarisation stage allows the listeners to familiarise themselves with the stimuli under investigation, the assessment scales and the user interface and test environment. No results are collected during this stage. After they have completed the familiarisation stage the listeners then commence the grading stage, from which the results are collected, under formal test conditions. To limit the effects of fatigue the standard recommends that a maximum of 10-15 comparisons (test pages) be used per listening session (using a minimum of 5 stimuli) and that the session should last no longer than 30 minutes in total. In addition to the experimental design, BS.1116-1 also recommends target specifications for the design of listening rooms suitable to achieve the critical listening conditions required. ITU-R BS.1116-1 is a useful method for testing small audible differences caused by audio codecs. However it is an inefficient method, and potentially inaccurate, if many stimuli are to be assessed that generally exhibit larger differences.

5.1.2 ITU-R BS.1534 (MUSHRA)

To allow the assessment of a larger number of stimuli more efficiently ITU-R BS.1534 (MUSHRA) [2001] was designed jointly by the ITU and EBU for the assessment of low and intermediate quality audio codecs that would fall into the lower half of the impairment scale used by ITU-R Recommendation BS.1116-1. The abbreviation MUSHRA stands for Multi Stimulus test with Hidden Reference and Anchors.

The method presents the listener with a number of stimuli for assessment. The listener is asked to compare these processed stimuli against an unprocessed signified reference stimulus using a continuous 100 point grading scale. The scale has five labels which describe intervals corresponding to different levels of perceived quality and is often known as a continuous quality scale (CQS). A typical example of a MUSHRA GUI is depicted in figure 5.2. The stimuli are synchronously looped and the listener can switch between the stimuli as many times as they wish. Amongst the stimuli at least one hidden reference is included. The listener is informed that one or more hidden reference stimuli are present in the test, and that these should be given a grade of 100. The standard also recommends that at least one hidden (or indirect) anchor should be included. The first choice for the hidden anchor should be a low-pass filtered version of the reference stimulus with a bandwidth of 3.5kHz. If more anchors are required, further recommendations are incorporated in the standard. The additional anchors are intended to provide a context to the test by giving an indication of how the test conditions compare to well-known audio quality levels.

MUSHRA is usually employed for the assessment of BAQ but, similarly to BS.1116-1, it can also be used to assess attributes similar to spatial quality.

The same protocol for the selection and training of listeners, and running of tests, recommended in BS.1116-1 is recommended for MUSHRA tests. The standard also suggests that there be a maximum of 15 stimuli per page and as a general rule an experiment should consist of a

minimum of 5 test pages and a maximum number of 1.5 times the number of test stimuli. Each stimulus should be a maximum of 20 seconds long in order to reduce fatigue.

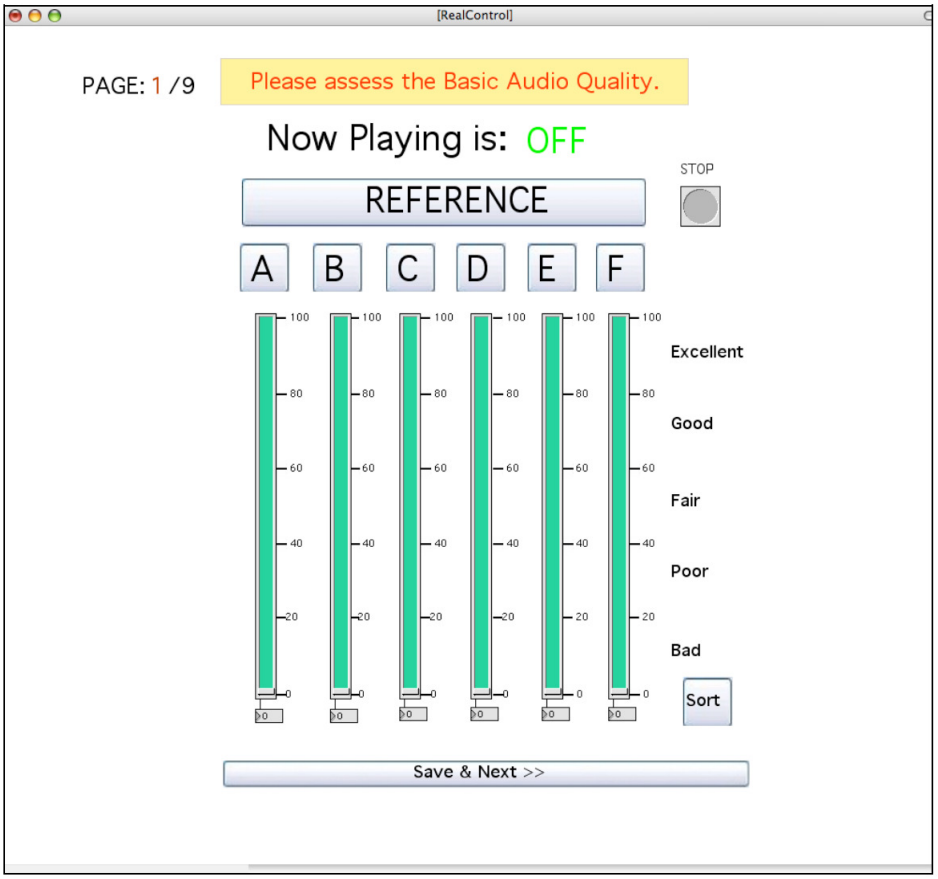


Fig. 5.2 An example of a typical ITU-R BS.1534 GUI [Jiao *et al*, 2007].

5.1.3 Listening test standards: summary and conclusions

BS.1116-1 was designed for the detection of small impairments between stimuli and the level of annoyance they create, for this purpose it is limited to the evaluation of a single test condition per test page. As it is desirable to collect subjective data, using the BS.1116-1 method would be inefficient and very time consuming and therefore it is not suitable for use in this research project. By comparison MUSHRA, which is a multistimulus test, allows several stimuli to be compared side-by-side. This is a much more efficient way of collecting the amount of subjective data required for this project.

However Zielinski *et al* have noted that data collected from experiments employing the MUSHRA method suffer from biasing. Using biased data to calibrate a model would potentially limit its validity and generalisability. The different types of, causes of and possible solutions to biases known to affect listening tests are discussed in the following section.

5.2 Biases affecting audio quality listening tests

Biases are systematic errors which influence the mapping process that a listener uses to transfer their opinion of a stimulus to the test scale. Bias can affect the scores of every listener, revealing itself, for example, as a continuous shift in the scores or an exaggeration of the difference between perceptually similar stimuli [Zielinski *et al*, 2008]. Random errors are common in listening tests, and can often occur through isolated mistakes that a listener makes during the test, such as forgetting to grade one stimulus on a page. These errors are easily identified and are often removed during statistical analysis of the results. However because bias reveals itself as systematic errors affecting all of the results it is not easy to identify and is difficult to remove once it has "infected" the data [Zielinski *et al*, 2008].

It is desirable for sake of the validity of the QESTRAL model to minimise or reduce the influence of bias on the data collected for its calibration. Zielinski *et al* [2008] and Bech and Zacharov [2006] provide an overview of various biases that can affect audio quality listening tests.

5.2.1 Biases affecting MUSHRA and multistimulus tests

There are a number of ways that bias in MUSHRA and other multistimulus tests can be created. These are summarised in tables 5.2 and 5.3. Six types of bias are known to affect the results collected using multistimulus tests, four of which have been shown to influence the results collected from tests using the MUSHRA method (see Table 5.2).

5.2.1.1 Stimulus spacing bias

Stimulus spacing bias (Fig 5.3) can be created when the perceptual distribution range of the stimuli under test is skewed by the dominance of perceptually similar stimuli on a particular page of the test. Listeners have been shown to over-estimate the differences between the similar stimuli while under-estimating the differences between the other stimuli, leading to a skewed usage of the scale. Although rank order information is preserved an interpretation of the relative differences between the stimuli is unreliable.

Zielinski *et al* [2007a] showed how this bias could occur in MUSHRA tests. They observed that when additional stimuli of low quality were added to the stimulus set, the other stimuli were scored higher (and the scores were positively skewed). This was because the differences in perceived quality between the additional lower quality stimuli were exaggerated or over-estimated. However, because the perceptual distribution range had not expanded, this reduced the scale area over which the higher quality stimuli could be scored and hence the differences between them were under-estimated. Conversely the distribution of the scores was negatively skewed when additional stimuli of high quality were added to the stimulus set.

Zielinski *et al* [2008] suggested that stimulus spacing bias can be reduced by selecting stimuli that are perceptually equally spaced across the range of the scale. This might be possible for the

stimuli used in an entire test but in practice it is difficult to achieve for every page of the test. Although randomising the presentation order of the stimuli might help, if enough tests are conducted. It might be possible to diagnose whether this bias is present in the data for a particular listener by including an anchor for the middle of the scale and comparing the assessment score for this anchor between different test pages.

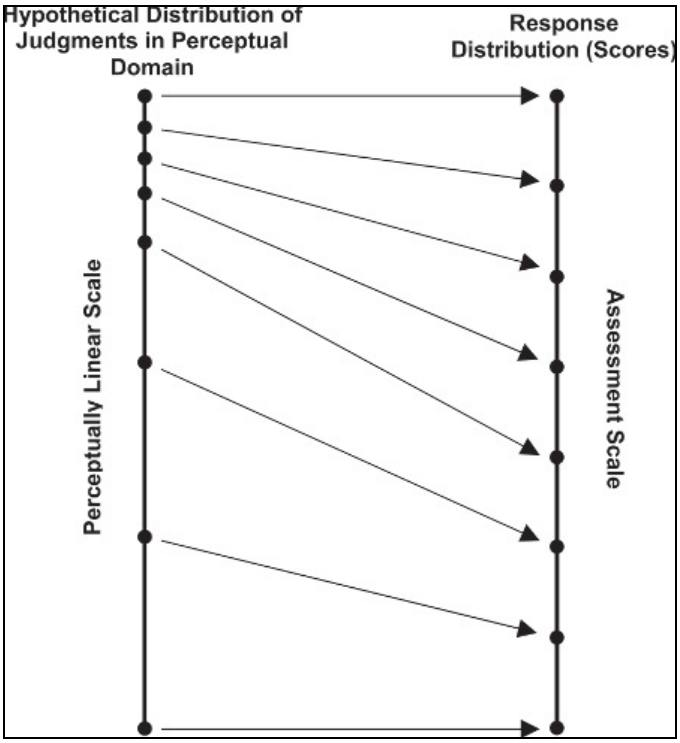


Fig 5.3 The effect of stimulus spacing bias [Zielinski *et al*, 2008].

5.2.1.2 Range-equalising bias

Also known as the 'rubber-ruler effect', range-equalising bias (Fig 5.4) is created when the range of the stimuli is extended or decreased (eg. by the addition or removal stimuli at the top or bottom of the range) between pages of the test. It occurs because listeners often like to use the full range of the scale when assessing a large number of stimuli. The test scale is fixed so the listeners adapt their usage of the scale to accommodate the new stimuli, and hence the scores are comparatively squashed together if the range is extended or spread out if the range is decreased. Therefore the scores may span the entire range of the scale regardless of their actual perceptual range. Similarly to stimulus spacing bias, although rank order information is preserved, range-equalising bias makes an interpretation of the relative differences between the stimuli unreliable.

Zielinski *et al* [2007a] revealed that range-equalising bias could occur in MUSHRA tests when additional stimuli of lower quality than the suggested low anchor (3.5kHz low-pass filtered) were added to the stimulus range. They showed that this resulted in the scores for all stimuli being "pushed up". In the MUSHRA method this occurs because the top of the scale is fixed, but to

accommodate the extra low quality stimuli at the bottom of the range, the scores for the rest of the stimuli are "pushed up".

The occurrence of range-equalising bias in MUSHRA or multistimulus tests can be reduced by using direct (signified) or indirect (hidden) anchoring, to standardise the perceptual range of the scale [Zielinski *et al*, 2008].

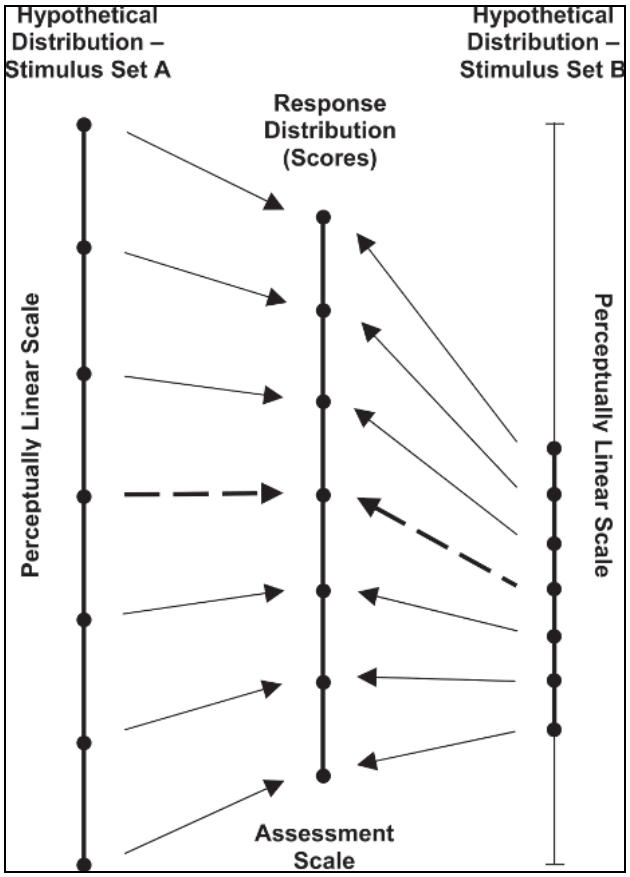


Fig 5.4 The effect of range equalising bias [Zielinski *et al*, 2008].

5.2.1.3 Bias due to perceptually non-linear scale

Although the continuous quality scale of the MUSHRA GUI is numerically linear there is evidence to suggest that the labels employed to describe it are neither perceptually or semantically linear (Fig 5.5). This can lead to non-linear responses from the listeners [Zielinski *et al*, 2007b] and therefore an interpretation of the relative differences between the stimuli may become unreliable. Zielinski *et al* [2008] have also shown how the interpretation of the labels can differ between languages. They suggest that this type of bias can be reduced by removing the labels or by employing a polarity scale, whereby only the top and bottom of the scale are labelled with opposing descriptors (eg. excellent and bad).

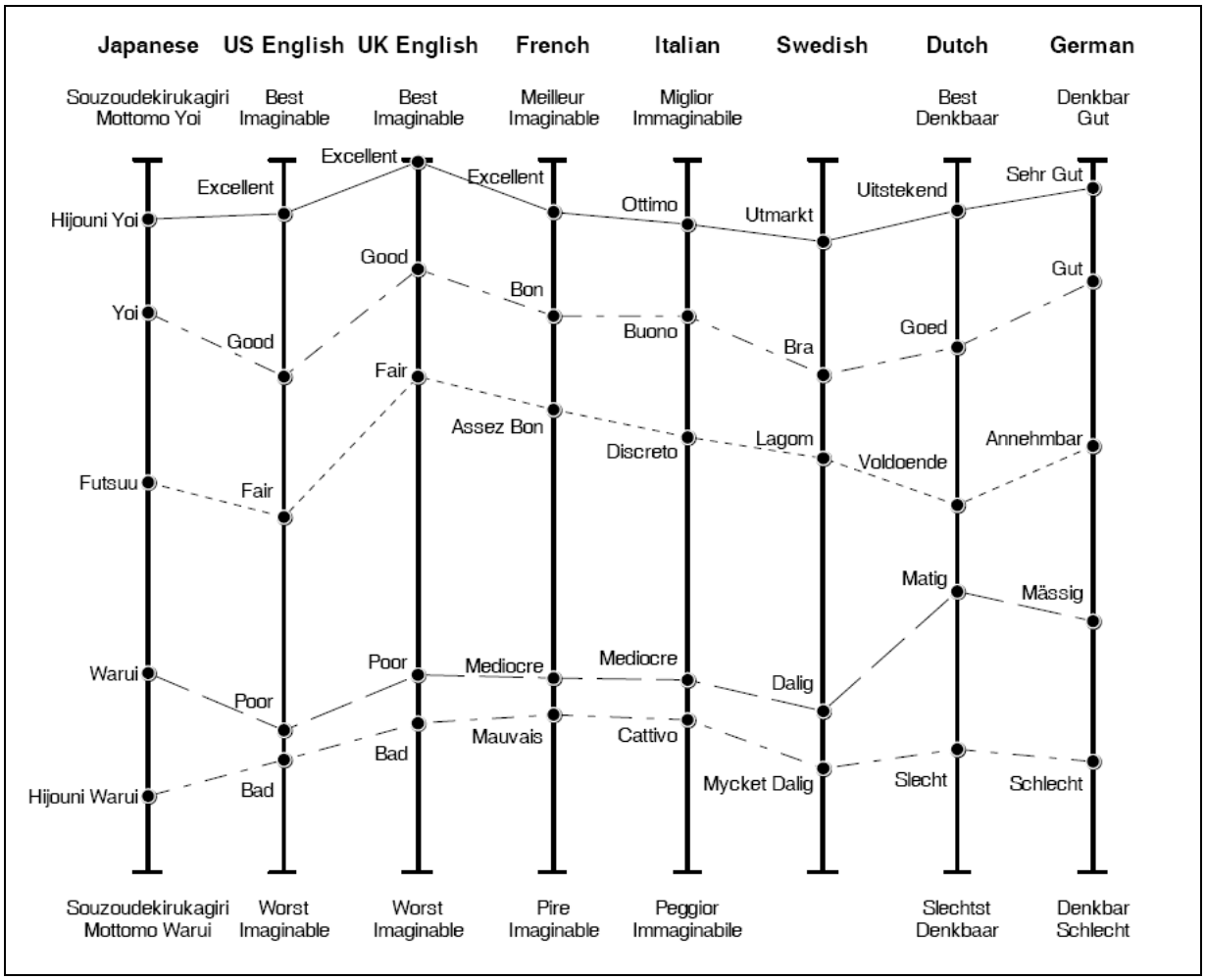


Fig 5.5 A comparison, between languages, of the interpretation of the perceptual weighting of the MUSHRA GUI CQS labels [Zielinski *et al*, 2008].

5.2.1.4 Interface bias

Interface bias is caused by the ergonomics of the MUSHRA GUI. It is sometimes known as quantisation bias due to the visual appearance of the scores clustered around markings on the scale such as markings, numbers or labels (see Fig 5.6) [Zielinski *et al*, 2007b]. Although the rank order information is preserved it makes an interpretation of the relative differences between the stimuli unreliable. Zielinski *et al* [2008] indicate that interface bias can be avoided by removing the markings, numbers or labels from the scale, or reduced by using a large population of listeners.

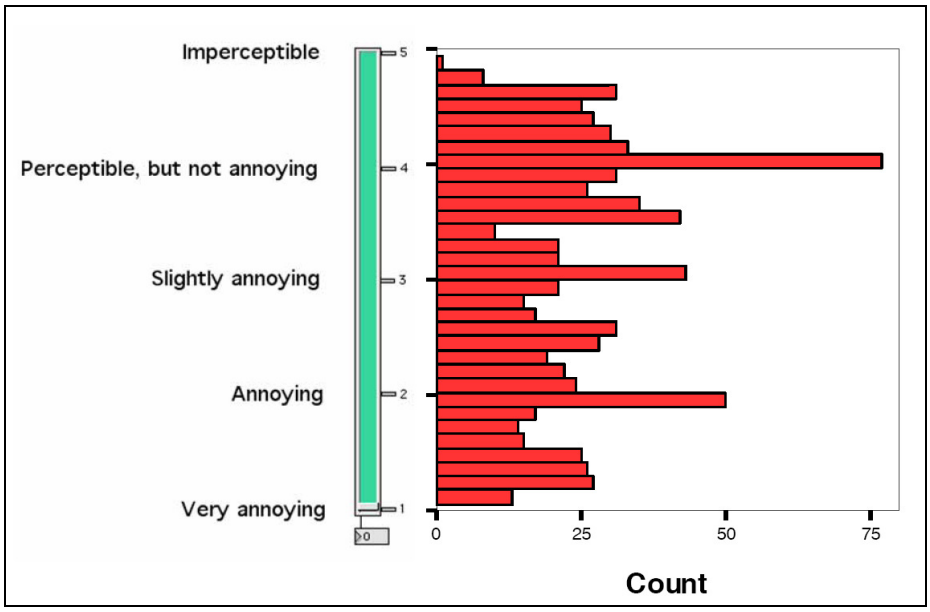


Fig 5.6 Histogram of scores exhibiting interface bias caused by the tick marks on the BS.1116-1 ITU impairment scale [Zielinski *et al*, 2007b].

Biases known to affect MUSHRA tests	Manifestations	Potential implications	Examples of bias reduction
Stimulus spacing bias	Subjects use the entire range of the scale, equalising the differences between the stimuli, regardless of the perceptual difference.	Distorted information about the genuine differences between the stimuli. Information about rank order is preserved.	Select stimuli that are perceptually equally spaced. Randomisation
Range equalising bias – "Rubber ruler" effect.	Subjects use the entire range of the scale, regardless of the perceptual range of the stimuli.	Cannot assess absolute quality. Information about rank order is preserved.	Use direct or indirect anchoring
Bias due to perceptually non-linear scale	Non-linear effect in the distribution of the scores.	Distorted information about genuine differences between the stimuli. Information about rank order is preserved.	Use a label-free scale or only label the top and bottom of the scale.
Interface bias	Quantisation effect in the distribution of the scores.	Distorted information about genuine scores. Only rank order	Remove labels, numbers or markings from the interface. Use a large population of listeners.

Table 5.2 Biases affecting MUSHRA method (adapted from Zielinski *et al* [2008]).

Biases known to affect multistimulus tests	Manifestations	Potential implications	Examples of bias reduction
Stimulus frequency bias	Expansion effect in scores.	Overestimated differences between most frequent stimuli.	Use a balanced design (avoid presenting perceptually similar or identical stimuli more often than other stimuli).
Centring bias	Systematic shift of all the scores.	Cannot assess absolute quality. Rank order preserved	Use direct or indirect anchoring.

Table 5.3 Biases affecting multistimulus tests (adapted from Zielinski *et al* [2008]).

5.2.1.5 Stimulus frequency bias

Similarly to stimulus spacing bias, stimulus frequency bias (Fig 5.7) occurs when there are a large number of perceptually very similar or identical stimuli on a test page. Rather than give these stimuli the same score the listeners over-estimate the perceptual differences between them, spreading the scores out on the scale. Although rank order information is mostly preserved it makes interpreting the relative differences between the stimuli unreliable. This problem can be removed by employing a balanced test design in which very similar or identical stimuli are not presented more than once per test page [Zielinski *et al*, 2008].

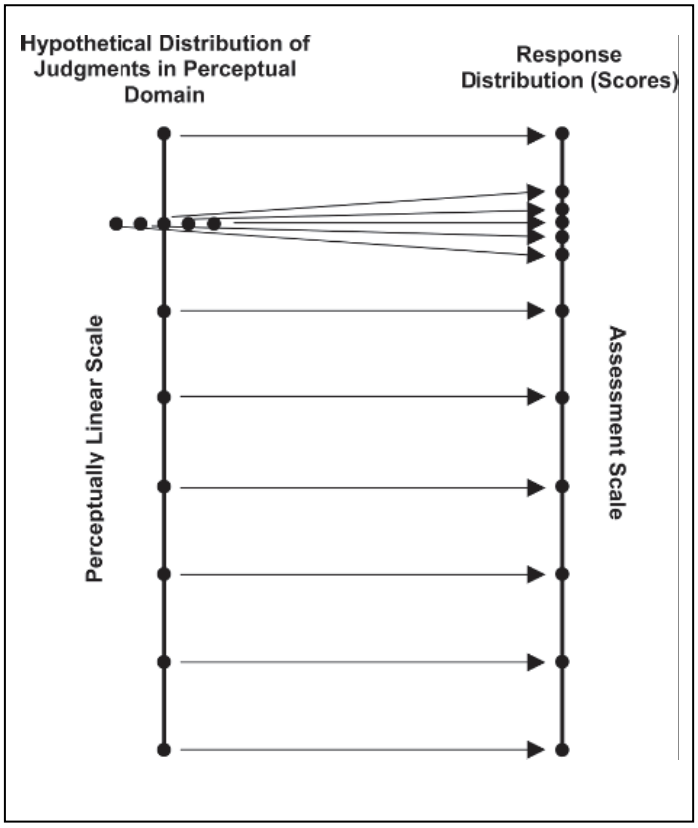


Fig 5.7 The effect of stimulus frequency bias [Zielinski *et al*, 2008].

5.2.1.6 Centring bias

Centring bias (Fig 5.8) reveals itself as shift in scores towards the centre of the scale due to the lack of a reference to describe the assessment scale. Rank order information is preserved but it makes an interpretation of the relative differences between the stimuli unreliable. Hence centring bias is a problem in multi-stimulus tests if the scale is not calibrated properly and can be reduced by using direct or indirect anchoring.

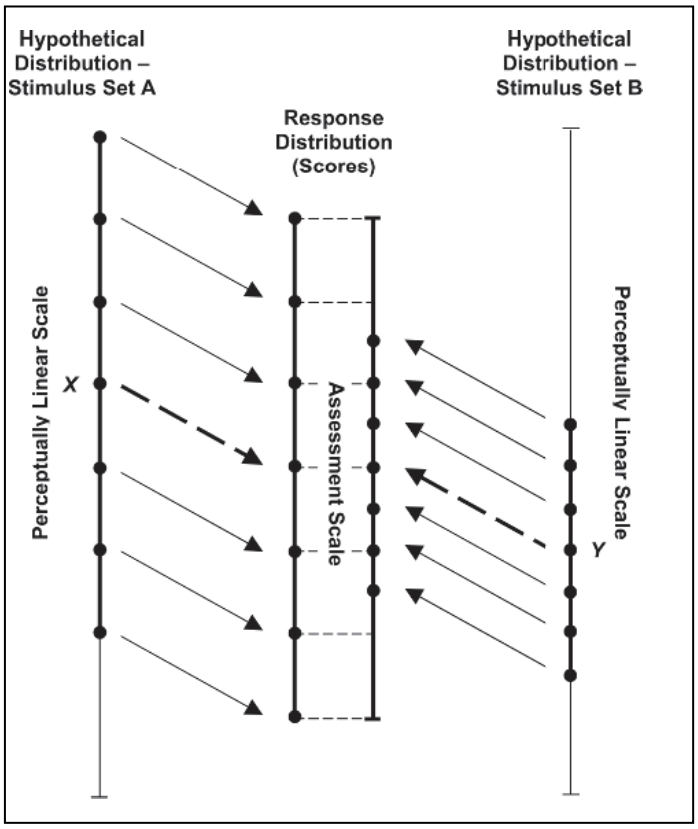


Fig 5.8 The effect of centring bias [Zielinski *et al*, 2008].

5.2.2 Other biases

There are a number of other biases which are not exclusive to MUSHRA or multistimulus tests, but should be considered during the collection of subjective data for this research project. These are summarised in table 5.4.

Other biases known to affect audio quality listening tests	Manifestations	Potential implications	Examples of bias reduction
Recency effect bias (halo bias)	Assessment of stimulus is influenced by scaling of previous or recent stimulus or by the perceived quality of the part of the audio excerpt auditioned most recently	Over or under-estimation of audio quality. differences between stimuli.	Use short looped recordings with consistent characteristics. Randomise the stimuli. Synchronously loop the stimuli.
Equipment bias, Listener expectation bias	Systematic shift in the distribution of the scores, due to listener expectation, overtraining, liking of stimuli, or distracting objects.	Over or under-estimation of audio quality.	Use blind listening tests. Use a large population of listeners from different backgrounds.
Unfamiliarity with magnitude/stimuli	Inconsistency in the scoring of stimuli	Over or under-estimation of audio quality.	Familiarise or train the listeners before the test.

Table 5.4 Other biases affecting subjective tests (adapted from Zielinski *et al* [2008] and Bech and Zacharov [2006]).

5.2.3 Biases: summary and conclusions

Biases are systematic errors which influence the mapping process that listeners use to transfer their opinion of a stimulus to the test scale. Bias is not easy to identify and is difficult to remove once it has "infected" the data. It is desirable for the sake of the validity of the QESTRAL model to minimise or reduce the influence of bias on the subjective data collected for its calibration. Of the standard test methods MUSHRA is the most suitable for use in this project. However a number of biases have been shown to potentially affect the scores collected using it (these are summarised in Table 5.5). Therefore a new listening test method should be developed that incorporates methods of reducing bias in audio quality listening tests discussed above.

Biases known to affect audio quality listening tests	Examples of bias reduction
Stimulus spacing bias	Select stimuli that are perceptually equally spaced. Randomise the presentation of stimuli
Range equalising bias – "Rubber ruler" effect.	Use direct or indirect anchoring
Bias due to perceptually non-linear scale	Use a label-free scale or only label the top and bottom of the scale.
Interface bias	Remove labels, numbers or markings from the interface. Use a large population of listeners.
Stimulus frequency bias	Use a balanced design (avoid presenting perceptually similar or identical stimuli more often than other stimuli).
Centring bias	Use direct or indirect anchoring.
Recency effect bias (halo bias)	Use short looped recordings with consistent characteristics. Randomise the stimuli. Synchronously loop the stimuli.
Equipment bias, Listener expectation bias	Use blind listening tests. Use a large population of listeners from different backgrounds
Unfamiliarity with magnitude/stimuli	Familiarise or train the listeners before the test.

Table 5.5 Summary of biases affecting audio quality tests and examples of methods of reducing them (adapted from Zielinski *et al* [2008]).

5.3 Creation of a listening test method to reduce bias

This section details the various steps taken to reduce the potential for bias in the listening test method.

5.3.1 Alteration of the MUSHRA graphical user interface

In order to reduce the influence of biases related to the appearance and contents of the user interface, the MUSHRA interface has been altered, using the information discussed above, to create a novel user interface (see Fig 5.9).

Biases are said to result from the perceptually non-linear quality labels used in the MUSHRA interface, so to reduce this problem the labels have been removed and replaced by a downward pointing arrow labelled 'Worse'. This is similar in concept to what Watson has termed a 'Polar scale' [Zielinski *et al*, 2008]. The arrow indicates that a stimulus of lower quality than the reference should

be graded below the top position on the scale, the magnitude of its position depending upon the severity of the degradation.

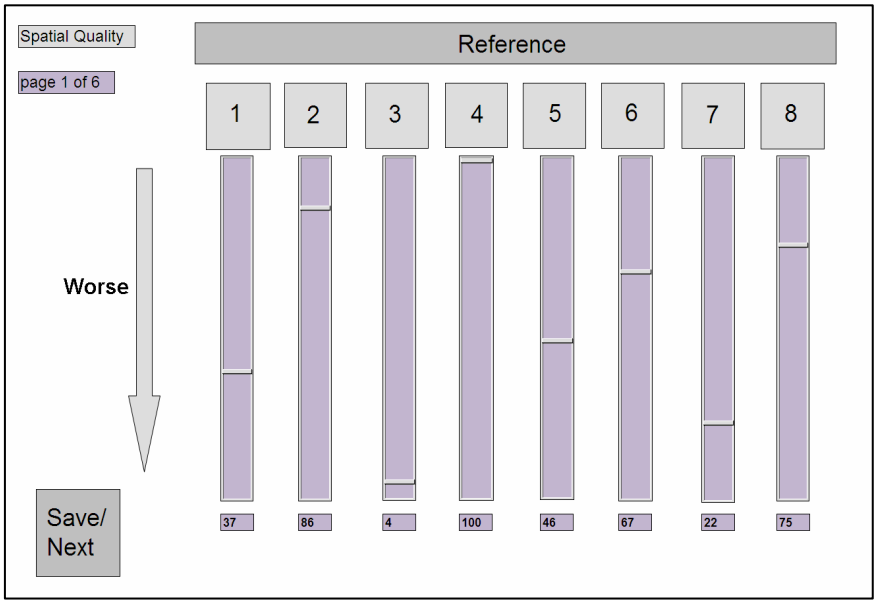


Fig 5.9 Screenshot of the proposed GUI.

Additionally the scale markings which define the numerical value of the labels on the MUSHRA interface, which have been shown to create interface bias [Zielinski *et al*, 2008], have also been removed. However the numerical counter indicating slider position has been kept to give the listeners guidance in their scoring, and consequently this bias may not be completely eliminated.

5.3.2 Indirect anchoring

Three indirect (hidden) anchors will be used to calibrate or define the top, middle and bottom of the scale on every test page. This will provide the listener with a perceptual reference for the range of the scale which will have a stabilising effect on the scale, helping to reduce both centring bias and range equalising bias. The anchors will be included on every page of the test. As a large number of stimuli are required for the calibration of the QESTRAL model, too many to be assessed on a single page and a single test, the inclusion of the anchors on every test page will allow comparisons of the subjective scores between different test pages and different tests to be made and will also encourage listeners to utilise the full range of the scale. However, as anchors will be included on every page of the test, listeners will make their assessments of the stimuli in the context of these anchors, so they should be selected carefully. Similarly to MUSHRA and BS.1116-1, an unprocessed version of the reference recording will be used as the high anchor to calibrate the top of the scale. This high anchor could also be used as a method of determining a listener's ability to discriminate differences between stimuli. The audio processes chosen for use as middle and low anchors will be carefully selected during a series of pilot studies.

5.3.3 Reducing other bias

Stimulus spacing bias will be reduced by carefully selecting SAPs to create stimuli that equally cover the range of the spatial quality scale. This will be achieved during listening sessions conducted by the author during the creation of the listening tests and informed by the results of pilot studies. Randomising the presentation order of the stimuli presented to the listeners will also help to reduce stimulus spacing bias, as well as stimulus frequency bias and recency effect bias. Synchronously looping the playback of each stimulus to ensure that the entire recording is evaluated and that switiching between them is seamless, will also help to reduce recency effect bias. Expectation bias will be removed by obscuring all test equipment from the listeners. Any bias due to listener unfamiliarity with the task of evaluating of spatial quality and the GUI will be reduced by instructing the listeners on the task and also by allowing them to have a practice run before each test.

5.3.4 Reduced-bias listening test method: summary and conclusions

The new listening test method is based on MUSHRA but incorporates methods for reducing each of the biases discussed in section 5.2. Table 5.6 summarises the methods used to reduce each bias.

Biases known to affect audio quality listening tests	Examples of bias reduction	Method of reduction
Stimulus spacing bias	Select stimuli that are perceptually equally spaced. Randomise the presentation of stimuli	Stimuli will be carefully selected and their presentation order will be randomised.
Range equalising bias – “Rubber ruler” effect.	Use direct or indirect anchoring	Indirect anchoring
Bias due to perceptually non-linear scale	Use a label-free scale or only label the top and bottom of the scale.	GUI labels removed
Interface bias	Remove labels, numbers or markings from the interface. Use a large population of listeners.	GUI labels and markings are removed
Stimulus frequency bias	Use a balanced design (avoid presenting perceptually similar or identical stimuli more often than other stimuli).	The presentation order of stimuli will be randomised
Centring bias	Use direct or indirect anchoring.	Indirect anchoring
Recency effect bias (halo bias)	Use short looped recordings with consistent characteristics. Randomise the stimuli. Synchronously loop the stimuli.	Stimuli will be synchronously looped and their presentation order will be randomised.
Equipment bias, Listener expectation bias	Use blind listening tests. Use a large population of listeners from different backgrounds	An acoustically transparent curtain will be used to disguise the test equipment
Unfamiliarity with magnitude/stimuli	Familiarise or train the listeners before the test.	Listeners will be given test instructions and a familiarisation session.

Table 5.6 Summary of biases affecting audio quality tests (adapted from Zielinski *et al* [2008]) and methods of reducing them employed in the new listening test method.

5.4 Summary and conclusions

Formal subjective testing is currently regarded as the most reliable method for the evaluation of audio quality. This research requires a suitable method for reliably investigating spatial quality, and so

existing standards for the subjective assessment of audio quality were studied. The International Telecommunication Union (ITU) has developed and standardised methods for the evaluation of audio quality that are used extensively in research. These are BS.1116-1, BS.1534. BS.1116-1 was designed for the detection of small impairments between stimuli and is therefore limited to the evaluation of a single test condition per test page. For the correct calibration of the QESTRAL model for the prediction of spatial quality subjective data need to be collected on a large number of spatial audio processes (SAPs) representing the wide range of impairments to spatial quality. Using the BS.1116-1 method for this task would have been inefficient and very time consuming and therefore it was decided that it was not suitable for use in this research. By comparison BS.1534 (MUSHRA), a multistimulus test, allows several stimuli to be compared simultaneously. This was seen as a much more efficient way of collecting the amount of subjective data required for this project. However it has been observed that results collected from experiments employing the MUSHRA method suffer from biasing. Biases are systematic errors which influence the mapping process that listeners use to transfer their perception of a stimulus to the test scale. Bias is not easy to identify and is difficult to remove once it has "infected" the data. Using biased data to calibrate a model would limit its validity and generalisability and so it was desirable to remove or reduce the appearance of bias in the data collected for this project. Therefore a new listening test method was developed that incorporates methods of reducing bias in audio quality listening tests discussed. Table 5.6 summarises the methods used to reduce each bias.

Chapter 6 – Pilot studies

At the outset of the QESTRAL project it was not known if it was possible to subjectively assess spatial quality robustly. Chapter 3 proposed a suitable approach to developing the QESTRAL model; this required the collection of subjective data using listening tests, and a novel listening test method was developed in chapter 5. Prior to conducting a large scale listening test this chapter describes and discusses four listening tests conducted as pilot studies with the aim of:

(i) determining the suitability of the proposed listening test method and GUI for evaluating spatial quality,

(ii) assessing the difficulty of the task required of the listening test subjects, at two listening positions, using a wide range of different SAPs,

(iii) trialling a method for the selection of suitable SAPs prior to the large scale listening test,

(iv) addressing the question raised in section 2.1.1 about whether changes to timbral quality might affect the assessment of spatial quality,

(v) identifying and investigating variables in the experiments that influence perceived spatial quality, and determine their relevance for calibrating of the QESTRAL model,

(vi) selecting suitable SAPs for use as indirect anchors.

6.1 Pilot study 1 – An initial investigation of the spatial quality listening test method

This section describes and discusses the aims, methodology and results of pilot study 1, which was conducted to address aims (i), (ii), (v) and (vi) from the list above.

6.1.1 Aims of pilot study 1

The aims of pilot study 1 are as follows:

i) Test the suitability of the listening test method designed for the assessment of spatial quality. Suitability will be determined by analysing the listeners' discrimination ability and consistency in repeated assessments and by comparison with other similar listening tests.

ii) As described in section 1.1, a unique function of the QESTRAL model, allows it to evaluate the reproduced soundfield at a number of different listening positions across the listening area. As this could be useful for audio system designers and researchers it may be important for the QESTRAL model to be calibrated for the objective evaluation of spatial quality at multiple listening positions. Therefore the second aim is to determine whether the perception of spatial quality at a central listening position differs significantly from

that at on off-centre position (1 metre to the right). If it does then calibration at multiple positions will be required. The suitability of the listening test method will be examined for both listening positions.

iii) The third aim is to identify which variables in the experiment have an influence on the perceived spatial quality. This will be achieved by statistical analysis of the results.

iv) The fourth aim is to evaluate the suitability of the SAPs chosen to be used as indirect anchors. This will be achieved by analysis of the subjective scores.

6.1.2 Creation of stimuli for pilot study 1

This section describes the creation of the stimuli used in pilot study 1.

6.1.2.1 Programme material evaluated in pilot study 1

Four 5-channel programme items were chosen for assessment. Descriptions of the programme items are provided in table 6.1.

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, C and R. Audience applause in 360°.
2	Classical	F-B	Excerpt from Felix Mendelssohn – A Midsummer Night's Dream - Symphony No. 4 "Italian" (BBC catalogue). Wide continuous front stage, Ambient surrounds with reverb from front stage.
3	Pop/Rock	F-F	Excerpt from Steely Dan – Jack of Speed. Wide continuous front stage (including Drums, Bass, Guitars). Brass in Surrounds.
4	Pop	F-F	Excerpt from The Eagles – Seven Bridges Road. 5 harmony voices only, one in each channel. Audience in gaps.

Table 6.1 Description of programme items evaluated in pilot study 1.

The different programme items were chosen with the intent to span a representative range of ecologically valid programme material, likely to be listened to by typical audiences of consumer multichannel audio reproduction, while also covering typical genres and spatial audio mixing styles or 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 multichannel audio. There are two commentators panned slightly left and right of the front centre position where the television set would likely be placed. Audience applause and ambience can be heard in 360° around the listening position. This recording represents a typical F-F scene type as all audio sources are either close or clearly perceivable. In comparison programme item 2 (Classical) is a classical recording which exhibits a different mix style, typical of many recordings from this genre, whereby the front three loudspeakers (i.e. left, centre and right) contain a wide continuous mix of the orchestra while the rear or surround loudspeakers contain ambient or reverberant energy. This recording represents a typical F-B scene type.

6.1.2.2 Spatial audio processes (SAPs) investigated in pilot study 1

Eight different SAPs (Table 6.2) were selected to be applied to each programme item to create 32 stimuli.

No.	Spatial audio process	Description
1	Altered loudspeaker locations A	Ls and Rs re-positioned at -90° and 90°
2	Altered loudspeaker locations B	L and R re-positioned at -10° and 10°
3	Channel removed A	Ls removed
4	Inter-channel crosstalk A	1.0 downmix in all channels
5	1.0 downmix	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.
6	Anchor recording A	High Anchor - Unprocessed reference.
7	Anchor recording B	Mid Anchor - 2.0 downmix: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.
8	Anchor recording C	Low Anchor – 1.0 downmix reproduced asymmetrically by the rear left loudspeaker only.

Table 6.2 List of spatial audio processes investigated in pilot study 1.

The audio processes were selected to enable the investigation of a wide range of different spatial qualities with the intention that they would lead to listener responses covering the full range of the spatial quality scale. All processes were chosen in the light of an informal listening session conducted by the author and discussions amongst the QESTRAL project team. Anchor recording A was chosen to define the very top of the scale and was identical to the reference stimulus. Anchor recording B, a 2.0 downmix was chosen to define the middle portion of the scale, while anchor recording C, a 1.0 downmix reproduced asymmetrically through the left surround (Ls) loudspeaker only, was chosen to define the lower portion of the scale.

6.1.2.3 Stimulus loudness equalisation

Effective models for loudness equalising time-varying mono and multi-channel audio signals exist [Glasberg and Moore, 2002][Seefeldt et al, 2004][Seefeldt et al, 2006]. The accuracy of these models is most often compared against judgements made by a listening panel. Therefore considering the complexity and varied range of the SAPs it was decided that the most appropriate method of loudness equalising the stimuli (SAP and programme item combinations) would be to use a listening panel. Using a specially designed GUI with a gain slider that adjusted each channel equally and simultaneously, the listeners were asked to make each stimulus equally loud to the reference recording (unprocessed programme item). The listener's gain adjustments were averaged and applied to the stimuli. This corresponded to a playback level of approximately 75-80dB $L_{AEQ(1\text{-}3mins)}$.

6.1.3 Apparatus employed for pilot study 1

Pilot study 1 was conducted at the Institute of Sound Recording in a listening room which meets ITU-R BS.1116-1 [1997] requirements. A 5-channel loudspeaker system was used as a reference system (see Fig 6.1). The loudspeakers were arranged in 3/2 stereo configuration according to the requirements described in ITU-R BS.775 [1992-1994]. A number of additional loudspeakers were also

employed, when required, for SAPs 1 and 2 (see Table 6.2). Bang and Olufsen Beolab 3 loudspeakers (Frequency response: 50 – 20,000 Hz [Bang & Olufsen, 2011]) were used in all cases. Listeners selected stimuli and recorded their responses using a laptop situated at the listening position. Prior to each test all channel gains were calibrated individually to have the same sound pressure level, at listening position 1, using a pink noise signal. Not shown in the diagram is an acoustically transparent but visually opaque curtain, used to disguise the loudspeaker positions and type from the listener.

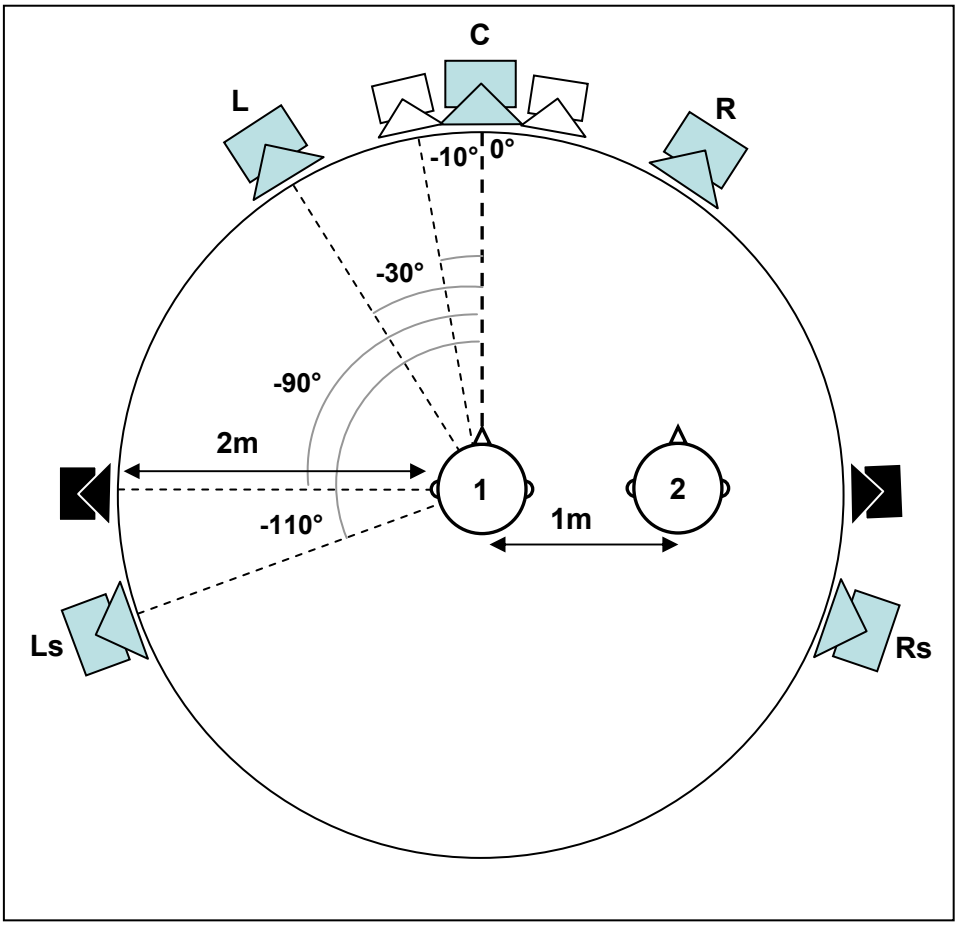


Fig 6.1 Schematic illustrating the listening positions and loudspeaker positions employed for pilot study 1. Loudspeakers labelled L, C, R, Ls and Rs indicate the 3/2 loudspeaker array used as the reference system. Other loudspeaker positions indicate those employed for processes 1 and 2 (see Table 6.2).

6.1.4 Methodology employed for pilot study 1

The listeners sat two tests, the first conducted at listening position 1 (LP1) in the centre of the loudspeaker system and the second conducted at listening position 2 (LP2), 1 metre to the right of centre (as labelled in figure 6.1). Listeners sat the test at listening position 1 first. The listeners were instructed to assess the spatial quality of each stimulus compared against the reference using the graphical user interface (GUI) described in chapter 5 (the full listener instructions are given in Appendix A). The presentation order of the stimuli in each case was randomised. A full test consisted of two assessments of all stimuli and lasted approximately 30-40 minutes. Before commencing each

test listeners completed a familiarisation session using the same GUI. This enabled them to hear, and practise the assessment of, each stimulus featured in the test using the interface. Seven Tonmeisters or experienced listeners from the Institute of Sound Recording (IoSR) at the University of Surrey took part in the test.

6.1.5 Listener selection

It is accepted that using only Tonmeister students from the IoSR may limit the generalisability of the model but it was felt that the task of assessing spatial quality would have been too difficult for an inexperienced listener and that they would not be capable of providing consistent results. Hence Tonmeisters or post-graduate students from the IoSR were used as listeners because of their experience with critical listening (NB. ITU-R BS.1116-1 [1997] and BS.1534 [2001] both suggest using expert listeners).

As mentioned above, each listener received a small amount of training on the task before each test. This took the form of detailed instructions (Appendix A) and a familiarisation session, whereby the listeners could familiarise themselves with process of assessing spatial quality using the GUI and also with the stimuli featured in that test.

6.1.6 Discussion of the results of pilot study 1

This section presents and discusses the results of pilot study 1.

6.1.6.1 Assessment of listener performance in pilot study 1

Each listener's responses were assessed, so that the most reliable data could be selected for analysis and investigation. Two methods of assessment were used:

1) Discrimination ability determined by conducting a one-sample t-test on each listener's scores for 'Anchor recording A' (high anchor – unprocessed reference). A one-sample t-test tests whether a mean is statistically significantly different ($p < 0.05$) from a specified value. If a listener was capable of identifying this stimulus and scoring it as instructed, they were deemed as having suitable discrimination ability.

2) Consistency was determined by investigating the magnitude of a listener's error in repeat judgements. Root mean square error was calculated between repeated assessments of stimuli. To pass this test a listener's RMS error must not be greater than 15% (based on a 100 point test scale). Although smaller values of RMS error such as 10% have been considered as acceptable in similar experiments [Rumsey, 1998] a higher threshold was chosen due to the expected difficulty of the task. (NB. The anchor recordings are assessed many more times than the other stimuli so to balance the assessment they are removed). Figures 6.2 and 6.3 illustrate the results of both assessments. The listeners who were removed from the results are circled. Listener 2 being removed from the database

for LP2 (NB. There was no data for listener 1 at LP2 because they were unavailable on the day of the test).

Listener	Mean score	p
1	100	N/A
2	100	N/A
3	100	N/A
4	100	N/A
5	97.63	0.125
6	99.88	0.351
7	95.63	0.172

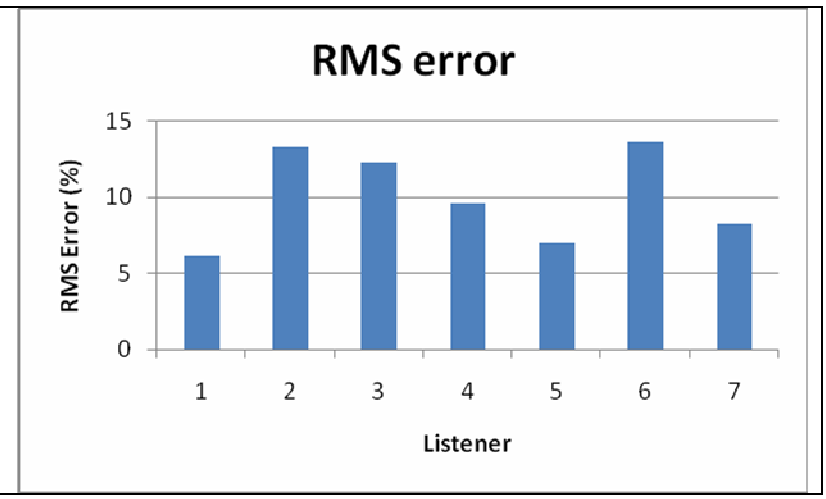


Fig 6.2 Pilot study 1, listening position 1 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
2	100	N/A
3	100	N/A
4	100	N/A
5	98.88	0.197
6	100	N/A
7	98.63	0.120

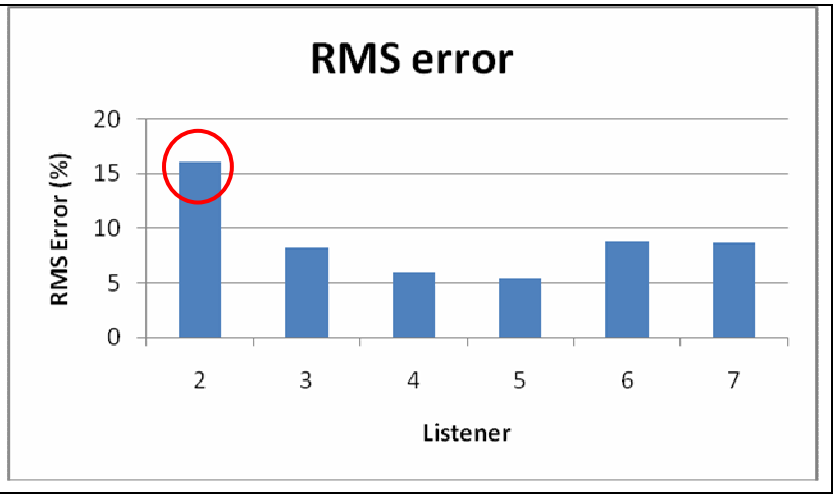


Fig 6.3 Pilot study 1, listening position 2 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

6.1.6.2 Analysis of Variance (ANOVA) of the results of pilot study 1

A univariate ANOVA was conducted to investigate the main effects of the experimental variables on spatial quality (dependent variable) and their 1st order interactions (Table 6.3). SAP (Process), listening position (LP), programme item (ProgItem) and listener (Listener) were included in the model as independent variables. The structure of the ANOVA model is shown in equation B1 (Appendix B).

The variable Process (SAP) had a statistically significant effect ($p < 0.05$) on perceived spatial quality. The main effects and 1st order interactions reveal that listening position (LP), programme item (ProgItem) and listener all had a significant effect on perceived spatial quality. To illustrate the most important experimental factors or interactions, figure 6.4 depicts main effects and interactions with an effect size (partial eta squared) greater than 0.1. These are discussed in the proceeding sections. The

effect size describes the total amount of variance in the dependent variable attributable to each independent variable.

Tests of Between-Subjects Effects

Dependent Variable: SQ

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1061697.377[a]	112	9479.441	87.544	.000	.937
Intercept	2641236.039	1	2641236.039	24392.22	.000	.974
Process	794013.940	7	113430.563	1047.549	.000	.918
LP	406.406	1	406.406	3.753	.053	.006
ProgItem	994.572	3	331.524	3.062	.028	.014
Listener	16041.576	6	2673.596	24.691	.000	.184
Process * LP	20077.519	7	2868.217	26.488	.000	.221
Process * ProgItem	15305.486	21	728.833	6.731	.000	.177
Process * Listener	20270.796	42	482.638	4.457	.000	.222
LP * ProgItem	1024.931	3	341.644	3.155	.024	.014
LP * Listener	253.438	4	63.359	.585	.674	.004
ProgItem * Listener	4252.142	18	236.230	2.182	.003	.057
Error	70924.642	655	108.282			
Total	3948415.000	768				
Corrected Total	1132622.020	767				

a. R Squared = .937 (Adjusted R Squared = .927)

Table 6.3 Univariate ANOVA results output for pilot study 1.

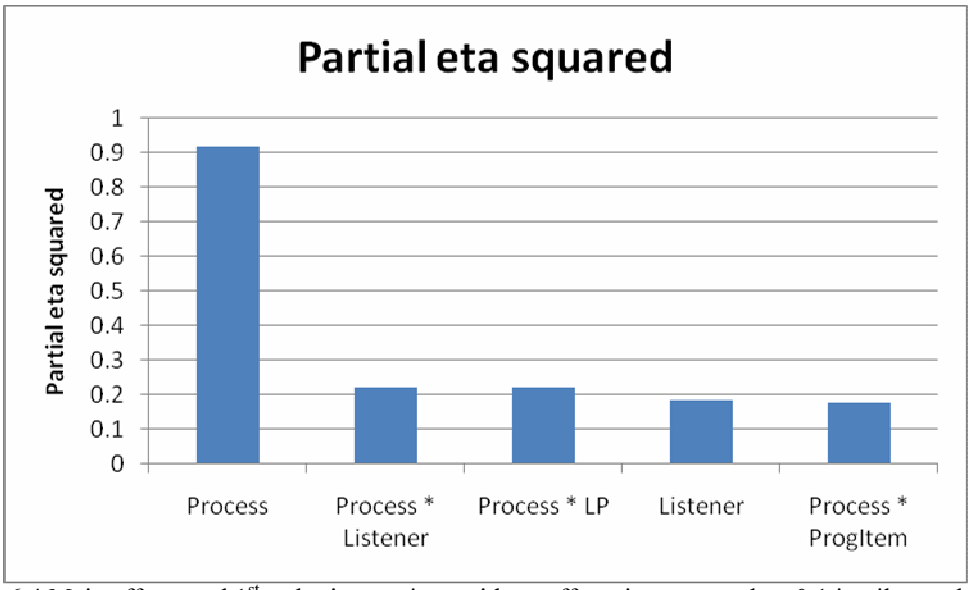


Fig 6.4 Main effects and 1st order interactions with an effect size greater than 0.1 in pilot study 1.

6.1.6.3 The influence of spatial audio process on spatial quality in pilot study 1

SAP has the largest effect on spatial quality. Figure 6.5 shows means and 95% confidence intervals for all processes (including anchor recordings), averaged across both listening positions and all programme items and listeners. Although this method of presentation is oversimplified and hides the influence of listening position, programme item type and listener, it does allow the mean scores for individual audio processes to be observed and compared. The mean scores and confidence intervals for

the SAPs span the entire range of the test scale and have 95% confidence intervals narrower than 10 points (10%) of the scale.

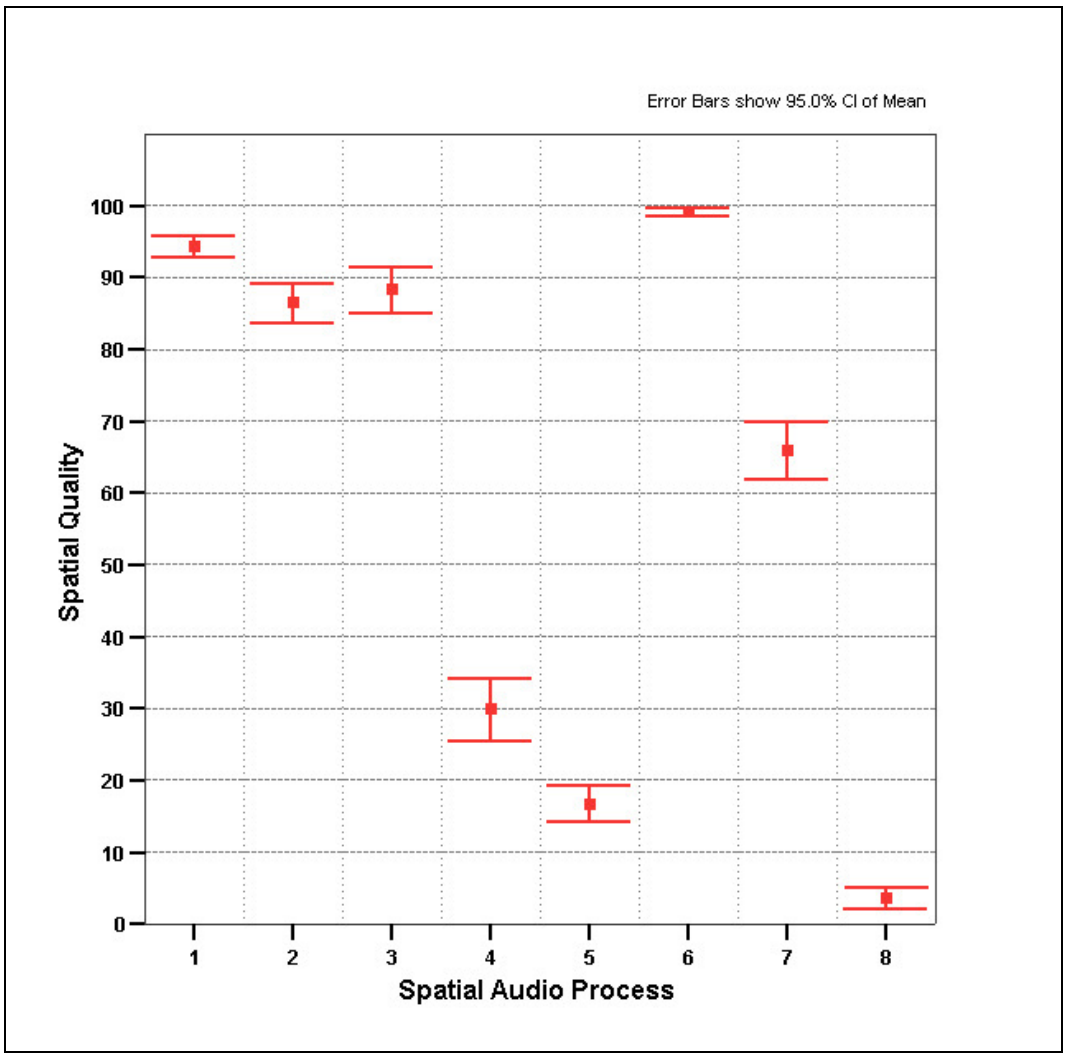


Fig 6.5 Pilot study 1 means and 95% confidence intervals for all audio processes averaged across programme item type, listening position and listener.

SAP 6 (Anchor recording A – high anchor (unprocessed reference)) is scored at the top of the scale, and SAP 8 (Anchor recording C – low anchor) is scored at the very bottom of the scale. The scoring of the low anchor at the very bottom of the scale, as hoped, is remarkable, but can be explained by comparing it with the score of SAP 5 (1.0 downmix from 5-channel to the centre channel (C)) which was scored as the next lowest process. Hence if SAP 5 is repositioned to the left surround channel (Ls) it is naturally perceived as creating a larger impairment to spatial quality. The score positions for the high and low anchors support using them as high and low anchors in future tests. However SAP 7 (Anchor recording B – mid anchor) is scored higher on the scale than expected (66%). SAP 1 is scored at 95%, which shows that symmetrically altering the locations of the left and right surround loudspeakers from ±110º to ±90º only slightly impairs the perceived spatial quality. This is also true for SAP 2 (L and R repositioned at ±10º), scored at 85% and SAP 3 (Ls removed), scored at 89%.

SAP 4 (1.0 downmix in all channels) which was scored at 30% is an example of a process which creates a substantial impairment to spatial quality. Informal listening revealed that this process substantially reduces the perceived spaciousness and ability to localise audio sources in the programme items.

Importantly none of the SAPs were scored in the middle of the scale. This highlights a potential problem for collecting data that spans the entire range of spatial quality.

6.1.6.4 The influence of listener on spatial quality in pilot study 1

The interaction of listener with SAP has the second largest effect on perceived spatial quality and suggests that there is a difference in opinion or lack of consensus between listeners for certain stimuli. The subjective scores for these stimuli will exhibit a multi-modal or platykurtic distribution of data. This is of particular importance for calibrating the QESTRAL model, as subjective score averages will be used to describe the spatial quality of each SAP predicted by the model, and stimuli which elicit a large difference in opinion between the listeners will not have a meaningful or reliable score average. Such stimuli should therefore be considered for removal from the calibration database. A method of investigating this is to analyse the distribution of the subjective scores for each stimulus using statistical and visual analysis techniques, such as assessments of normality and modality. A summary of the results of this analysis is displayed in table 6.4 (A full analysis is presented in Appendix C).

Listening position	Programme item	Spatial audio process		
		Mean	Median	Remove
1	1	2, 5, 7, 8	1, 3, 4, 6	-
	2	4	1, 3, 5, 6, 7, 8	2
	3	-	1, 2, 5, 6, 7, 8	3, 4
	4	2, 4, 5	1, 6, 8	3, 7
2	1	1, 2, 3, 4, 7	5, 6, 8	-
	2	1, 2, 4, 5, 7	3, 6, 8	-
	3	2, 3, 7	1, 5, 6, 8	4
	4	1, 2, 3, 5, 7	6, 8	4

Table 6.4 Stimuli in pilot study 1 that should be removed from a database used to calibrate the QESTRAL model.

6.1.6.5 The influence of listening position on spatial quality in pilot study 1

The interaction of listening position with SAP is shown to have the third largest effect on perceived spatial quality. This suggests that certain stimuli create an impairment to spatial quality that is different at the second listening position. A one-way ANOVA using listening position as the factor was used to statistically assess which stimuli exhibited this effect. The list of processes where this test was found to be statistically significant ($p < 0.05$), are given in table 6.5.

Figures D1-4 (Appendix D) illustrate this list as means and 95% confidence intervals. Interestingly SAP 7 (mid anchor - 2.0 downmix) was scored statistically significantly different between listening positions for every programme item type. It was scored as much as 30% lower at listening position 2 (LP2) than at listening position 1 (LP1). It has been shown in previous research

that 2-channel stereo recordings are often perceived as being quite enveloping [George, 2009][Conetta, 2007]. Although many of the source locations in the audio scene will be altered, it is suggested that from LP1 (centralised) a 2.0 downmix has a similar perceived envelopment to the 5-channel reference recording, hence it is scored higher on the scale here. However from LP2 (off-centre), this illusion is broken because the listener is seated closer to one channel (in this case the right front loudspeaker). Another interesting example is SAP 3 (Ls removed) where listeners' scores are significantly different between LP1 and LP2 for both programme items 3 and 4 (both F-F recordings). SAP 3 is scored significantly higher (approx. 20%) at LP2 than LP1. This could be because the removal of Ls, noticeable at LP1, was masked when the listener was seated further away at LP2.

Programme item	Spatial audio process
1	7
2	2, 4, 7
3	3, 7
4	3, 4, 7

Table 6.5 Stimuli which create a statistically significant difference in perceived spatial quality between listening positions in pilot study 1.

6.1.6.6 The influence of programme item type on spatial quality in pilot study 1

The interaction of programme item type with SAP is also shown to have a significant effect on perceived spatial quality. This suggests that certain audio processes create an impairment to spatial quality that is different between programme items. A one-way ANOVA using programme item as the factor was used to statistically assess which stimuli exhibited this effect. The list of SAPs where this test was found to be statistically significant ($p < 0.05$), are given in table 6.6.

Listening position	Spatial audio process
1	2, 3, 7
2	2

Table 6.6 Stimuli which create a difference in perceived spatial quality between programme item types in pilot study 1.

Figures E1 and 2 (Appendix E) illustrate this list as means and 95% confidence intervals. Listeners scored SAP 2 (L and R repositioned at ±10º) statistically significantly differently between programme items at both listening positions. At LP1 listeners scored this process 10-20% lower when combined with programme item 1 and 2 than items 3 and 4 (Fig. E1). An explanation for this could be that in programme item 1 (TV/Sport), scored at 84%, the commentators are moved noticeably closer together. This is not necessarily annoying but it is an obvious change to the spatial scene that is not preferred when compared to the reference. In programme item 2 (Classical), which was scored the lowest, at 73%, the front scene width is reduced. This is very noticeable because the front scene is the dominant audio scene in this recording. Also interesting is that the listener score confidence intervals for these programme items are much wider than those for items 3 and 4, signifying that there was a greater

spread in the opinion of the spatial quality, meaning that some listeners were less annoyed by the degradation created by this SAP than others.

At LP2, programme item 1 is scored the lowest, at 83% (similar to its score at LP1) and again significantly differently from programme item 4 (scored at 96%), however programme item 2 is scored similarly to items 3 and 4 (Fig. E2). This could be explained by the angle of listening at LP2 making the degradation created with programme item 2 less noticeable but, in the case of programme item 1, the changed location of the commentators is still perceived.

Listeners scored SAP 3 (Ls removed) statistically significantly higher (approx. 20-30%) when it was applied to programme items 1 and 2 than items 3 and 4. This could be because the content in rear channels of items 1 and 2 is very diffuse applause, or room reverberance. The removal of a channel containing these sorts of audio sources seems to create a minor impairment to the perceived spatial quality. There might for example be a small change in the feeling spaciousness or envelopment. This was not the case when the process was applied to items 3 and 4 which contain predominantly foreground sources in their rear channels and whose removal is much more perceivable and detrimental as these sources are very localisable.

6.1.7 Pilot study 1: conclusions

Analysing each listener's performance revealed that consistency levels similar to other listening tests were achieved, and analysing their discrimination ability indicated that they were capable of identifying the hidden reference correctly. This indicates that listeners can reliably assess the spatial quality of the stimuli investigated using the listening test method and graphical user interface developed in section 5.3 and that this method and interface is therefore suitable.

A univariate ANOVA showed that the interaction of SAP with listening position had a statistically significant effect on the perception of spatial quality. This suggests that certain SAPs create an impairment to spatial quality at LP2 that is different from that at LP 1. The ANOVA also revealed that listener and programme item type influenced the perception of spatial quality. The interaction of listener with SAP had the second largest effect (after SAP) on perceived spatial quality and suggests that there was a difference in opinion between listeners for certain stimuli. The stimuli listed in table 6.4 elicited a statistically significant difference in opinion or lack of consensus between the listeners and are deemed to have unreliable score averages. Therefore as the data used to calibrate the QESTRAL model will consist of SAP score means, the stimuli where this effect is observed should be considered for removal from the database. The interaction of programme item with SAP was also shown to have a statistically significant effect on perceived spatial quality, suggesting that certain SAPs create an impairment to spatial quality that differs between programme items types. Table 6.6 lists the SAPs where this occurred. In consideration of the database used to calibrate the QESTRAL model aggregated scores for the SAPs whose impairment to spatial quality differs between listening positions and/or programme items will have unreliable means. Therefore from this evidence listening

position and programme item should be included as separate variables in the QESTRAL model, which could be achieved either by creating different calibrations for each or by using a subjective database that incorporates scores collected from both listening positions separately.

Anchor recording A and anchor recording C were scored in their intended locations. This result supports using these processes as high and low anchors in future tests. However anchor recording B was scored higher on the scale than expected (at 66%). Interestingly none of the SAPs evaluated in pilot study 1 were scored in the middle of the scale; only the top 20% and lowest 30% of the scale were used by the listeners. This indicates that the evaluated SAPs were limited to only small or large degradations to the lower level spatial attributes that contribute to the perception of spatial quality. As discussed in section 3.1 this is a known risk of using a direct method for the QESTRAL model development and highlights the need for a method of selecting suitable SAPs.

6.2 Pilot study 2 – Further investigation of spatial quality

This section describes and discusses the aim, methodology and results of pilot study 2, which was conducted to address the problem with limited middle-of-scale usage in the previous pilot study and to further address aims (i), (ii), (v) and (vi) set out at the beginning of this chapter.

6.2.1 Aims of pilot study 2

The aims of pilot study 2 are as follows:

i) As none of the SAPs evaluated in pilot study 1 was scored in the middle of the scale. The first aim is to identify audio processes which create a medium level of impairment to perceived spatial quality and would be scored in the middle of the scale. This will also help to identify a more suitable indirect anchor for the middle of the scale,

ii) Investigate a wider range of additional SAP types not evaluated in pilot study 1, including low bit-rate multichannel audio codecs and virtual surround algorithms,

iii) The third aim is to continue to test the suitability of the listening test method,

iv) The fourth aim is to investigate which variables in pilot study 2 have an influence on the perceived spatial quality. This will be achieved by statistical analysis of the results.

6.2.2 Creation of stimuli for pilot study 2

This section describes the creation of the stimuli used in pilot study 2.

6.2.2.1 Programme material evaluated in pilot study 2

The same recordings as pilot study 1 were used. However, to reduce the overall test length programme item 4 was removed because of similarity to programme item 3. Descriptions of the programme items are provided in table 6.7.

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, C and R. Audience applause in 360°.
2	Classical	F-B	Excerpt from Felix Mendelssohn – A Midsummer Night's Dream - Symphony No. 4 "Italian" (BBC catalogue). Wide continuous front stage, Ambient surrounds with reverb from front stage.
3	Rock/Pop music	F-F	Excerpt from Steely Dan – Jack of Speed. Wide continuous front stage (including Drums, Bass, Guitars). Brass in Surrounds.

Table 6.7 Description of programme items evaluated in pilot study 2.

6.2.2.2 Spatial audio processes (SAPs) investigated in pilot study 2

Thirteen different SAPs were selected to be applied to each programme item to create a number of stimuli exhibiting a range of impairments to spatial quality. These are described in table 6.8.

No.	Spatial audio process	Description
1	Altered loudspeaker locations C	Ls and Rs re-positioned at -170° and 160°
2	Altered loudspeaker locations D	C is skewed; re-positioned at 20°
3	Channel removal B	R is removed
4	Channel rearrangements A	Channel order is randomised
5	Channel rearrangements B	Channel order rotated 1 channel to the left
6	Virtual surround algorithms A	2-channel virtual surround – Trusurround
7	Multichannel audio coding A	Audio codec (80kbs)
8	Multichannel audio coding B	3-stage cascaded audio codec (64kbs)
9	Combination A	3.0 downmix + Channel removal B
10	Combination B	Multichannel audio coding A + Altered loudspeaker locations C
11	Anchor recording A	High Anchor - Unprocessed reference.
12	Anchor recording B	Mid Anchor - 2.0 downmix: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.
13	Anchor recording C	Low Anchor – 1.0 downmix reproduced asymmetrically by the rear left loudspeaker only.

Table 6.8 List of spatial audio processes investigated in pilot study 2.

The audio processes were selected to extend the selection investigated in pilot study 1 to cover all of the key types SAP likely to be introduced by real audio equipment. They were chosen to cover a wide range of different spatial qualities with the intention that they would cover the range of the test scale more evenly than in the earlier experiment, with particular priority to cover the middle of the scale. All processes were chosen in the light of an informal listening session conducted by the author and discussions amongst the QESTRAL project team. The anchor recordings remained the same as those used in pilot study 1. All stimuli were loudness equalised using the method described in section 6.1.2.3. This corresponded to a playback level of approximately 75-80dB $L_{AEQ(1\text{-}3mins)}$.

6.2.3 Apparatus employed for pilot study 2

The apparatus for pilot study 2 (Fig 6.6) was similar to that used in pilot study 1, with a standard 3/2 stereo configuration plus additional loudspeakers for SAPs 1 and 2 (Table 6.8). Bang and Olufsen Beolab 3 loudspeakers (Frequency response: 50 – 20,000 Hz [Bang & Olufsen, 2011]) were used in all cases. Listeners selected stimuli and recorded their responses using a laptop situated at the listening position. Prior to each test all channel gains were calibrated individually to have the same sound pressure level, at the listening position, using a pink noise signal. Not shown in the diagram is an acoustically transparent but visually opaque curtain, used to disguise the loudspeaker positions and type from the listener.

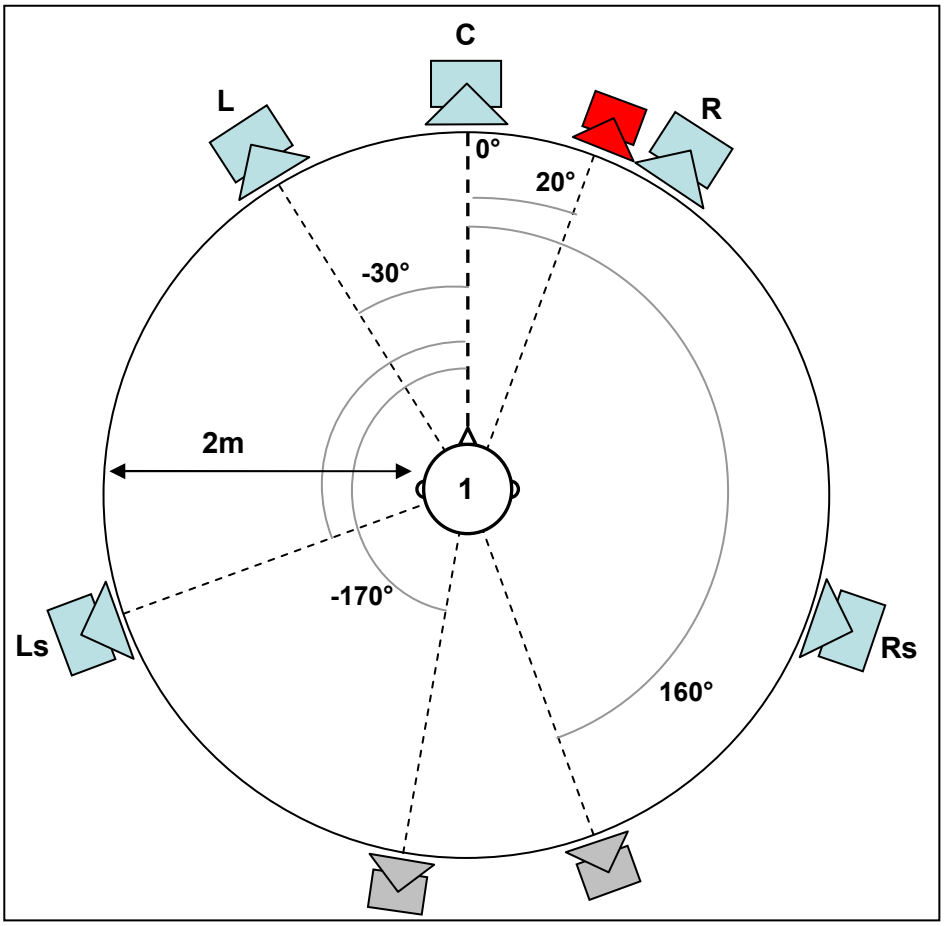


Fig 6.6 Schematic illustrating the listening position and loudspeaker positions employed for pilot study 2. Loudspeakers labelled L, C, R, Ls and Rs indicate the 3/2 loudspeaker array used as the reference system. Other loudspeaker positions indicate those employed for processes 1 and 2 (see Table 6.8).

6.2.4 Methodology employed for pilot study 2

The GUI and method were as for pilot study 1 except that listeners sat one test at listening position 1 only (Fig 6.6). Again, the listeners were instructed to assess the spatial quality of each stimulus, presentation order was randomised, a full test consisted of two assessments of all stimuli and lasted

approximately 30-40 minutes, and an initial familiarisation session was employed. Ten Tonmeisters or experienced listeners from the IoSR at the University of Surrey took part in the test.

6.2.5 Discussion of the results of pilot study 2

This section presents and discusses the results of pilot study 2.

6.2.5.1 Assessment of listener performance in pilot study 2

As in pilot study 1 each listener's responses were assessed, so that the most reliable data could be selected for analysis and investigation (Fig 6.7).

Listener	Mean score	p
1	99.42	0.339
2	99.75	0.082
3	99.50	0.339
4	99.75	0.339
5	100	N/A
6	100	N/A
7	99.58	0.339
8	99.17	0.339
9	94.33	0.339
10	100	N/A

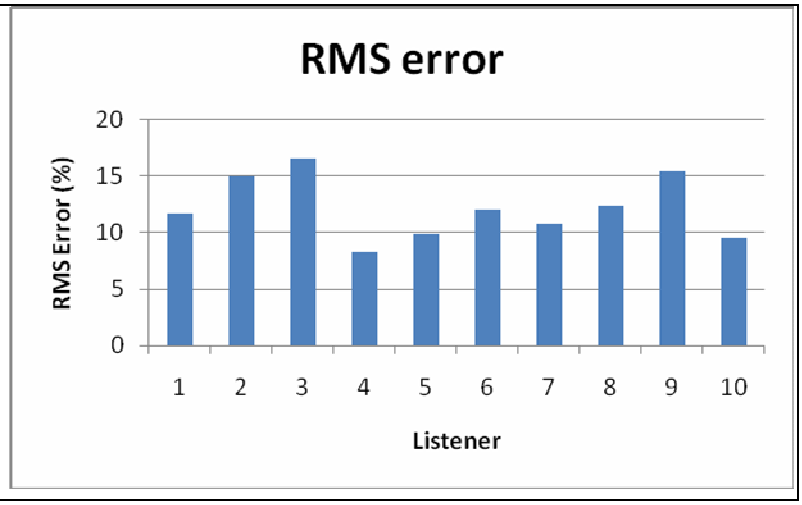


Fig.6.7 Pilot study 2, listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

The outcome of the listener assessment resulted in listener 3 being removed from the results. The RMS error score for listener 9 was 15%. Closer investigation revealed that they had made a single error when identifying Anchor recording A; furthermore the results of the discrimination t-test were not statistically significant ($p < 0.05$); therefore this listener was not removed.

6.2.5.2 Analysis of Variance (ANOVA) of the results of pilot study 2

A univariate ANOVA was conducted to investigate the main effects and 1st order interactions of the experimental variables on spatial quality (dependent variable) (Table 6.9). Spatial audio process (Process), programme item (ProgItem) and listener (listener) were included in the model as independent variables. The structure of the ANOVA model is shown in equation B1 (Appendix B).

The factor Process (SAP) had a significant and the largest effect on the perceived spatial quality. The main effects and 1st order interactions reveal that programme item (ProgItem) and listener (listener) both had a significant effect on spatial quality. To illustrate the most important experimental factors or interactions, figure 6.8 depicts the main effects and interactions with an effect size greater than 0.1. These are discussed in the proceeding sections.

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	822401.913[a]	158	5205.075	36.789	.000	.892
Intercept	2675851.362	1	2675851.362	18912.85	.000	.964
Process	665685.916	12	55473.826	392.088	.000	.870
ProgItem	11797.437	2	5898.719	41.692	.000	.106
Listener	26251.877	8	3281.485	23.193	.000	.208
Process * ProgItem	40005.602	24	1666.900	11.782	.000	.286
Process * Listener	73239.345	96	762.910	5.392	.000	.423
ProgItem * Listener	5217.553	16	326.097	2.305	.003	.050
Error	99745.669	705	141.483			
Total	3876749.000	864				
Corrected Total	922147.582	863				

a. R Squared = .892 (Adjusted R Squared = .868)

Table 6.9 Univariate ANOVA results output for pilot study 2.

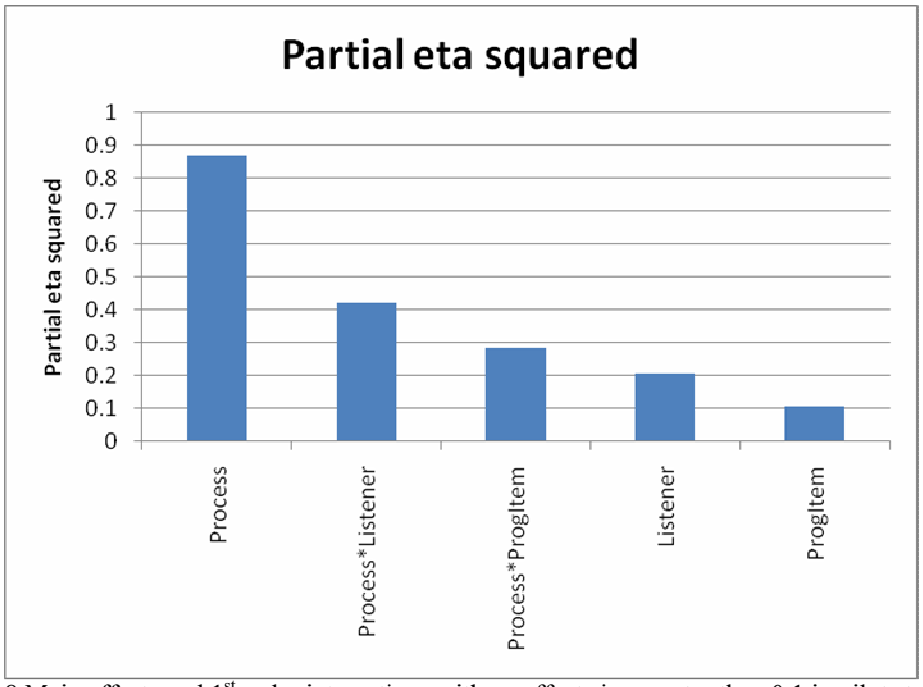


Fig 6.8 Main effects and 1st order interactions with an effect size greater than 0.1 in pilot study 2.

6.2.5.3 The influence of spatial audio process on spatial quality in pilot study 2

SAP has the largest effect on spatial quality. Figure 6.9 shows means and 95% confidence intervals for all processes and anchors, averaged across all programme items and listeners. Although this method of observation is oversimplified and hides the influence of programme item type and listener previously revealed by the ANOVA, it does allow the mean scores for individual audio processes to be observed and compared. The mean scores and confidence intervals for the SAPs cover the entire range of the test scale and mostly exhibit 95% confidence intervals narrower than 10 points (10%) of the scale.

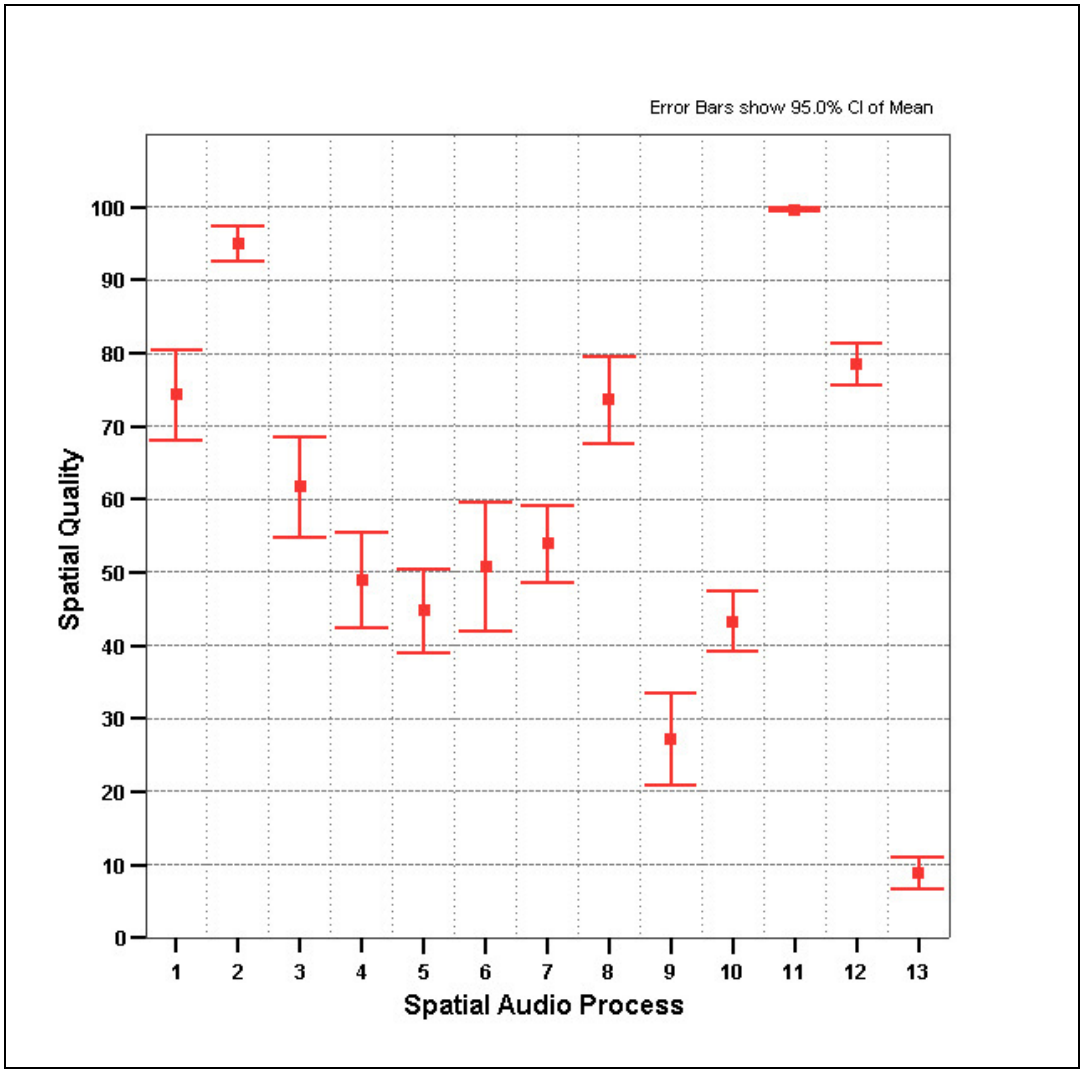


Fig 6.9 Pilot study 2 means and 95% confidence intervals for all audio processes averaged across programme item type, and listener.

Observing figure 6.9 shows that the first aim of pilot study 2, to identify SAPs that created a medium impairment to the spatial quality, was fulfilled. Eight of the SAPs were perceived as creating a medium level of impairment and hence were scored in the middle of the scale, between 30-80%. SAP 1 (Ls and Rs repositioned at -160º and 170º) is perceived as creating a larger impairment (74%) to spatial quality tha SAP 1 in pilot study 1 (Ls and Rs repositioned to ±90º) (94%). This is likely to be because the change is more severe, and repositioning the loudspeakers behind the head drastically impairs the surround image. SAP 3 (R removed) was scored at 60%. The removal of this channel was perceived as more degrading tha SAP 3 in pilot study 1 where the Ls channel was removed (89%). SAP 4 (channel order randomised) is scored at 51%; the random re-ordering of channels destroys the intended locations of the sound sources in the audio scene making it confusing. Interestingly SAP 5 (channels rotated) is scored slightly lower, although not statistically significantly ($p < 0.05$) so, at 45%. It may have been assumed that this SAP would create a lesser impairment than when the channel order is randomised; however the lower score suggests that if the audio scene remains similar to the

reference (e.g. with the majority of the sound source locations unchanged) but is rotated or skewed it is perceived as more annoying than if the image is completely random. SAP 6 (TruSurround) was scored at 48%. The TruSurround process is a virtual surround algorithm that downmixes the 5-channel audio signal to 2-channels and adds reverberation to enhance the downmixed scene and give the impression of a spacious and enveloping surround image. This is not the same as the reference and it is in fact perceived as causing a greater impairment than the 2-channel downmix. SAP 7 (Audio codec – Aud-X (80kbs)) is scored at 58%. Informal listening revealed that this multichannel audio codec reduces the spaciousness, blurs the perceived source locations and also creates a substantial change to the perceived timbral quality. Interestingly SAP 8 (3-stage cascaded audio codec - AAC (64kbs)) is scored at 74%. For this SAP the audio coding process is repeated 3 times at a bit-rate of 64kbs. The author imagined that this would create a greater impairment to spatial quality tha SAP 7. However SAP 8 uses a professional quality audio codec, AAC, whereas the Aud-X codec is a downloadable freeware codec.

Anchor recordings A-C, SAPs 11-13, are scored similarly to pilot study 1. This is particularly important in the case of SAP 13 (Anchor recording C) as it confirms its use as the low anchor. SAP 12 (Anchor recording B) is again scored above the middle of the scale and therefore it is not suitable for use as a middle anchor, and will be replaced. A possible alternative SAP could be SAP 7 (Audio codec – Aud-X (80kbs)) which is scored very close to the centre of the scale (~57%) and with narrow confidence intervals, indicating that there is reasonable agreement between the listeners.

Combination processes SAP 9 and 10 are scored the lowest (excluding the low anchor). SAP 9 (3.0 downmix with R removed) is scored at 25%. In this process not only is the audio recording downmixed from 5-channels to 3 but in addition channel R is removed. Intuitively this changes the audio scene substantially creating a large impairment to spatial quality. This could be similar to a 5-channel broadcast that has been downmixed by the listener's distribution amplifier, and reproduced through a loudspeaker system in which channel R is not connected. SAP 10 (Aud-X (80kbs) with Ls and Rs repositioned at -160º and 170º), which is scored at 45%, again compounds the impairments of two SAPs. This process could occur if a consumer reproduces an internet streamed 5-channel audio recording using a loudspeaker system that is not arranged as defined in ITU-R BS.775-1 [1992-1994].

6.2.5.4 The influence of listener on spatial quality in pilot study 2

The interaction of listener with process has the second largest effect on perceived spatial quality and again suggests that there was a difference in opinion between listeners for certain stimuli. The distribution of the subjective scores was analysed as per pilot study 1. A summary of the results of this analysis is displayed in table 6.10 (A full analysis is presented in Appendix C).

Programme item	Spatial audio process		
	Mean	Median	Remove
1	1, 2, 5, 6, 8, 9	7, 10, 11, 12, 13	3, 4
2	5, 7, 8, 10	1, 2, 3, 11, 12, 13	4, 6, 9
3	1, 4, 5, 10	2, 3, 11, 12, 13	6, 7, 8, 9

Table 6.10 Stimuli in pilot study 2 that should be removed from a database used to calibrate the QESTRAL model.

The SAPs listed in table 6.10 have been shown to exhibit large differences or a lack of consensus between listeners and should be removed if the data are to be used for calibrating the QESTRAL model.

6.2.5.5 The influence of programme item type on spatial quality in pilot study 2

The interaction of programme item type with SAP is also shown to have a significant effect on perceived spatial quality. This suggests that certain audio processes create an impairment to spatial quality that is different between programme items. A one-way ANOVA using programme item as the factor was used to statistically assess 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 6.11.

Listening position	Spatial audio process
1	1, 2, 3, 4, 6, 7, 12

Table 6.11 Stimuli which create a difference in perceived spatial quality between programme item types in pilot study 2.

Figure E3 (Appendix E) illustrates this list as means and 95% confidence intervals. For SAP 3 (R removed) programme item 3 is scored significantly differently from programme items 1 and 2. Programme item 3 is a Rock/Pop music recording with an F-F audio scene that has localisable sources surrounding the listener, mixed with background or ambient content, which creates the impression of being very enveloped. However the focus of the audio scene is contained in the front 3 channels L, C and R and hence the removal of channel R removes some of the most important sound sources and also destroys the sensation of envelopment. By comparison the commentator located in channel R of programme item 1 is also present in channel C, so the removal of R only slightly alters his location. The audience applause in this item, although considered as foreground audio content, is quite diffuse and hence its location is not important and so the gap is created by the removal of channel R is less impairing. A similar reasoning can also be assumed for programme item 2.

Another interesting example is SAP 7 (Multichannel audio codec – Aud-X (80kbs)), where programme item 1 is scored significantly differently from programme item 3. As discussed previously this multichannel audio codec reduces the spaciousness, blurs the perceived source locations and also creates a substantial change to the perceived timbral quality. Despite the changes to spatial quality it is believed that the difference between the scores is created by the change to the timbral quality. Due to the high frequency content in the applause sound sources the effect of this low quality low bit-rate

codec on programme item 1 is quite severe and annoying. However when applied to programme item 3 (and also programme item 2) the effect is not so marked.

6.2.6 Pilot study 2: conclusions

Eight of the SAPs were identified which were scored in the middle of the scale (30-80%) not covered by the SAPs investigated in pilot study 1, showing that some SAPs can create a medium impairment to spatial quality. Anchor recording B (mid anchor) was scored higher than in pilot study 1 (75%). The SAP (2.0 downmix) is therefore not appropriate for this purpose and a new process was identified. The replacement SAP is SAP 7 (multichannel audio coding A – 80kbs) which was scored very close to the centre of the scale (~57%) and with narrow confidence intervals, indicating that there was reasonable agreement between the listeners.

An additional three SAP types were investigated (virtual surround algorithms, multichannel audio codecs and SAP combinations). These were perceived by the listeners as creating a medium level of impairment to spatial quality.

Analysis of each listener's performance showed that consistency levels similar to other listening tests were achieved and that they were capable of identifying the hidden reference correctly. This further supports the use of the proposed listening test method and graphical user interface.

A univariate ANOVA of the collected data identified that, as also observed in pilot study 1, in addition to SAP, listener and programme item type influenced the perception of spatial quality. The stimuli identified as exhibiting these effects are listed in tables 6.9 and 6.10 respectively. This again supports the conclusions reached in pilot study 1 that any stimuli which elicit a statistically significant difference in opinion or lack of consensus between the listeners should be considered for removal from the database used to calibrate the QESTRAL model and also experimental variables such as listening position and programme item type should be included as an independent variables in the QESTRAL model, either by creating different calibrations for each or by using a subjective database that incorporates scores collected from both listening positions separately.

6.3 Pilot study 3 – Investigating the extent to which the spatial audio processes create changes to lower level spatial attributes

The pilot study documented in this section addresses aim (iii) set out at the beginning of this chapter. As discussed in section 3.1, it was decided that a direct method of model development would be used to develop the QESTRAL model. Therefore, to be suitable for the calibration of the QESTRAL model, the SAPs chosen for study should exhibit changes to a range of lower level spatial attributes.

6.3.1 Aim of piot study 3

The aim of pilot study 3 is to trial a method for the selection of suitable SAPs prior to conducting a large scale listening test. The method needs to show whether low-level attributes are stressed, and how even the stress distribution is, and will be deemed suitable if (i) there are no experimental or analysis problems and (ii) results are representative of pilot study 1 results.

6.3.2 Lower level spatial attributes chosen for assessment in pilot study 3

Eight lower level spatial attributes (Table 6.12) were identified as being the most important attributes of interest. The selection process was based upon the findings of the elicitation experiments discussed in section 2.2, Rumsey's scene-based paradigm and discussions amongst members of the QESTRAL project group. The eight attributes represent what could be considered as the main components of a spatial audio scene in the reproduced sound environment.

No.	Spatial attribute	Description
1	Audio scene coverage angle	The extent to which the audio scene physically surrounds the listener.
2	Individual source width	The perceived width of an individual sound source(s) within the audio scene.
3	Ensemble width	The perceived width of a group of sound sources.
4	Scene Envelopment	The perceived envelopment created by the audio scene.
5	Scene Spaciousness	The feeling of being present in the audio scene rather than absent.
6	Scene or source Distance	The perceived distance between the listener and the audio scene or sound source.
7	Scene or source Depth	The perceived distance between sound the front and rear of the entire audio scene or of a sound source(s) within an audio scene.
8	Individual source location	The perceived location of an individual sound source(s) within the audio scene.

Table 6.12 List of spatial attributes assessed in pilot study 3.

6.3.3 Stimuli and apparatus employed in pilot study 3

The stimuli and apparatus from pilot study 1 were used.

6.3.4 Methodology employed in pilot study 3

Listeners were asked to assess, at listening position 1, the differences between each stimulus and the unprocessed reference in terms of each of the eight spatial attributes. Judgements were recorded over four assessment levels (1. no changes, 2. slight changes, 3. moderate changes and 4. large changes) using pen and paper. The presentation order of the stimuli was randomised. The assessment of each spatial attribute took approximately 30 minutes. Before commencing each assessment listeners completed a short familiarisation session to ensure that they understood the task. Due to time constraints two experienced Tonmeisters from the IoSR were used rather than a large panel.

6.3.5 Discussion of the results of pilot study 3

Figure 6.10 shows how each of the investigated spatial attributes were stressed by the SAPs. Observations from figure 6.10 show that the SAPs mostly created 'no changes' or 'large changes' to the spatial attributes. Ideally, to optimise the calibration of the QESTRAL model, the SAPs should stress each attribute equally across the range of assessment levels. However only the attributes 'envelopment' and 'source location changes' come close to being stressed equally across the four assessment levels.

Figure 6.11 illustrates that there is similarity, in the distribution of the subjective scores, between the results of this study and those of pilot study 1. This suggests that the method of assessment employed in pilot study 3 could also be a useful tool for forecasting the distribution of scores that the SAPs might elicit in a more thorough assessment of their effect on spatial quality; hence it might be a useful tool for pre-selecting processes that cover the range of the test scale evenly. This will help to reduce the appearance of stimulus spacing bias in the listener scores, as discussed in section 5.2.1.1.

6.3.6 Pilot study 3: conclusions

The aim of pilot study 3 was to trial a method for the selection of suitable SAPs prior to conducting a large scale listening test. This was achieved by using a listening test method to determine the extent to which SAPs exhibit changes to a range of lower level spatial attributes. The assessment of the changes to the spatial attributes revealed that the SAPs examined (from pilot study 1) did stress all of the 8 lower level attributes tested, with 6 of the 8 being stressed, to some degree, across the full range of assessment levels. Ideally the attributes should be stressed across the entire range evenly. However the results indicate this method could be used to identify suitable SAPs for the calibration of the QESTRAL using a direct prediction method. Additionally the distribution of the results when the stimuli are assessed in this manner seems indicative of the results obtained in pilot study 1, which suggests that this method could also be used to select suitable SAPs to elicit subjective scores across the whole range of spatial quality and help to reduce stimulus spacing bias.

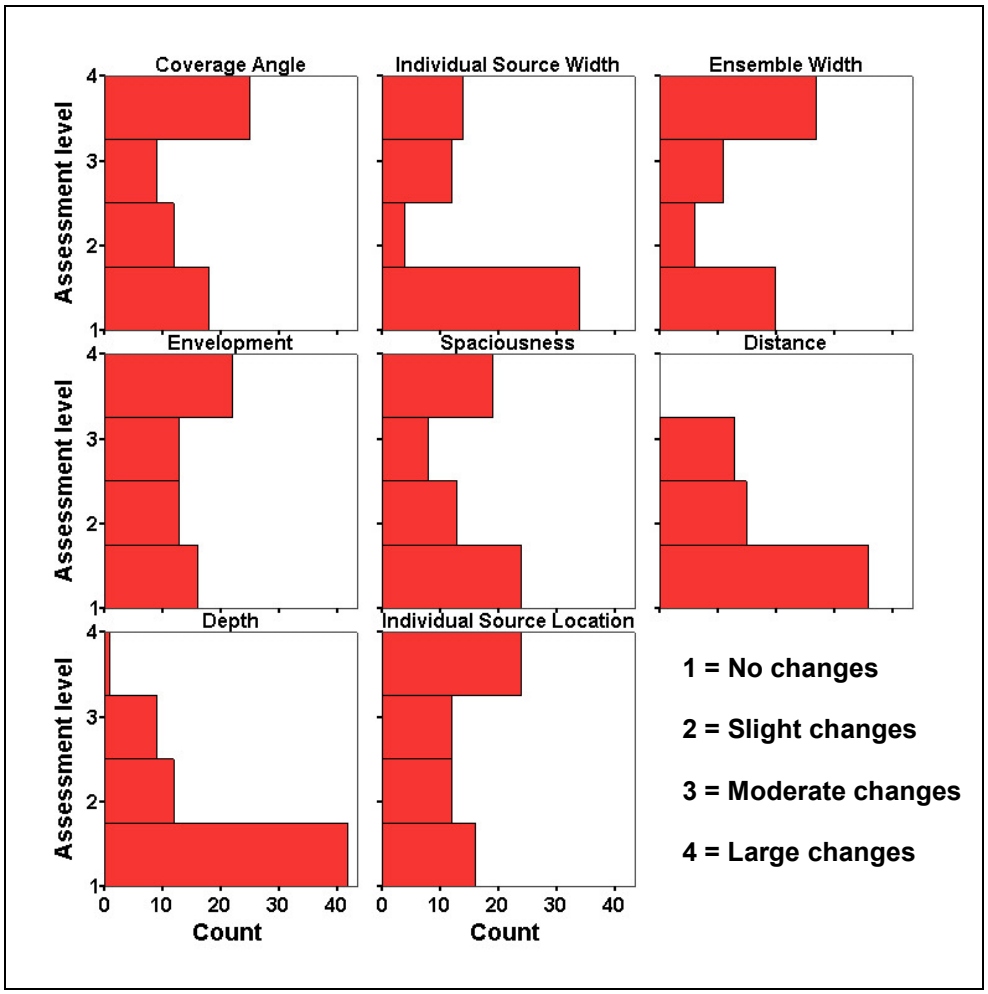


Fig 6.10 Histograms illustrating the assessment level results for the spatial attributes investigated in pilot study 3.

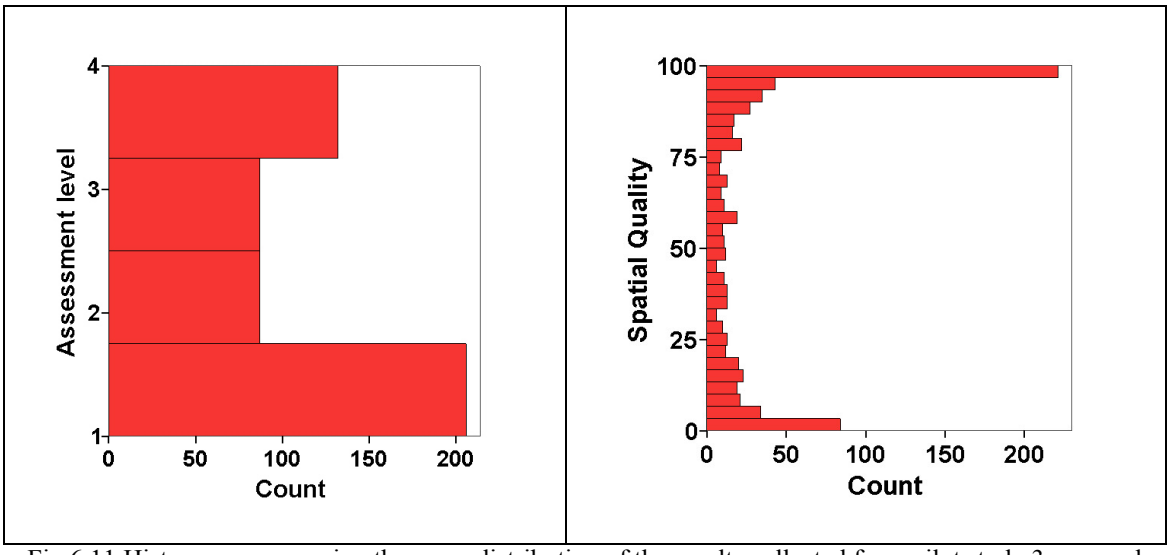


Fig 6.11 Histograms comparing the score distribution of the results collected from pilot study 3 summed across all 8 attributes (left) and pilot study 1 (right)(NB.The meaning of the y-axis between the plots is inverted).

6.4 Pilot study 4 – Is the perceived spatial quality of a stimulus influenced by its timbral quality?

The pilot study documented in this section addresses aim (iv) set out at the beginning of this chapter. As discussed in section 2.1.1, Zielinksi *et al* [2005b] observed that the audio processes they investigated degraded both timbral fidelity and spatial fidelity. Letowski [1989] also suggests that we have limited ability to evaluate quality when different domains vary simultaneously. Therefore it might be possible for listeners to become confused if a SAP causes a change in the quality across both domains, which could result in their opinion of the spatial quality being influenced by the perceived timbral quality (e.g. downmixes, bandwidth limitations and multichannel audio codecs). It is not possible to completely separate the two domains in the context of this research project. So it is important to establish if changes, created by different SAPs, to the timbral quality of an audio recording (programme item) have an affect on a listener's perception of the spatial quality.

6.4.1 Aims of pilot study 4

The aims of pilot study 4 are to determine whether:

(i) the SAPs investigated in this project affect spatial and timbral quality together or separately.

(ii) listeners can assess timbral and spatial quality separately if the two domains are separately affected.

6.4.2 Creation of stimuli for pilot study 4

This section describes the creation of the stimuli used in pilot study 4.

6.4.2.1 Programme material evaluated in pilot study 4

Three 5-channel programme items were chosen for assessment. Descriptions of the programme items are provided in table 6.13.

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, C and R. Audience applause in 360°.
2	Classical Music	F-B	Excerpt from Johann Sebastian Bach – Concerto No.4 G-Major. Wide 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 continuous front stage, including guitars, bass and drums. Main vocal in C. Harmony vocals, guitars and drum cymbals in Ls and Rs.

Table 6.13 Description of programme items evaluated in pilot study 4.

6.4.2.2 Spatial audio Processes (SAPs) investigated in pilot study 4

Thirteen different SAPs were selected to be applied to each programme item to create a number of stimuli exhibiting a range of impairments to spatial quality (Table 6.14). Some of these had been previously used in pilot studies 1 and 2; others were new additions.

No.	Spatial audio process	Description
1	Downmix 1	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.
2	Downmix 2	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.
3	Multichannel audio coding	Audio codec (160kbs)
4	Channel rearrangements	L and R reversed
5	Inter-channel level mis-alignment	L, C and R -6dB quieter than Ls and Rs
6	Inter-channel out-of-phase	C 180° out-of-phase
7	Channel removal	Ls removed
8	Spectral filtering	500Hz HPF on all channels
9	Inter-channel crosstalk	1.0 downmix in all CH
10	Combination	L and R re-positioned at -10° and 10° + Ls and Rs re-positioned at -170° and 160°
11	Anchor recording A	High Anchor - Unprocessed reference.
12	Anchor recording B	Mid Anchor – Audio codec (80kbs)
13	Anchor recording C	Low Anchor – 1.0 downmix reproduced asymmetrically by the rear left loudspeaker only.

Table 6.14 List of spatial audio processes investigated in pilot study 4.

The processes were chosen primarily to provide an example of each type of SAP investigated in this research project (see table G1) but also with the intention that they would elicit listener assessments covering the full range of the test scale. The processes were chosen via an informal listening session conducted by the author. All stimuli were loudness equalised using the method described in section 6.1.2.3. This corresponded to a playback level of approximately 75-80dB $L_{AEQ(1\text{-}3mins)}$.

6.4.3 Apparatus employed in pilot study 4

The apparatus for pilot study 4 (Fig 6.12) was similar to that used in pilot study 1, with additional loudspeakers for SAP 10 (Table 6.14). Bang and Olufsen Beolab 3 loudspeakers (Frequency response: 50 – 20,000 Hz [Bang & Olufsen, 2011]) were used in all cases. Listeners selected stimuli and recorded their responses using a laptop situated at the listening position. Prior to each test all channel gains were calibrated individually to have the same sound pressure level, at the listening position, using a pink noise signal. Not shown in the diagram is an acoustically transparent but visually opaque curtain, used to disguise the loudspeaker positions and type from the listener.

6.4.4 Methodology employed in pilot study 4

The listeners sat one test each at listening position 1 (see Fig 6.12). The listeners were instructed to assess the spatial quality and timbral quality of each stimulus compared against an unprocessed reference on alternate pages of the GUI (the full listener instructions are given in Appendix A). The order in which listeners assessed spatial or timbral quality was alternated. As in the previous pilot studies, presentation order was randomised and a preliminary familiarisation session was employed.

Each test consisted of a single judgement of the spatial quality and timbral quality of each stimulus and lasted approximately 30-40 minutes. Seventeen Tonmeisters from the Institute of Sound Recording (IoSR) at the University of Surrey took part in the test.

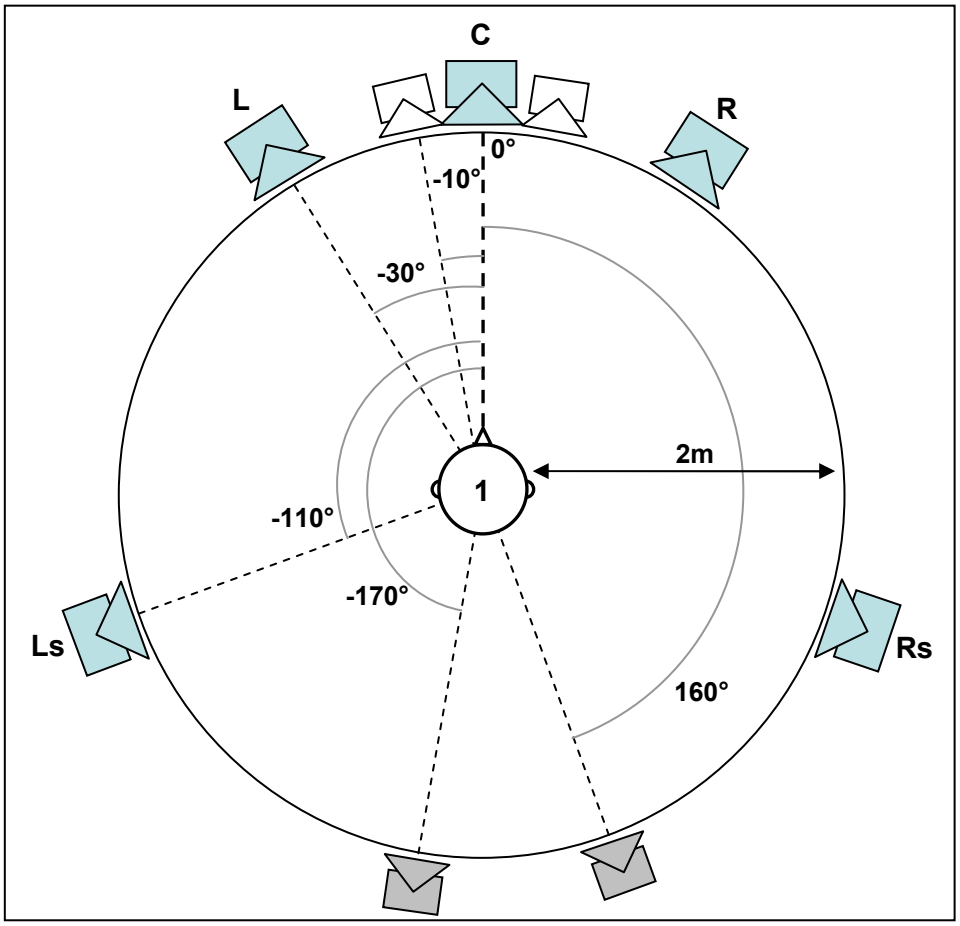


Fig 6.12 Schematic illustrating the listening position and loudspeaker positions employed during plot study 4. Loudspeakers labelled L, C, R, Ls and Rs indicate the 3/2 loudspeaker array used as the reference system. Other loudspeaker positions indicate those employed for SAP 10 (see Table 6.13).

6.4.5 Discussion of the results of pilot study 4

This section presents and discusses the results of pilot study 4.

6.4.5.1 Analysis of Variance (ANOVA) of the results of pilost study 4

A univariate ANOVA was conducted to investigate the main effects and 1st order interactions of the experimental variables on spatial quality (dependent variable) (Table 6.15). Spatial audio process (Process), Assessment type (Assessment), programme item type (ProgItem) and listener (listener) were included in the model as independent variables. The structure of the ANOVA model is shown in equation B1 (Appendix B).

Tests of Between-Subjects Effects

Dependent Variable: Quality

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1378848.736[a]	309	4462.294	18.669	.000	.814
Intercept	6075916.347	1	6075916.347	25419.98	.000	.951
Process	892771.021	12	74397.585	311.259	.000	.739
Assessment	2966.054	1	2966.054	12.409	.000	.009
ProgItem	8190.799	2	4095.400	17.134	.000	.025
Listener	97798.817	16	6112.426	25.573	.000	.236
Process * Assessment	176051.829	12	14670.986	61.379	.000	.358
Process * ProgItem	67725.783	24	2821.908	11.806	.000	.177
Process * Listener	104173.654	192	542.571	2.270	.000	.248
Assessment * ProgItem	333.735	2	166.867	.698	.498	.001
Assessment * Listener	7857.541	16	491.096	2.055	.008	.024
ProgItem * Listener	11804.368	32	368.886	1.543	.028	.036
Error	315986.126	1322	239.021			
Total	8029773.000	1632				
Corrected Total	1694834.862	1631				

a. R Squared = .814 (Adjusted R Squared = .770)

Table 6.15 Univariate ANOVA results output for pilot study 4.

The factor Process (SAP) had a significant and the largest effect on spatial quality. The main effects and 1st order interactions reveal that assessment type (Assessment), programme item type (ProgItem) and listener (listener) all had a significant effect on perceived quality. The 1st order interaction of Process and Assessment had the second largest effect suggesting that the perceived quality of a SAP differed depending upon whether it was assessed for spatial or timbral quality. To illustrate the most important experimental factors or interactions, figure 6.13 depicts the main effects and interactions with an effect size greater than 0.1. The effects of Process and of the 1st order interaction of Process and Assessment are discussed below.

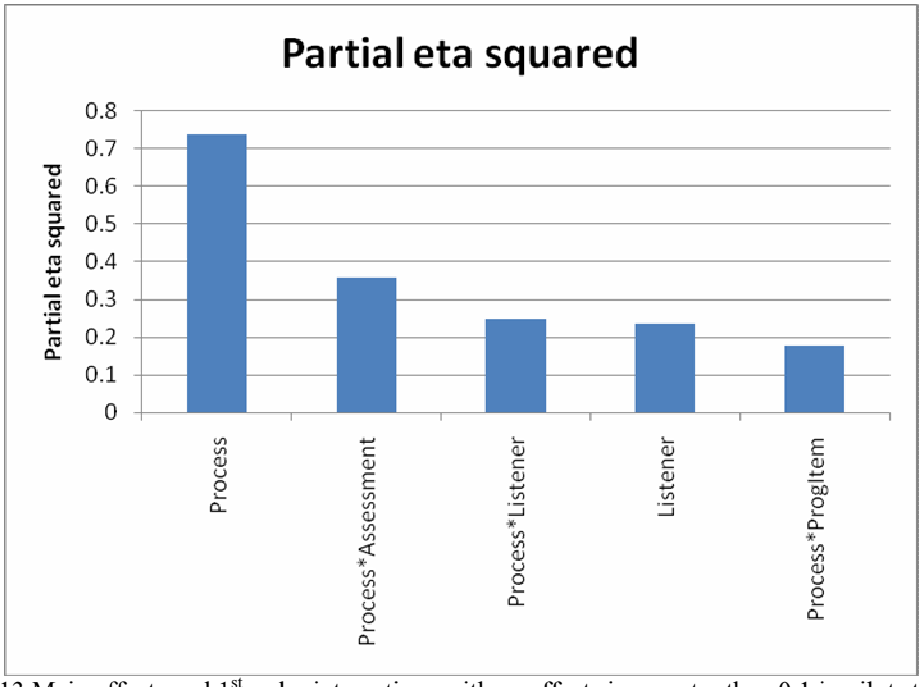


Fig 6.13 Main effects and 1st order interactions with an effect size greater than 0.1 in pilot study 4.

6.4.5.2 The influence of SAP on spatial and timbral quality in pilot study 4

Spatial audio process had the largest effect on the results. Figure 6.14 shows means and 95% confidence intervals for all SAPs and including anchor recordings, averaged across all programme items and listeners. Although this method of observation is oversimplified and hides the influence of programme item type and listener previously revealed by the ANOVA, it does allow the mean scores for individual audio processes to be observed and compared. The mean scores and confidence intervals for the SAPs cover the entire range of the test scale and mostly have 95% confidence intervals narrower than 10 points (10%) of the scale.

It is found that for the majority of SAPs there is no significant difference ($p < 0.05$) in the listeners' scores between the spatial and timbral domains. This suggests that these SAPs impaired both spatial quality and timbral quality. For example SAP 1 (2.0 downmix from 5-channels) which has been shown to impair spatial quality (scored 67% here) in pilot studies 1 and 2 impairs the timbral quality as the tonal balance of the recording is changed due to comb filtering effects caused by the combining of previously separately mixed channels together [Zielinski *et al*, 2005b].

6.4.5.3 The influence of domain assessment type in pilot study 4

The 1st order interaction of Process and Assessment has the second largest effect on the results. The SAPs were scored statistically significantly different ($p < 0.05$) when assessed for spatial quality compared to their scores when assessed for timbral quality. Hence this suggests that certain SAPs create an impairment to sound quality that is different between assessment types. A one-way ANOVA using Assessment as the factor was used to statistically assess which stimuli exhibited this effect. The processes where this test was found to be statistically significant, are listed in table 6.16.

Six SAPs were shown to be statistically significant when tested for the factor Assessment. It is for these processes that listeners perceived the spatial and timbral quality as being impaired differently. SAPs 2, 3, 8, 12 and 13 are downmixes, bandwidth limitations or multichannel audio codecs, and have been previously identified by Zielinski *et al* [2005b] as creating an overlap between the domains. The appearance of SAP 7 (Ls removed) in this analysis was slightly less obvious. However this is explainable because removing a channel from a recording reproduction would not only change the perceived spatial quality (as shown in pilot study 1) but could also change the timbre of the recording because a part of the audio mix has been removed. This evidence suggests that when a SAP affects one domain more than the other, listeners can assess them differently.

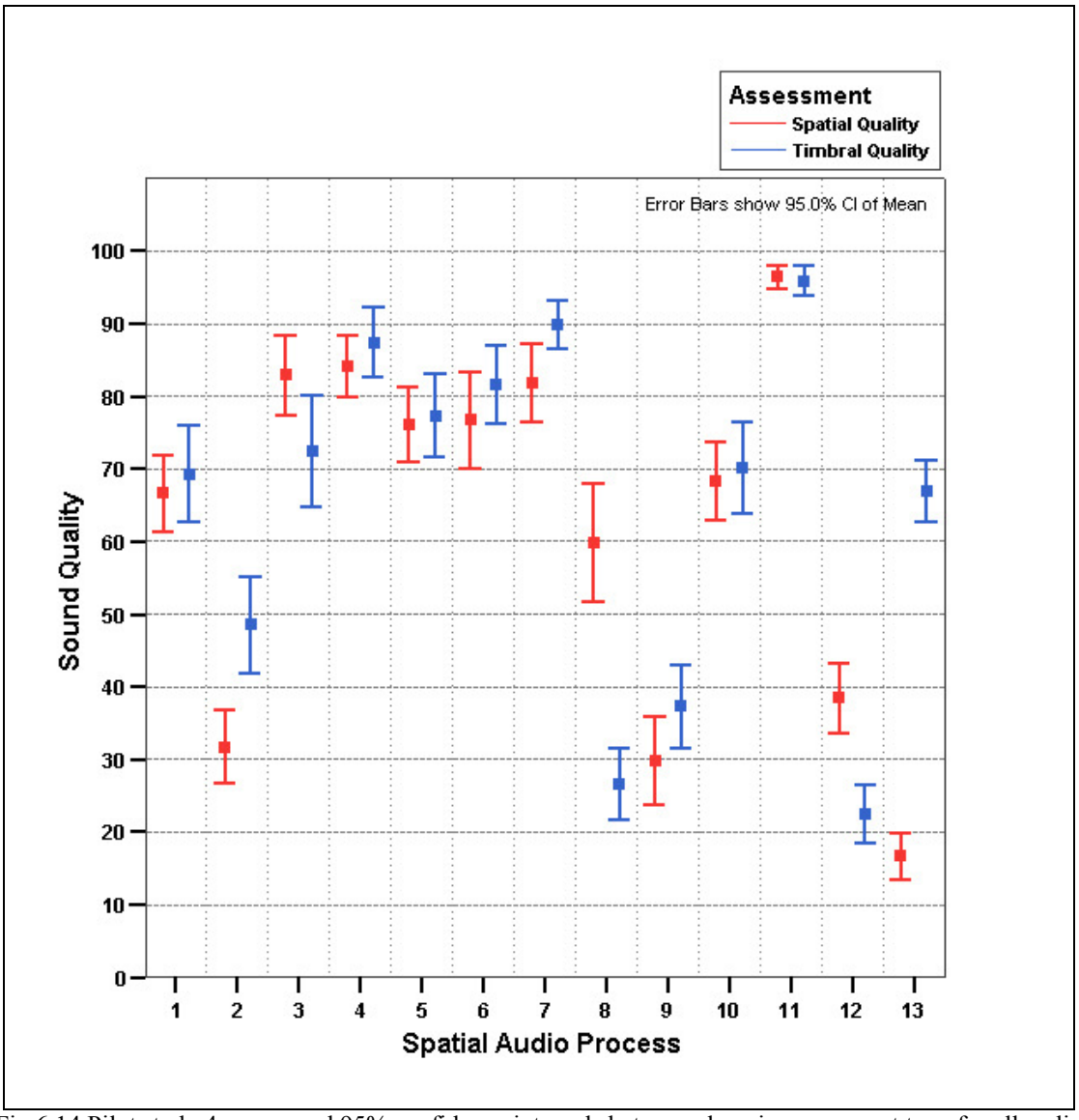


Fig 6.14 Pilot study 4 means and 95% confidence intervals between domain assessment type for all audio processes averaged across programme item type, and listener.

No.	Spatial audio process	Description
2	Downmix 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.
3	Multichannel audio coding 1	Audio codec (160kbs)
7	Channel removal 2	Ls removed
8	Spectral filtering 1	500Hz HPF on all channels
12	Anchor recording B	Mid Anchor – Audio codec (80kbs)
13	Anchor recording C	Low Anchor – 1.0 downmix reproduced asymmetrically by the rear left loudspeaker only.

Table 6.16 Stimuli which create a difference in perceived spatial quality between domain assessment type in pilot study 4.

6.4.6 Pilot Study 4: conclusions

The SAPs investigated in this study impaired both the perceived spatial quality and the perceived timbral quality of the programme items. Although it was shown that when a SAP affects one domain more than the other, listeners can assess the domains separately, in the majority of SAPs investigated

there was statistically no significant difference ($p < 0.05$) in the listeners' scores between the two domains. This suggests that the SAPs impaired spatial quality and timbral quality similarly and therefore it is possible that the perceived spatial quality of a stimulus is influenced by its timbral quality. However in the context of this research project it is not possible to separate the two domains so as the QESTRAL model aims to be a perceptual model, it is reasonable to argue that an objective metric to measure changes to timbral quality may be useful to predict the subjective spatial quality scores collected from the SAPs (NB. George used timbral metrics in his models to predict perceived SSF and FSF, as discussed in section 4.2.2). This objective metric will used in the QESTRAL model calibration process.

6.5 Analysis of listener questionnaires

In pilot studies 1, 2 and 4 listeners were asked to complete a questionnaire at the end of the test. In pilot studies 1 and 2 the listeners were asked to indicate how difficult they found the task, on a ten point scale, one being easy and ten being hard. In pilot study 4, listeners were asked to indicate how easy/hard they found the task of scaling spatial quality and also timbral quality, this time using a five point scale; one being easy and five being hard. The opinions of the listeners, particularly regarding the difficulty of the task of scaling spatial quality, are of interest for the development of a robust and usable test paradigm.

6.5.1 Questionnaire results

The results from these questionnaires are displayed in figures 6.15 to 6.17 as means and 95% confidence intervals. Fig 6.15 shows that listeners found the task of scaling spatial quality in pilot study 1 moderately difficult. The slightly lower mean value for listening position 2 may suggest that the listeners found the task easier at listening position 2. However using a one-way ANOVA the difference was found not to be statistically significant ($p < 0.05$) and therefore this suggests that, although the means are slightly different, the listeners found the task equally difficult for both listening positions. The confidence intervals are relatively wide for both, covering approximately 50% of the scale for both tests. The scores ranged from 2 to 8. This can most likely be attributed to the small number of listeners used.

Figure 6.16 shows that listeners found the task in pilot study 2 slightly easier than that in pilot study 1, although they cannot be directly compared, because mostly different listeners were used. The confidence intervals cover approximately 40% of the scale with scores ranging from 1 to 8.

The opinions collected in pilot study 4 show that the listeners found the assessment of timbral quality easier than that of spatial quality (fig 6.17). Although there was only a small difference shown between the mean values, a one-way ANOVA reveals that this difference is statistically significant (p

< 0.05). Interestingly the mean value for the difficulty of assessing spatial quality here (mean value = 3.1) was similar to that seen in pilot study 2 (mean value = 3).

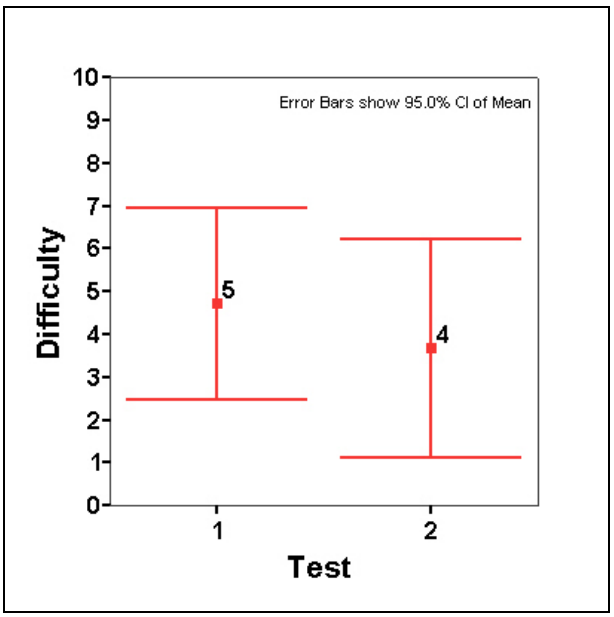


Fig 6.15 Listener opinion of the difficulty of assessing spatial quality at listening positions 1 and 2 in pilot study 1.

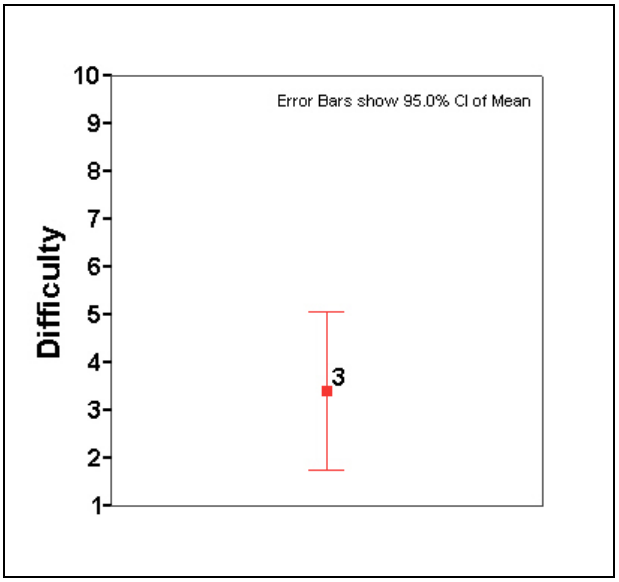


Fig 6.16 Listener opinion of the difficulty of assessing spatial quality in pilot study 2.

6.5.2 Analysis of listener questionnaires: conclusions

In pilot studies 1, 2 and 4 listeners were asked to complete a questionnaire at the end of the test. In pilot studies 1 and 2 the listeners were asked to quantify how difficult they found scaling spatial quality. In pilot study 4, listeners were asked to interpret how easy/hard they found the task of scaling spatial quality and also timbral quality. The results of these questionnaires established that listeners

found the task of scaling spatial quality easy to moderately difficult at both listening positions and found assessing spatial quality slightly more difficult than timbral quality.

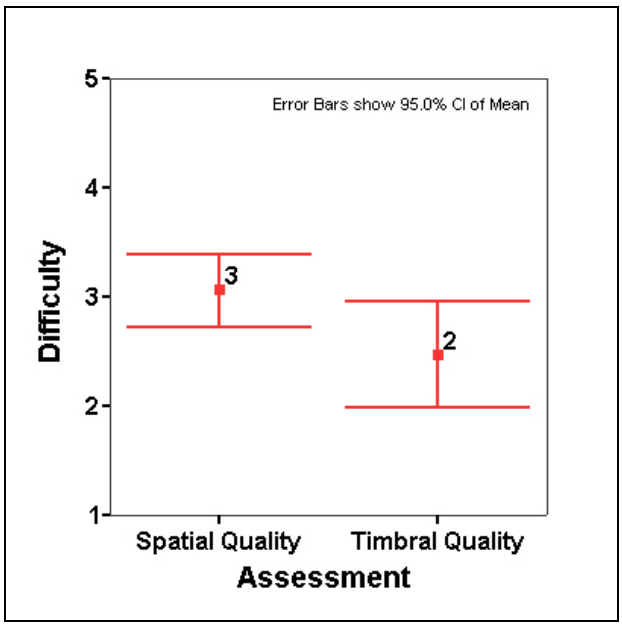


Fig 6.17 Listener opinion of the difficulty of assessing spatial quality and timbral quality in pilot study 4.

6.6 Summary and conclusions

This chapter described and discussed four listening tests conducted as pilot studies with the aims of:

(i) determining the suitability of the proposed listening test method and GUI,

(ii) assessing the difficulty of the task required of the listening test subjects at two listening positions using a wide range of different SAPs,

(iii) investigating the extent to which the SAPs create changes to lower level spatial attributes, and thus their suitability for the development of the QESTRAL model using a direct prediction method,

(iv) addressing the question raised in section 2.1.1 about whether changes to timbral quality might affect the assessment of spatial quality,

(v) identifying and investigating variables in the experiments that influence perceived spatial quality, and determine their relevance for calibrating of the QESTRAL model,

(vi) selecting suitable SAPs for use as indirect anchors.

Pilot studies 1 and 2 tested the use of the listening test method and graphical user interface (GUI) design proposed in section 5.3. It was tested with a wide range of different SAPs applied to a varied selection of different programme items at two listening positions. Analysing the listeners' performance in both studies, found that consistency levels similar to other listening tests were achieved. This indicates that the listeners could use the GUI to consistently assess the spatial quality of the stimuli

investigated. Therefore the test method and GUI are deemed suitable for the reliable assessment of SAPs in a large scale listening test.

The influence of different SAPs on the perceived spatial quality was discussed in the results of both pilot studies 1 and 2. The SAPs evaluated created impairments to spatial quality across the entire range of the test scale. Suitable SAPs for use as indirect anchors were indentified in pilot study 2 (Table 6.17)

Anchor	Anchor description
Anchor recording A	High Anchor - Unprocessed reference
Anchor recording B	Mid Anchor - Audio codec (80kbs)
Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only

Table 6.17 Description of indirect anchor recordings.

A univariate ANOVA of the collected data showed that, in addition to SAP, listener, listening position and programme item influenced the perception of spatial quality. The interaction of listener with process had the second largest effect (after SAP) on perceived spatial quality and this finding suggests that there was a difference in opinion between listeners for certain stimuli. The stimuli listed in tables 6.4 and 6.9 exhibit a statistically significant difference in opinion or lack of consensus between the listeners and are deemed to have unreliable score averages. Therefore, as the database used to calibrate the QESTRAL model will consist of SAP score averages, stimuli where this effect is observed should be considered for removal. The analysis method is described in Appendix C.

The interactions of both listening position and programme item with SAP were also shown to have a large effect on perceived spatial quality. This suggests that certain SAPs created an impairment to spatial quality that was different at the second listening position (LP2) than the first (LP1) and also different between programme items. These SAPs, listed in tables 6.5, 6.6 and 6.10 respectively, will also have unreliable means and therefore listening position and programme item should be included as separate variables in the QESTRAL model. This could be achieved either by creating different calibrations for each or by using a subjective database that incorporates scores collected from both listening positions separately.

In pilot study 3 a method was developed for the selection of suitable SAPs prior to conducting a large scale listening test. This was achieved by using a listening test method to determine the extent to which SAPs exhibit changes to a range of lower level spatial attributes. As a direct method of model development is being used, this is important for the models validity. The assessment of the changes to the spatial attributes revealed that the SAPs examined (from pilot study 1) did stress all of the lower level attributes tested, with 6 of the 8 being stressed, to some degree, across the full range of assessment levels. Additionally the distribution of the results when the stimuli were assessed in this manner was indicative of the results obtained in pilot test 1, suggesting that this method could also be

used to select SAPs to elicit subjective scores across the whole range of spatial quality. This method will be used to select optimal SAPs for a large scale listening test.

The aims of pilot study 4 were to identify whether the SAPs investigated in this project affect spatial and timbral quality together or separately and whether listeners can assess timbral and spatial quality separately when the two domains are separately affected. The results showed that when a SAP affects one domain more than the other, listeners do assess them differently. However in the majority of SAPs investigated in pilot study 4 there was no significant difference in the listeners' scores between the two domains. This suggests that these SAPs impaired spatial quality and timbral quality similarly and therefore it is possible that the perceived spatial quality of a stimulus is influenced by its timbral quality. To address this observation an objective metric capable of measuring changes to timbral quality will be included in the QESTRAL model calibration process.

In pilot studies 1, 2 and 4 listeners were asked to complete a questionnaire at the end of the test. Interestingly the results of these questionnaires established that listeners found the task of scaling spatial quality easy to moderately difficult at both listening positions and found assessing spatial quality slightly more difficult than timbral quality. However, as discussed above, analysing the listeners' responses has shown that it is possible for them to make reliable and consistent assessments of spatial quality.

Chapter 7 – Subjective assessment of spatial quality

Chapter 6 demonstrated the suitability of the proposed listening test method. This chapter describes and discusses the results of two large scale listening tests which use the developed listening test method to collect a reliable database of listener scores characterising the effects on perceived spatial quality of a large and varied range of 48 SAPs. This database will be used to calibrate the QESTRAL model. The aims of these listening tests are to:

(i) determine the effects of a wide range of SAPs on perceived spatial quality at two listening positions,

(ii) establish how the collected subjective data should be treated for calibrating the QESTRAL model;

   a. Determine which test variables should be included separately in the subjective database during the calibration process.

   b. Identify the most reliable subjective data for the calibration.

One of the aims for the QESTRAL model is that it will be calibrated to evaluate the spatial quality at two listening positions (LP1 – on-centre and LP2 – off-centre). This will require the effect of listening position to be quantified. Due to equipment restrictions, for some SAPs it will not be possible to set up both on-centre and off-centre loudspeaker arrays in order to make direct on-centre *vs* off-centre listening comparisons. Two listening tests will therefore be used: in listening test 1, for SAPs which require additional equipment (e.g. loudspeaker location alterations), the effect of listening position will be evaluated indirectly with seperate tests at listening position 1 and listening position 2 (see Fig 7.2); in listening test 2, for the less equipment-intensive SAPs, the effect of listening position will be evaluated directly to compare on-centre listening with off-centre listening (see Fig 7.3). Differences in reference conditions between listening test 1 and listening test 2 mean that a mathematical transform will be required to convert the subjective scores collected from listening position 2 in listening test 1, so that the scores from both tests can be combined into a single database.

7.1 Creation of stimuli for listening tests 1 and 2

This section describes the creation of the stimuli that were selected for listening tests 1 and 2.

7.1.1 Programme material evaluated in listening tests 1 and 2

Six 5-channel audio recordings were selected for the listening tests. Using the same criteria as in pilot studies 1 and 2, the different programme items were chosen with the intent of spanning a representative range of ecologically valid audio recordings, likely to be listened to by typical

audiences of consumer multichannel audio, while also covering typical genres and spatial audio scene types. Programme items 1 – 3 will be used in listening test 1 and programme items 4 – 6 will be used in listening test 2. Descriptions of the programme items are provided in table 7.1.

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, C and R. Audience applause in 360°.
2	Classical Music	F-B	Excerpt from Johann Sebastian Bach – Concerto No.4 G-Major. Wide 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 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 I've Got My Love To Keep Me Warm. Live music performance. Wide front stage, ambience from room and/or audience in the 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, SFX and Music in L, R, Ls, and Rs.

Table 7.1 Description of programme items evluated in listening tests 1 and 2.

7.1.2 Spatial audio processes (SAPs) investigated in listening tests 1 and 2

Forty-eight different SAPs were chosen to be applied to the programme items, to create a large number of stimuli exhibiting a range of impairments to spatial quality that would be typically encountered by consumers. The selection was informed by the results of the pilot studies dicussed in chapter 6 and discussions amongst the QESTRAL project group. The selection method fulfilled the criteria of the stimulus selection method described in pilot study 3. The results of this are illustrated in figure F1 (Appendix F) and established that the SAPs selected stressed a wide range of different spatial attributes and also spanned the range of the spatial quality scale. A full list of the chosen SAPs is given in table G1 (Appendix G), and can be divided into 12 groups (table 7.2).

Group	Process type
1	Down-mixing from 5 CH
2	Multichannel audio coding
3	Altered loudspeaker locations
4	Channel rearrangements
5	Inter-channel level mis-alignment
6	Inter-channel out-of-phase
7	Channel removal
8	Spectral filtering
9	Inter-channel crosstalk
10	Virtual surround algorithms
11	Combinations of 1-10
12	Anchor recordings

Table 7.2 Spatial audio process groups investigated in listening tests 1 and 2.

In listening test 1 forty SAPs (not including anchor recordings) were chosen for evaluation (Table G2) using programme items 1-3. Listening test 2 employs twenty SAPs (not including anchor recordings) using programme items 4-6 (Table G3). All stimuli were loudness equalised using the method described in section 6.1.2.3. This corresponded to a playback level of approximately 75-80dB $L_{AEQ(1\text{-}3mins)}$.

7.1.3 Indirect anchors employed in listening tests 1 and 2

Three indirect anchors were included in both listening tests. The use and selection of suitable anchors was discussed in chapters 5 and 6. Descriptions of the anchor recordings are given in table 7.3. All anchor stimuli were loudness equalised using the method described in section 6.1.2.3 to a comfortable listening level of approximately 75-80dB $L_{AEQ(1\text{-}3mins)}$.

Anchor	Anchor description
Anchor recording A	High Anchor - Unprocessed reference
Anchor recording B	Mid Anchor - Audio codec (80kbs)
Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only

Table 7.3 Description of anchor recordings employed for listening tests 1 and 2.

7.2 Graphical user interface employed for listening tests 1 and 2

The GUI developed and tested in chapters 5 and 6 was employed for listening tests 1 and 2 (Fig. 7.1).

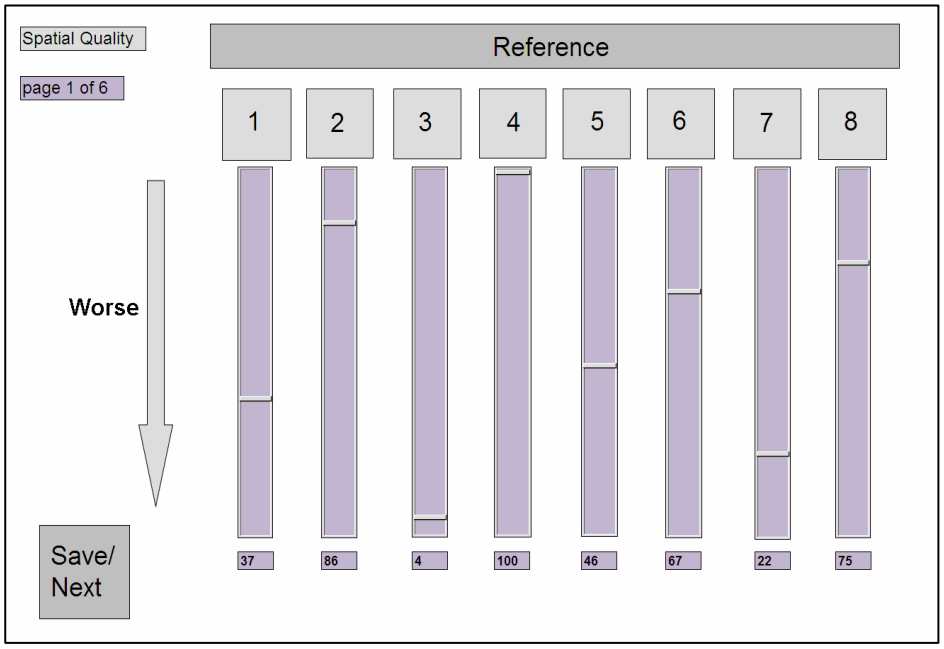


Fig. 7.1 Graphical user interface employed for listening tests 1 and 2.

7.3 Apparatus employed for listening tests 1 and 2

Both listening tests were conducted at the Institute of Sound Recording in a listening room which meets ITU-R BS.1116-1 [1997] requirements.

In listening test 1 a single 5-channel loudspeaker system was used as a reference system. The loudspeakers were arranged in 3/2 stereo configuration according to the requirements described in ITU-R BS.775 [1992-1994] (Fig. 7.2). A number of additional loudspeakers were also employed, when required, for SAPs 10 to 13 (see table G2). Bang and Olufsen Beolab 3 loudspeakers (Frequency response: 50 – 20,000 Hz [Bang & Olufsen, 2011]) were used in all cases. Not shown in figure 7.2 is an additional array loudspeaker system used for SAP 27 (see Table G2) and an acoustically transparent but visually opaque curtain, was used to conceal the loudspeaker positions and types from the listener.

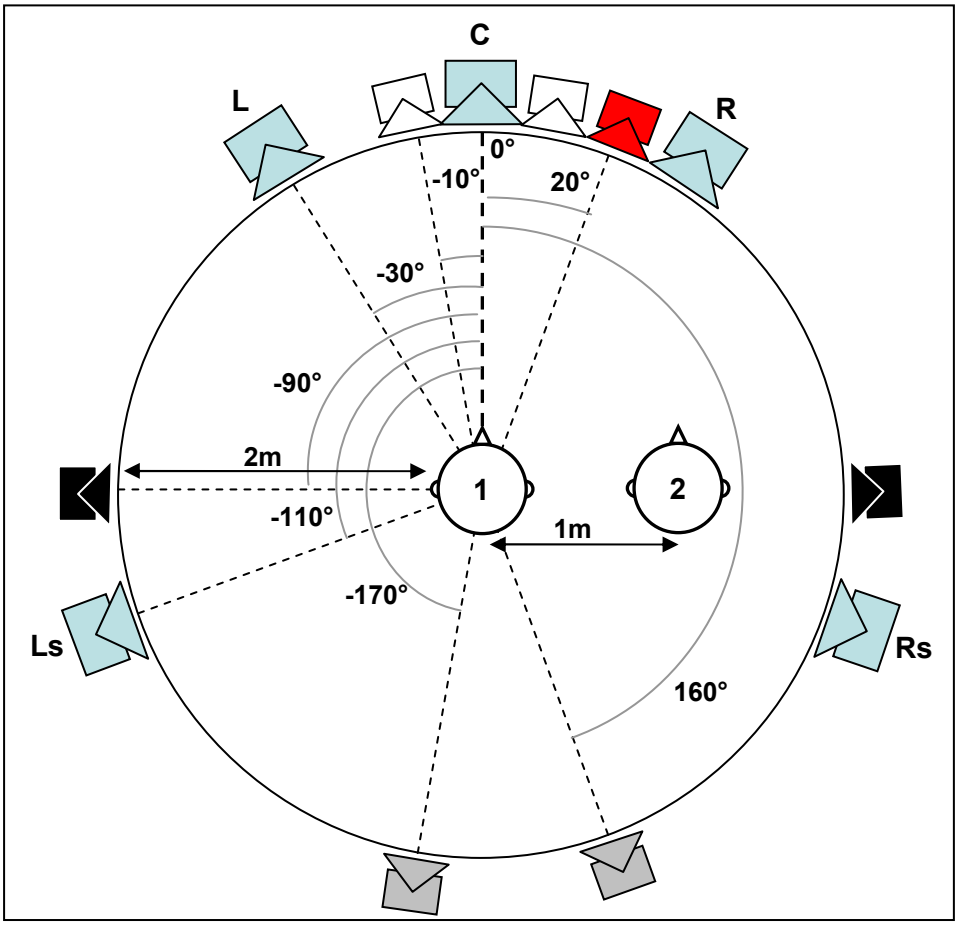


Fig 7.2 Schematic illustrating the listening positions and loudspeaker positions employed during listening test 1. Loudspeakers labelled L, C, R, Ls and Rs indicate the 3/2 loudspeaker array used as the reference system. Other loudspeaker positions indicate those employed for processes 10-13 (see Table G2). Also included in the diagram are listening positions 1 (centre) and 2 (off-centre).

In listening test 2 two 5-channel loudspeaker systems were used, one as a reference system (LP1) and one to provide an off-centre listening position (LP2) for comparison. Each loudspeaker system was arranged in a 3/2 stereo configuration according to the requirements described in ITU-R BS.775

[1992-1994] (see Fig. 7.3). Again, Bang and Olufsen Beolab 3 loudspeakers and an acoustically transparent but visually opaque curtain were used.

Listeners selected stimuli and recorded their responses using a laptop situated at the listening position. Prior to each test all channel gains were calibrated individually to have the same sound pressure level, at listening position 1, using a pink noise signal (NB. The off-centre system was calibrated separately from a listening position at its centre).

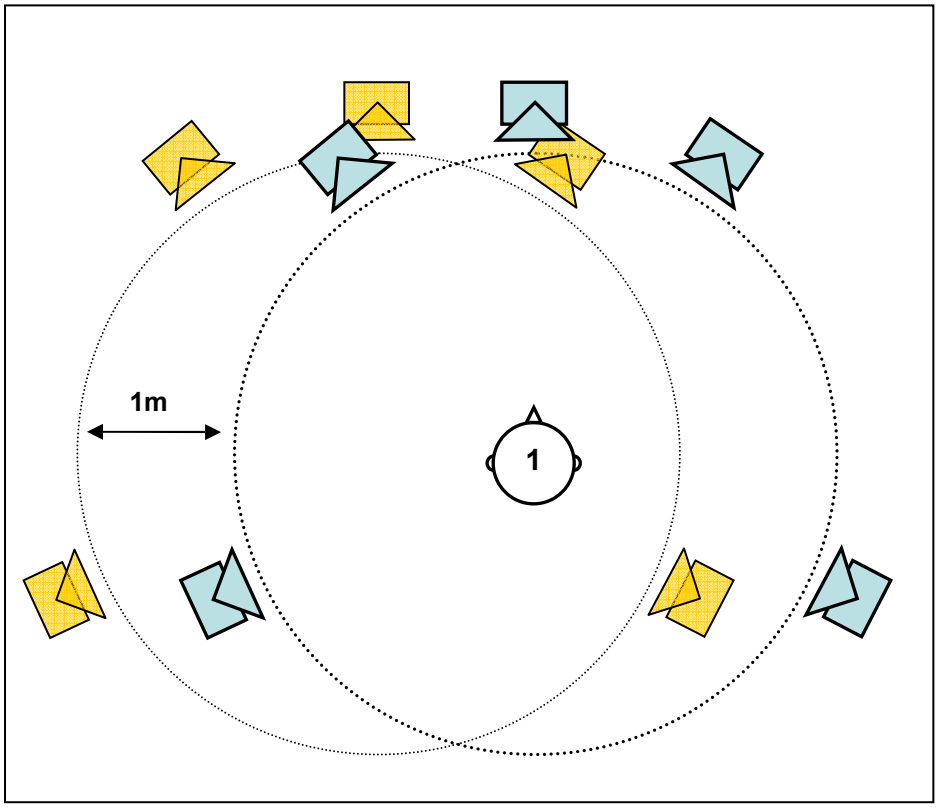


Fig 7.3 Schematic illustrating the listening position and loudspeaker positions employed during listening test 2. The blue coloured loudspeakers represent the 3/2 loudspeaker array used as the reference system. The orange coloured loudspeakers represent the 3/2 loudspeaker array used as the off-centre system.

7.4 Listening test 1

This section describes and discusses the aims, methodology and results of listening test 1.

7.4.1 Aims of listening test 1

The aims of listening test 1 are to:

i) determine the effect of a wide range of SAPs on perceived spatial quality when applied to programme items 1, 2 and 3,

ii) investigate how listeners score the SAPs differently when listening on-centre (LP1) and off-centre (LP2),

iii) identify which test variables in listening test 1 have an influence on the perceived spatial quality. This will be achieved by statistical analysis of the results.

7.4.2 Methodology employed for listening test 1

To collect subjective data a full factorial experimental method was used whereby the listeners assessed every stimulus in every condition. This meant that each listener was required to assess a large number of stimuli. In order to avoid listener fatigue the stimuli were blocked into 4 sessions, each including 10 SAPs (as shown in tables G4-7), resulting in 8 tests over two listening positions per listener. The presentation order of the stimuli within each session was randomised. Listeners assessed the 10 SAPs as well as the three indirect anchors with all 3 programme items, which created a total of 48 stimulus assessments per session. One session consisted of the test and a repeat of the test, and lasted approximately 30 minutes. Before commencing each session listeners completed a familiarisation using the GUI. This enabled them to hear, and to practice the assessment of, each stimulus featured in the session. Fourteen Tonmeisters or other experienced listeners from the Institute of Sound Recording (IoSR) at the University of Surrey took part in the test, each completed the sessions in order as per figure H1. The instructions given to each listener are shown in Appendix A.

7.4.3 Discussion of the results of listening test 1

This section describes the results of listening test 1.

7.4.3.1 Assessment of listener performance in listening test 1

Each listener's responses were assessed, so that the most reliable data could be selected for analysis and investigation. As discussed in chapter 6 each listener's discrimination ability was determined by conducting a one-sampled t-test on their scores for 'Anchor recording A' (high anchor – unprocessed reference). Their consistency was assessed by calculating the RMS error in their scoring of spatial quality between repeat judgements of the same stimuli. A full description of the assessments is given in Appendix I. The outcome of this analysis resulted in data from a number of listeners being removed from the results (Table 7.4).

Listening position	Session	Listeners whose data was removed
1	1	1, 3
	2	no listeners removed
	3	13
	4	no listeners removed
2	1	no listeners removed
	2	13
	3	no listeners removed
	4	no listeners removed

Table 7.4 Listeners removed from the subjective database of listening test 1 before results analysis.

7.4.3.2 Analysis of Variance (ANOVA) of the results of listening test 1

A univariate ANOVA was conducted to investigate the main effects and 1st order interactions of the test variables on perceived spatial quality (dependent variable) (Table 7.5). SAP (Process), listening position (LP), programme item (ProgItem), session and listener were included in the model as independent variables. The structure of the ANOVA model is shown by equation B1 (Appendix B).

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	10219793.9[a]	828	12342.746	114.382	.000	.908
Intercept	25250196.3	1	25250196.27	233997.3	.000	.961
Process	8454183.977	42	201290.095	1865.385	.000	.891
LP	9557.781	1	9557.781	88.573	.000	.009
ProgItem	31923.224	2	15961.612	147.919	.000	.030
Session	686.766	3	228.922	2.121	.095	.001
Listener	193256.078	13	14865.852	137.764	.000	.158
Process * LP	128159.883	42	3051.426	28.278	.000	.111
Process * ProgItem	315127.818	84	3751.522	34.766	.000	.234
Process * Session	3722.288	6	620.381	5.749	.000	.004
Process * Listener	723974.070	546	1325.960	12.288	.000	.413
LP * ProgItem	2314.362	2	1157.181	10.724	.000	.002
LP * Session	1543.444	3	514.481	4.768	.003	.001
LP * Listener	10506.681	13	808.206	7.490	.000	.010
ProgItem * Session	600.905	6	100.151	.928	.473	.001
ProgItem * Listener	26951.520	26	1036.597	9.606	.000	.026
Session * Listener	10307.496	39	264.295	2.449	.000	.010
Error	1029335.087	9539	107.908			
Total	45787221.0	10368				
Corrected Total	11249129.0	10367				

a. R Squared = .908 (Adjusted R Squared = .901)

Table 7.5 Univariate ANOVA results output for listening test 1.

The variable Process has a significant and the largest effect on spatial quality. Session is not significant. As discovered in chapter 6, the main effects and 1st order interactions reveal that listening position (LP), programme item (ProgItem) and listener all have a significant effect on spatial quality. To illustrate the most important test variables or interactions, figure 7.6 depicts main effects and interactions with an effect size greater than 0.1. These are discussed in the proceeding sections.

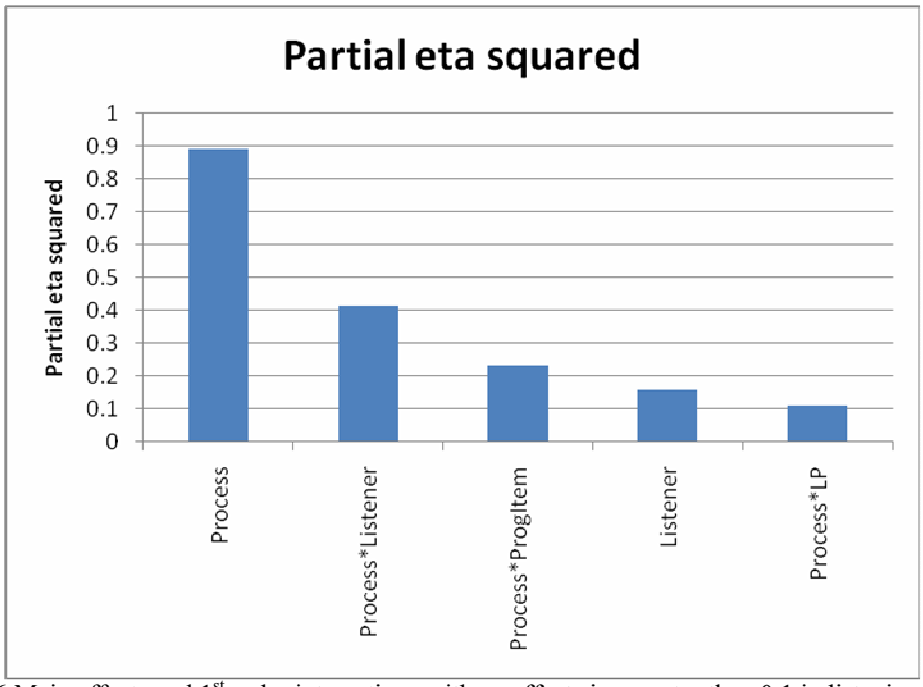


Fig 7.6 Main effects and 1st order interactions with an effect size greater than 0.1 in listening test 1.

7.4.3.3 The influence of spatial audio process on spatial quality

SAP has the largest effect on spatial quality. Figure 7.7 shows means and 95% confidence intervals for all SAPs (including the anchors), averaged across both listening positions and all programme items and listeners. Although this method of observation is oversimplified and hides the influence of these variables, it does allow the mean scores for individual audio processes to be observed and compared. To simplify analysis the results presented in figure 7.7 have been divided into SAP groups. 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.

SAP 41 (Anchor recording A – high anchor) is scored at the top of the scale, SAP 42 (Anchor recording B – mid anchor) is scored around the centre and SAP 43 (Anchor recording C – low anchor) at the bottom. SAP 1 (3/1 downmix) from group 1 creates the least impairment of all processes. In general groups 1-10 predominantly create small impairments to spatial quality while the SAPs in group 11 (combinations of 1-10) create severe impairments. This is not surprising as these processes compound the degradation created by two different SAPs. The majority of loudspeaker location change SAPs (group 3) and channel removal SAPs (group 7) do not create large impairments. Only the lowest bit rate multichannel audio coding SAPs create substantial impairments in group 2, possibly because of the combined effect of impairing both the spatial quality and timbral quality.

7.4.3.4 The influence of listener on spatial quality

The interaction of listener with SAP has the second largest effect on perceived spatial quality and, as discussed in chapter 6, this suggests that there is a difference in opinion between listeners for certain SAPs. Any stimuli which elicit a large difference in opinion or lack of consensus between the listeners

will not have reliable score averages and should be considered for removal from the subjective database. A summary of the results of this analysis is displayed in table 7.6, with a full analysis presented in Appendix C).

Listening position	Programme item	Spatial audio process
1	1	7, 28, 29
	2	7, 15, 17, 19, 23, 30, 32, 34, 40
	3	6, 7, 17, 28, 40
2	1	16, 17, 18
	2	4, 17, 23, 25
	3	4, 8, 17, 23, 25

Table 7.6 Stimuli in listening test 1 that should be considered for removal from the database.

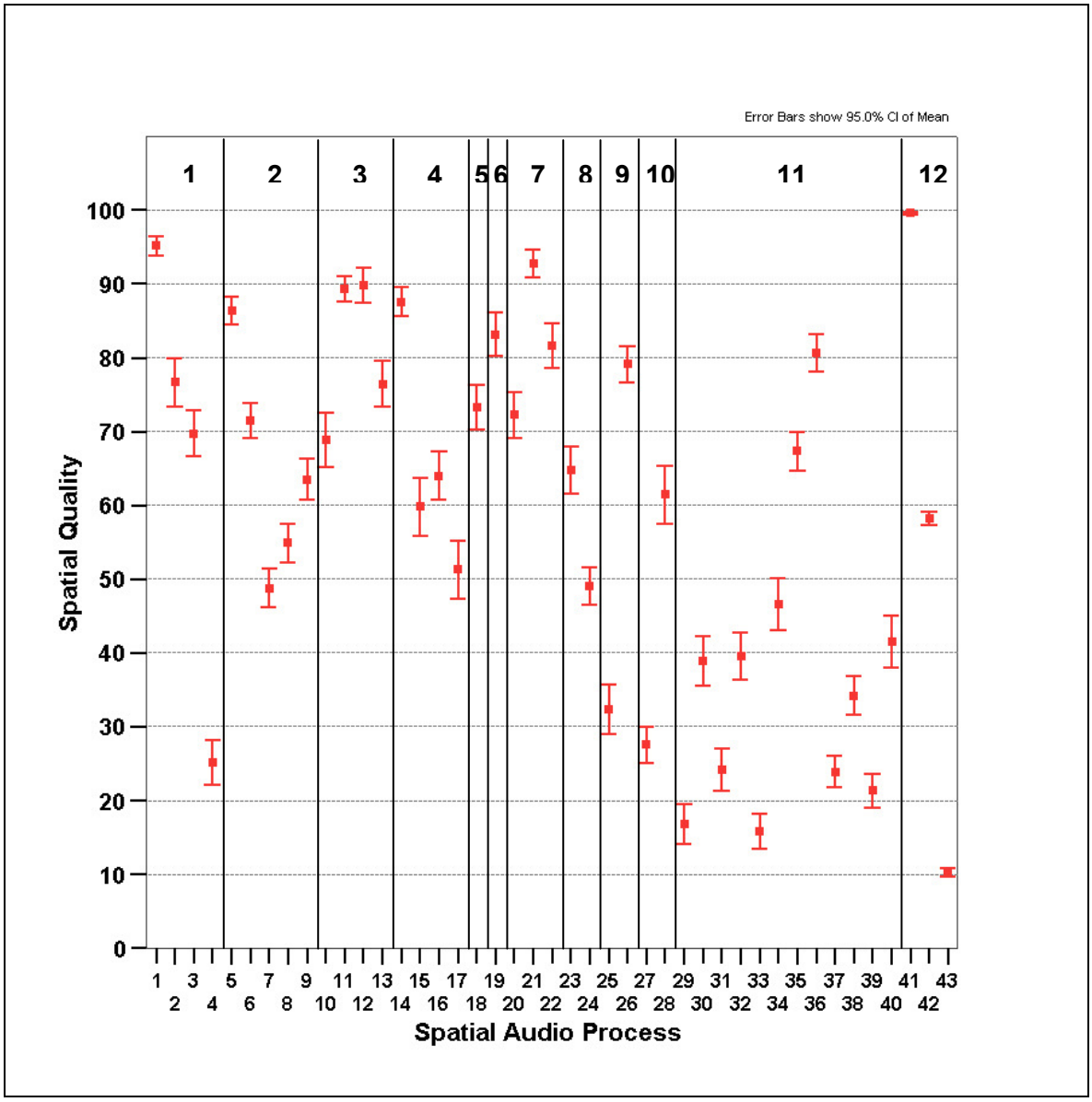


Fig 7.7 Listening test 1 means and 95% confidence intervals for all audio processes averaged across programme item type, listening position and listener.

7.4.3.5 The influence of programme item type on spatial quality

The interaction of programme item type with SAP is shown to have a significant effect on perceived spatial quality. This suggests that certain SAPs give rise to a difference in spatial quality between programme items. A one-way ANOVA using programme item as the factor was used to statistically assess 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 7.7. Figures E4 and E5 (Appendix E) illustrate this list as means and 95% confidence intervals.

Listening position	Spatial audio process
1	1, 2, 3, 5, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 22, 28, 29, 30, 34, 35, 36, 38, 40, 41, 42
2	1, 2, 3, 5, 9, 10, 11, 12, 13, 14, 15, 17, 19, 20, 22, 26, 28, 30, 35, 36, 37, 38, 39, 41, 42

Table 7.7 Stimuli which create a difference in perceived spatial quality between programme item types in listening test 1.

This difference in spatial quality can be created by differences in scene-type. For example, SAP 2 (3.0 downmix) created a far smaller impairment when applied to programme item 2 (classical) than when applied to items 1 and 3. This is likely to be because the rear channels of item 2 contain only ambient or reverberant information from the front audio scene, which is included to enhance the spaciousness or presence in the recording. As this background content is diffuse and not very localisable, downmixing it into the front channels does not create an overly degrading impairment. This is different to programme items 1 and 3 whose rear channels contain clearly identifiable foreground sources. This effect occurs at both listening positions (see Fig 7.8).

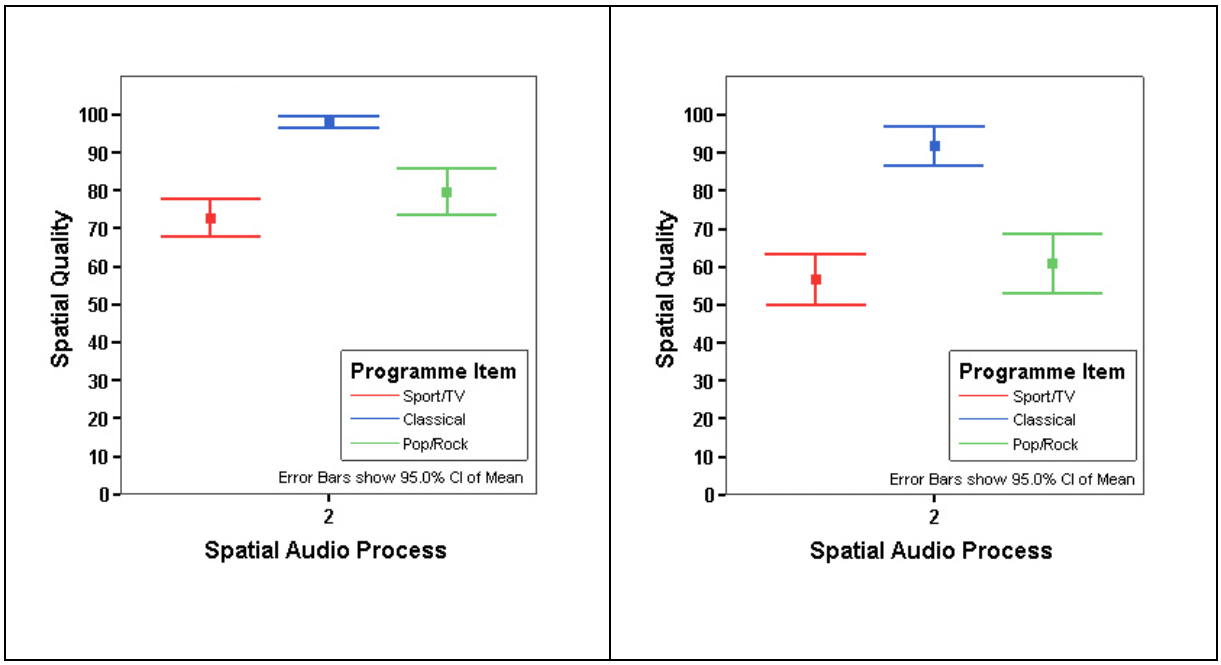


Fig 7.8 SAP 2 – Means and 95% confidence intervals illustrating an example of the influence of programme item type on the assessment of spatial quality at listening position 1 (left) and 2 (right).

This can also be influenced by the content. For example, SAP 17, where the channel order of the programme is randomly changed, created a lesser impairment to the spatial quality of programme item 1 than to programme items 2 and 3. This could be because the majority of the channels in programme item 1 contain audience applause which is very diffuse and does not carry much meaningful information in terms of location or image. Hence the channels can be re-routed at random without significant impairment to the overall spatial quality. It is likely that the perceived impairment is 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 audio image. Again this effect occurs at both listening positions (see Fig 7.9).

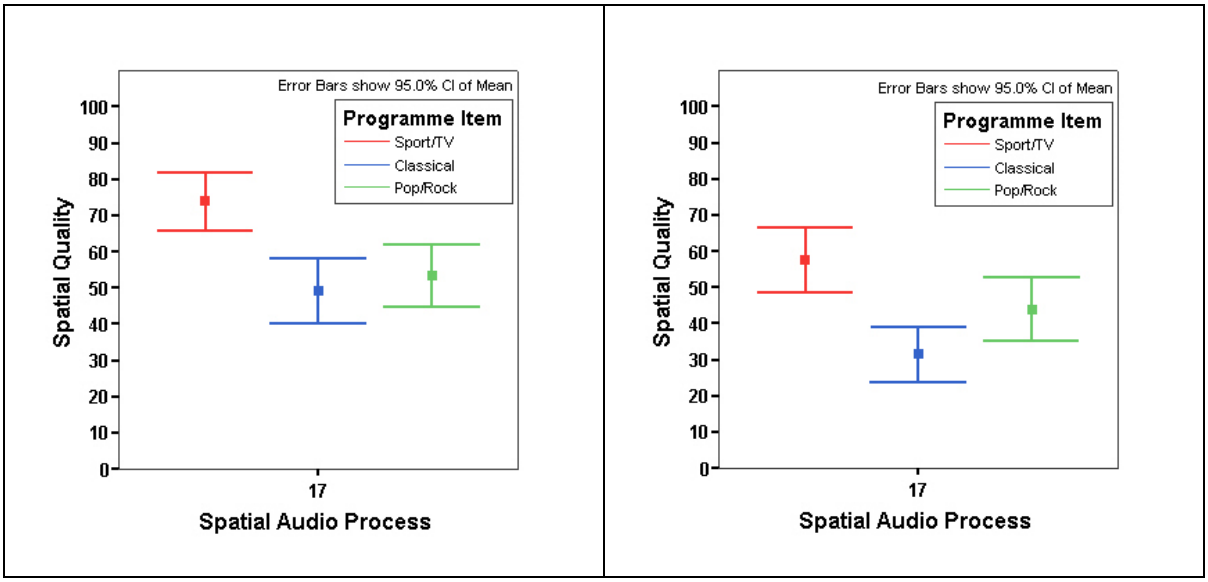


Fig 7.9 SAP 17 – Means and 95% confidence intervals illustrating an example of the influence of programme item type on the assessment of spatial quality at listening position 1 (left) and 2 (right).

7.4.3.6 The influence of listening position on spatial quality

The interaction of listening position with SAP is shown to have an effect on perceived spatial quality. This suggests that certain SAPs create an impairment to spatial quality that is different between listening positions. A one-way ANOVA using listening position as the factor was used to statistically assess 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 6.8. Figures D5 – D7 (Appendix D) illustrate this list as means and 95% confidence intervals.

Programme item	Spatial audio process
1	1, 2, 12, 13, 17, 18, 19, 21, 25, 26, 27, 29, 30, 34, 35, 36, 40, 42
2	1, 2, 5, 12, 13, 17, 19, 20, 22, 25, 27, 38, 40, 42
3	2, 3, 12, 13, 15, 20, 21, 27, 29, 30, 31, 32, 34, 35, 36, 40, 42

Table 7.8 Stimuli which create a difference in perceived spatial quality between listening positions in listening test 1.

This occurs because the physical location change in listening position between LP 1 and LP 2 alters the audio information that the listeners receive. For example, SAP 27 (Line array virtual surround) was perceived as creating a lesser impairment to spatial quality at LP1 than at LP2. This effect is observed with all three programme item types (see Fig 7.10). This occurred because the virtual surround effect created by the line array is achieved by processing the audio content and beam steering this signal behind the listener, by reflection from nearby walls, to give an impression of surrounding image. For this to work correctly it requires that the listener sits directly in front of it. However at LP2 this condition is compromised and the effect breaks down causing the SAP to be annoying.

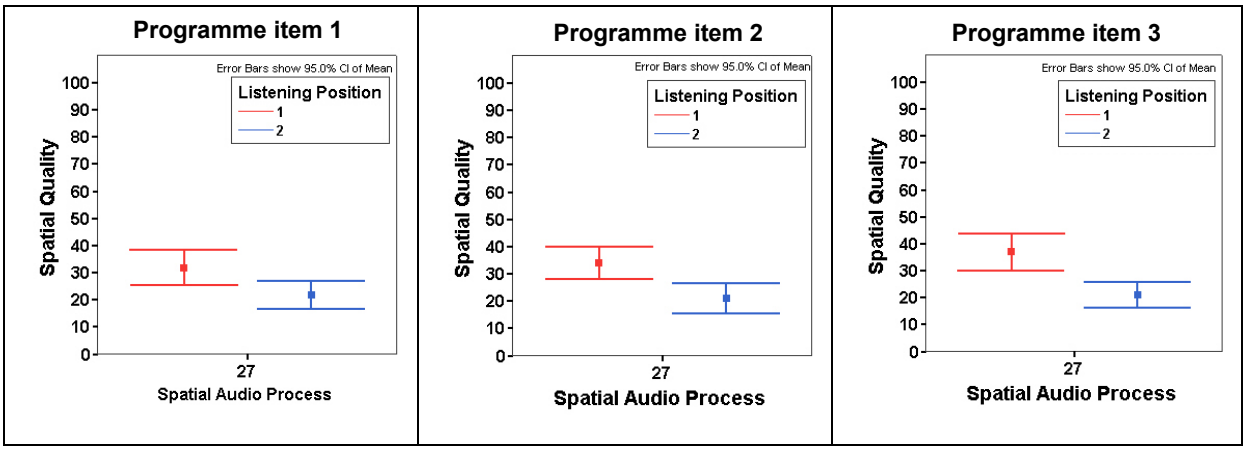


Fig 7.10 SAP 27 – Means and 95% confidence intervals illustrating an example of the influence of listening position on the assessment of spatial quality.

When the rear loudspeakers were misplaced to -90° and 90° respectively, in SAP 12, only a small impairment to spatial quality was perceived at LP1. This is possibly due to the inability of the human auditory system, as described by the 'minimum audible angle', to accurately locate sound sources positioned in the area around each ear (approximately ±90°) [Moore, 2003]. Conversely from LP2, which is closer to the right surround loudspeaker position, the misplacement of the loudspeakers is much more obvious and therefore the impairment becomes apparent and is scored lower. This effect is observed for all three programme item types (see Fig 7.11).

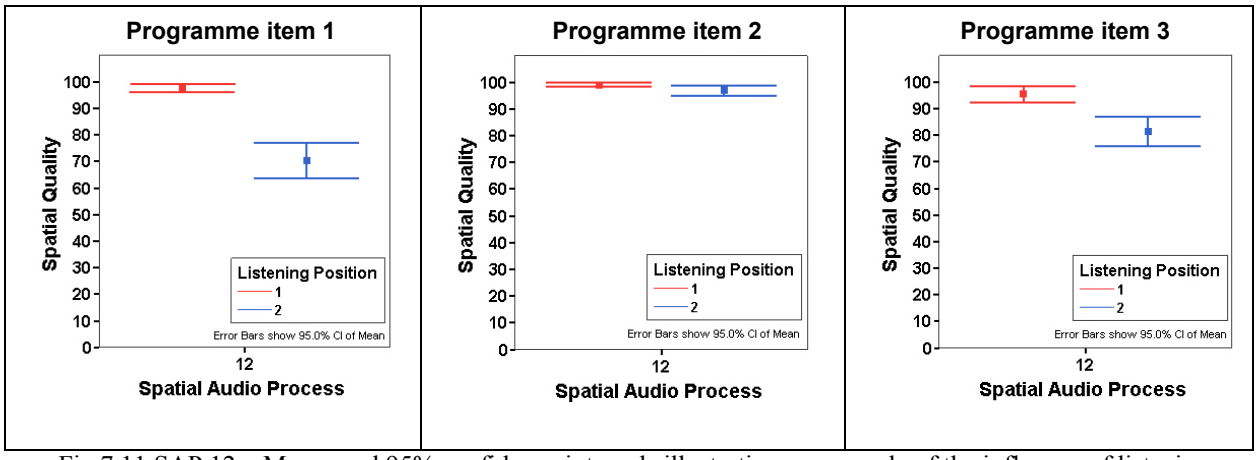


Fig 7.11 SAP 12 – Means and 95% confidence intervals illustrating an example of the influence of listening position on the assessment of spatial quality.

7.5 Listening test 2

This section describes and discusses the aim, methodology and results of listening test 2.

7.5.1 Aims of listening test 2

The aims of listening test 2 are to:

i) quantify the effect of off-centre listening from LP2 on perceived spatial quality when applied to programme items 4, 5 and 6,
ii) devise a mathematical transform to convert the subjective scores collected from listening position 2 in listening test 1,
iii) identify which test variables in listening test 2 have an influence on the perceived spatial quality. This will be achieved by statistical analysis of the results.

7.5.2 Methodology employed for listening test 2

To directly compare the perceived spatial quality when listening at LP2 with that at LP1, two 5-channel loudspeaker arrays were combined (Fig 7.3). The second loudspeaker array (used for LP2 and represented by orange loudspeakers) was arranged 1m to the left of the first array (reference system) (represented by blue loudspeakers). SAPs 1 – 20 (Table G3) were replayed through the reference system (LP1) (NB. SAPs 21 – 23 are the hidden anchor recordings) and SAPs 24 – 43 were replayed through the off-centre array (LP2).

As with listening test 1 a full factorial experimental method was used. To avoid listener fatigue the stimuli were blocked into 4 sessions, each including 10 processes (Tables G8 – 11). The presentation order of the stimuli within each session was randomised. Listeners assessed the 10 SAPs as well as 3 hidden anchors with all 3 programme items, creating a total of 48 stimulus assessments per session. One session consisted of the test and a repeat of the test, and lasted approximately 30 minutes. Before commencing each session listeners completed a familiarisation using the GUI. This enabled them to hear, and to practise the assessment of each stimulus featured in the session. Seventeen experienced listeners from the IoSR took part in the test. The order in which listeners complete the sessions was randomised. The instructions given to each listener are shown in Appendix A.

7.5.3 Discussion of the results of listening test 2

This section describes the results of listening test 2.

7.5.3.1 Assessment of listener performance in listening test 2

Each listener's responses were assessed in the same manner as listening test 1, so that the most reliable data could be selected for analysis and investigation. A full description of the assessment is given in Appendix I. The outcome of this analysis resulted in data from a number of listeners being removed from the subjective database (Table 7.9).

Session	Listeners whose data was removed
1	6, 7, 9, 16
2	3, 7, 9
3	7, 9
4	3, 7, 9, 15

Table 7.9 Listeners removed from the subjective database of listening test 2 before results analysis.

7.5.3.2 Analysis of Variance (ANOVA) of the results of listening test 2

A univariate ANOVA was conducted to investigate the main effects and 1st order interactions of the test variables on spatial quality (dependent variable) (Table 7.10). SAP (Process), listening position (LP), programme item (ProgItem), session and listener were included in the model as independent variables. The structure of the ANOVA model is shown in equation B1.

The variable Process has a significant and the largest effect on spatial quality. Session is not significant. The main effects and 1st order interactions reveal that listening position (LP), programme item (ProgItem) and listener all have a significant effect on spatial quality. To illustrate the most important test variables or interactions, figure 7.12 depicts main effects and interactions with an effect size greater than 0.1. These are discussed in the proceeding sections.

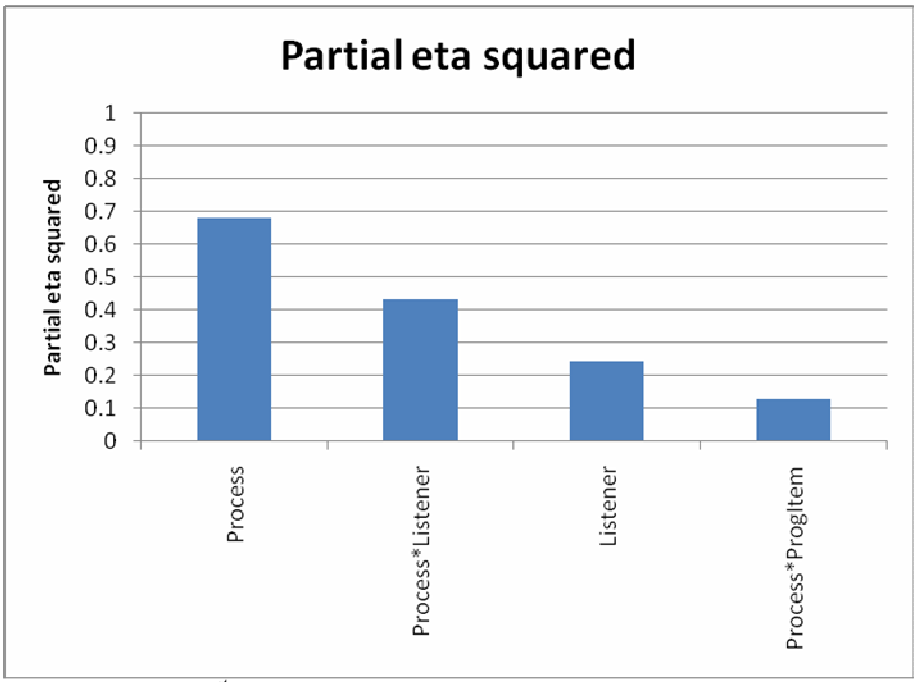


Fig 7.12 Main effects and 1st order interactions with an effect size greater than 0.1 in listening test 2.

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	4450114.973[a]	496	8972.006	59.860	.000	.861
Intercept	12145202.5	1	12145202.48	81030.58	.000	.944
Process	1539695.860	22	69986.175	466.935	.000	.682
LP	66548.515	1	66548.515	444.000	.000	.085
ProgItem	243.054	2	121.527	.811	.445	.000
Session	450.454	3	150.151	1.002	.391	.001
Listener	229905.501	14	16421.822	109.563	.000	.243
Process * LP	28885.799	3	9628.600	64.240	.000	.039
Process * ProgItem	105650.082	44	2401.138	16.020	.000	.128
Process * Session	649.055	6	108.176	.722	.632	.001
Process * Listener	547852.054	298	1838.430	12.266	.000	.433
LP * ProgItem	109.471	2	54.736	.365	.694	.000
LP * Session	.000	0	.	.	.	.000
LP * Listener	13759.516	14	982.823	6.557	.000	.019
ProgItem * Session	4675.419	6	779.237	5.199	.000	.006
ProgItem * Listener	11973.454	28	427.623	2.853	.000	.016
Session * Listener	27140.005	37	733.514	4.894	.000	.036
Error	716896.099	4783	149.884			
Total	24929962.0	5280				
Corrected Total	5167011.072	5279				

a. R Squared = .861 (Adjusted R Squared = .847)

Table 7.10 Univariate ANOVA results output for listening test 2.

7.5.3.3 The influence of spatial audio process on spatial quality

SAP has the largest effect on spatial quality. Figure 7.13 shows means and 95% confidence intervals for all processes and anchors for both LP 1 and LP 2. To allow the mean scores for individual SAPs to be observed and compared over both listening positions (LP1 in red, LP2 in blue) the scores for each stimulus are averaged across all programme items and listeners. To simplify analysis the results presented figure 7.13 have been divided into SAP groups (Table 7.2).

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. Separating the scores for LP1 (red) and LP2 (blue) illustrates how spatial quality is impaired when listening off-centre. A similar trend in the scoring of identical audio processes between LP1 and LP2 is noticed. However the range of the scores for LP2 is compressed to the lower half of the test scale. This compression is not linear, as shown in figure 7.13. The difference in perceived quality between the highest quality SAPs is large and is as much as 30% (e.g. SAP 1 circled in black), whereas the difference between the lowest rated SAPs is small, less than 5%, and is statistically not significant (e.g. SAP 18 circled in red). This smaller difference could suggest that the impairment to spatial quality created by these processes is so severe that a shift in the listening position does not influence the listener's opinion of it.

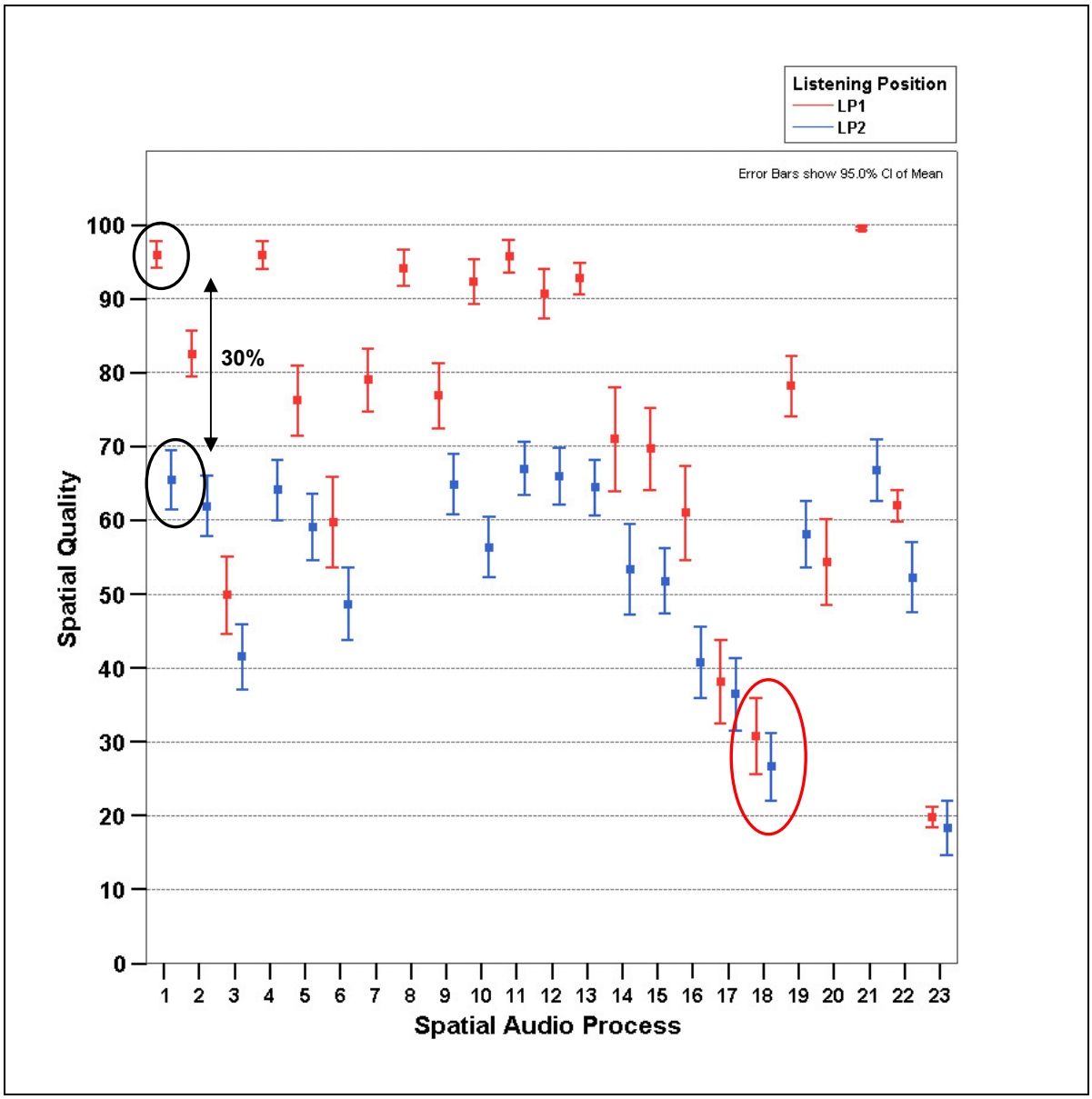


Fig 7.13 Listening test 2 means and 95% confidence intervals for all SAPs averaged across programme item type highlighting the non-linear compression in the scores of audio processes at LP2 (LP1 in red, LP2 in blue).

7.5.3.4 The influence of listener on spatial quality

Similarly to listening test 1, listeners' scores exhibited a difference in opinion and a lack of consensus for certain stimuli. This was investigated further in the same manner used in listening test 1. A summary of the results of this analysis is displayed in table 7.11 (A full summary of the analysis is presented in Appendix C).

Listening position	Programme item	Spatial audio process
1	4	3, 5, 6, 15, 16, 17, 20
	5	15, 17, 20
	6	3, 14, 15, 16, 17, 18
2	4	3, 16
	5	3, 9, 10, 16
	6	14, 17, 18

Table 7.11 Stimuli in listening test 2 that should be considered for removal from the database.

7.5.3.5 The influence of programme item type on spatial quality

The interaction of programme item type with process was again shown to have a significant effect on perceived spatial quality. A one-way ANOVA using programme item as the factor was used to statistically assess 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 7.12. Figures E6 and E7 illustrate this list as means and 95% confidence intervals.

Listening position	Spatial audio process
1	1, 2, 5, 7, 8, 12, 14, 19
2	5, 10, 14, 19

Table 7.12 Stimuli which create a difference in perceived spatial quality between programme items in listening test 2.

7.5.4 Calculating a mathematical transform to convert the scores from listening position 2 in listening test 1

One of the aims for the QESTRAL model is that it will use data collected at other listening positions to predict changes in spatial quality across the listening area. It will make these evaluations against an audio reference reproduced from a centralised listening position (i.e. LP1).

In listening test 1, the effect of listening position was evaluated indirectly with separate tests at listening position 1 and listening position 2 (see Fig 7.2); in listening test 2, the effect of listening position was evaluated directly to compare on-centre listening (LP1) with off-centre listening (LP2) (see Fig 7.3). Differences in the reference conditions between listening test 1 and listening test 2 resulted in two separate databases, one for the perception of spatial quality *vs* an on-centre reference and the other for the perception of spatial quality *vs* an off-centre reference, which could not be combined. Therefore a mathematical transform is required to convert the subjective scores collected from listening position 2 in listening test 1, so that the scores from both tests can be combined into a single database.

7.5.4.1 Transformation function

A transformation function was derived by plotting the score averages for stimuli evaluated off-centre from listening test 2, against corresponding data from listening test 1. To achieve this SAPs common to both tests were compared. However as identified in section 7.4.3.5 when a SAP was applied to programme items with different scene-types the spatial quality was perceived differently. So in consideration of this, only SAPs applied to programme items with similar scene-types were compared. Hence average listener scores for SAPs applied to programme item 5 (F-F) in listening test 2 were plotted against the corresponding and aggregated SAP scores for programme items 1 (F-F) and 3 (F-F) in listening test 1. This was repeated for F-B scene type material (programme items 2, 4 and 6). These data were plotted together (Fig 7.14) and a best-fit line was calculated, the equation of which was used as a transformation function (equation 7.1).

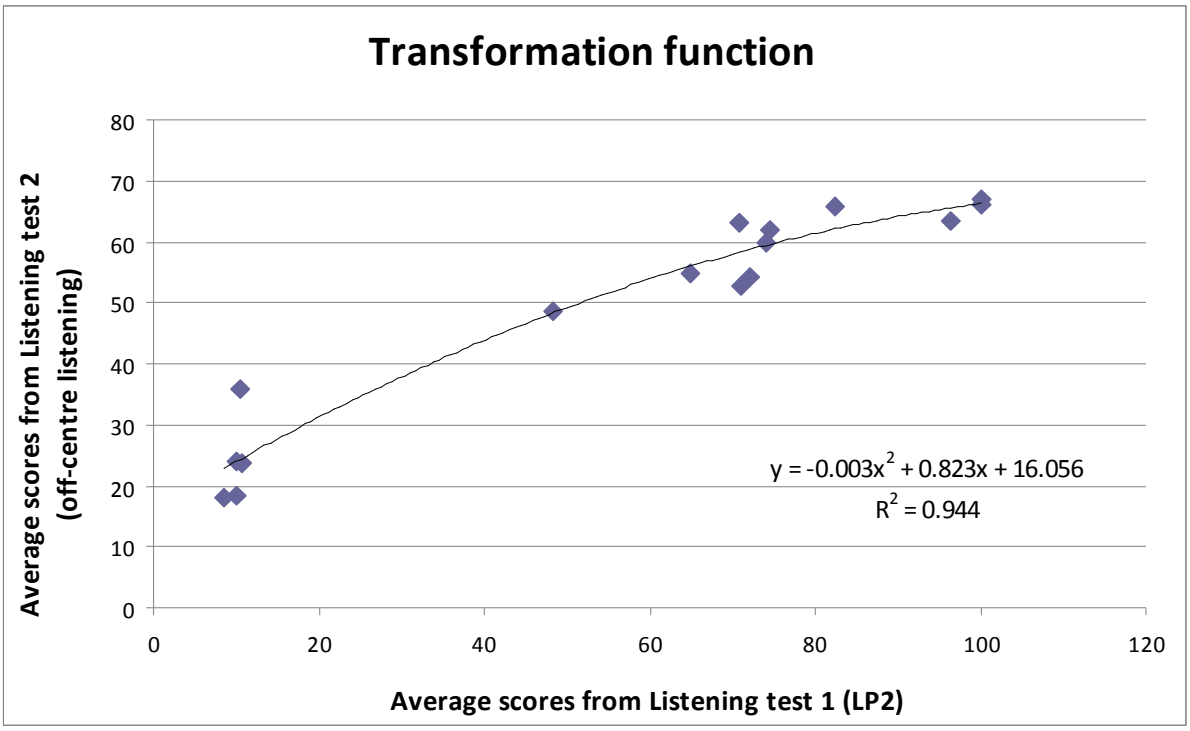


Fig 7.14 Scatterplot of average scores from listening test 2 (off-centre listening, on-centre reference) vs. average scores from listening test 1 (off-centre listening, off-centre reference) comparisons. Best fit line used to calculate 2nd order polynomial transformation function.

$$y = -0.003x^2 + 0.823x + 16.056 \qquad \text{(eq. 7.1)}$$

Where:

y = the score transformed to be with respect to an on-centre reference

x = the score from off-centre listening (LP2) in listening test 1 (off-centre reference)

7.6 The QESTRAL model subjective database

As shown by the results of both listening test 1 and 2 (and also in pilot studies 1 and 2) the scoring of perceived spatial quality was influenced by changing the listening position, the type of programme material that the SAP was applied to and differences in opinion between listeners, as well as by the SAP itself. Therefore these factors will be considered in the subjective database independently.

As has already been discussed, in sections 7.4.3.1 and 7.5.3.1, the influence of the differences in opinion between listeners leads to unreliable score averages. However this can be accounted for by analysing the data distributions of individual stimuli using a number of statistical and visual analysis techniques, the aim being to remove any stimuli where a large difference in opinion is observed and thereby identify the most reliable stimulus score averages. The results of this data screening are summarised in tables 7.6 and 7.11 and presented in full in Appendix C.

To incorporate listening position and programme item type in the calibration of the QESTRAL model independently, the stimulus score averages collected from both listening positions and all six programme items will be included separately in the database. This aims to make the calibrated model sensitive to the influence these test variables have on perceived spatial quality.

7.7 Summary and Conclusions

This chapter described and discussed the results of two large scale listening tests which used the developed listening test method to collect a reliable database of listener scores characterising the effects of a large and varied range of SAPs on perceived spatial quality, for calibrating of the QESTRAL model. The aims of these listening tests were to:

(i) determine the effects of a wide range of SAPs on perceived spatial quality at two listening positions,

(ii) establish how the collected subjective data should be treated for calibrating the QESTRAL model;

   a. Determine which test variables should be included separately in the subjective database during the calibration process.

   b. Identify the most reliable subjective data for the calibration.

Over two large scale experiments 48 SAPs were evaluated using six different programme items at two listening positions. The stimuli created impairments to spatial quality across the whole range of the test scale. The effects of these SAPs on spatial quality were examined and a number of examples were discussed. In listening test 1 listener responses were collected at an on-centre listening position (LP1) and an off-centre listening position (LP2) independently. In listening test 2 the effect of off-centre listening on spatial quality was examined and compared directly with on-centre listening; this lead to the development of a transform function which allowed the responses collected at listening position 2 (in listening test 1) to be converted and included in the subjective database.

Analysing the results of the listening tests using ANOVA it was identified that differences in listener opinion, listening position and programme item type influenced the perception of spatial quality. This had also been observed in the results of pilot studies 1 and 2. As the QESTRAL model will be calibrated as a perceptual model it was decided that it should be sensitive to the changes to perceived spatial quality created by listening position and programme item type. Therefore these variables will be incorporated into the calibration process by including separately the stimulus score averages collected at both listening positions and all six programme items. Any stimuli which elicit a large difference in opinion or lack of consensus between listeners will not have reliable score averages, and so stimuli where this effect is observed will be removed from the subjective database.

The entire database was analysed and the most reliable data were identified, leading to 308

scores which could be used for calibrating the QESTRAL model. The results of this data screening are summarised in tables 7.6 and 7.11 and presented in full in Appendix C.

Chapter 8 – Calibration of the QESTRAL model for the objective evaluation of spatial quality

In chapter 7 two large scale listening tests were discussed which were conducted to collect a reliable database of listener scores, characterising the effects of a large and varied range of SAPs on perceived spatial quality. The data was collected with the intention of using them for calibrating the QESTRAL model and were examined to determine which test variables should be included separately in the calibration and to identify which data were the most statistically reliable (308 reliable listener scores were identified for the calibration process).

This chapter describes the calibration and discusses the subsequent performance of the QESTRAL model for the automatic evaluation of spatial quality using the data collected in the listening tests discussed in chapter 7. The aims of chapter 8 are to:

i) establish if probe signals and objective metrics developed by the QESTRAL project team can be used to build a system that, after calibration against the listening test data from chapter 7, meets the target specifications proposed in section 3.3.

ii) determine if the calibrated QESTRAL model is generalisable and performs within target specifications for the prediction of spatial quality for each of the test variables (SAPs, programme items and listening positions).

8.1 Probe signals used for the prediction of spatial quality

It is currently not possible to automatically decompose the spatial scene elements of typical spatial audio recordings such as music. So the QESTRAL model evaluation scheme was designed to use probe signals specially designed to scrutinise aspects of the spatial scene. Probe signals have been shown to work successfully in similar applications [Mason, 2006][ITU-R BS.1387, 2001]. An advantage of using probe signals over commercially recorded audio is that they can be designed to emulate generic characteristics of audio recordings such as the programme items used in the listening tests 1 and 2. However their structure and characteristics can be controlled which allows changes created by a SAP to be detected and measured precisely.

Two probe signals were created by the QESTRAL project research team [Dewhirst *et al*, 2008], one to allow the QESTRAL model to measure changes, created by a SAP to spatial characteristics in the foreground stream and one for measuring these changes in the background stream (table 8.1). In the context of this study, changes in the foreground stream include changes to the locations of the sources and to the individual source width, ensemble width, source stability and

source focus for example [Rumsey, 2002], whereas changes in the background stream include changes in envelopment, scene width, spaciousness etc [Rumsey, 2002].

Probe signal	No. of channels	Description
1	5	36 pink noise bursts pairwise constant power panned from 0° to 360° in 10° increments.
2	5	Decorrelated pink noise (10 seconds in duration) replayed over all channels.

Table 8.1 Probe signals employed in the QESTRAL model.

Probe signal 1 was developed in a previous study by Dewhirst [2008] and was designed to allow the model to evaluate changes to the foreground stream. It consists of thirty-six one second pink noise bursts, positioned, using pairwise constant power panning, at 10° intervals in the horizontal plane. These are replayed sequentially from 0-360°. Probe signal 2 was designed to allow the QESTRAL model to evaluate changes in the background stream of the audio scene and consists of a 10 second burst of decorrelated pink noise replayed over all channels. This signal was designed to approximate the diffuse acoustic field of reverberant sound or room ambience. It was inspired by the work of Hiyama *et al* [2002] who reported that the spatial impression of a diffuse sound field could be reproduced from four loudspeakers corresponding to the front left and right, and left and right surround locations of a 3/2 stereo loudspeaker arrangement, and by that of George [2008] who later suggested that it was not possible to differentiate between the diffuse soundfields created by 5-channel and 4-channel uncorrelated pink noise recordings.

8.2 Objective metrics used for the prediction of spatial quality

A range of different metrics were developed by the QESTRAL project team to measure the changes in spatial quality created by the SAPs evaluated during listening tests 1 and 2. The metrics used were inspired by prior research conducted by the author and from work conducted by other researchers as discussed in chapter 4. Each metric was designed to be used with either probe signal 1, to measure changes to the foreground stream or probe signal 2, to measure changes to the background stream. In addition (as discussed in section 1.1) it was desirable for the QESTRAL model to be reproduction format independent. To achieve this, the metrics were developed to analyse the probe signals as received by a virtual binaural simulator or other virtual microphone receivers at the listening position simulated in the QESTRAL model.

8.2.1 Identification of attributes that are significantly impaired by the SAPs investigated

Suitable metrics must respond to changes in the attributes most affected by SAPs. It was therefore necessary to identify which spatial attributes had been impaired by the SAPs evaluated in listening tests 1 and 2. The results are summarised in figure 8.1.

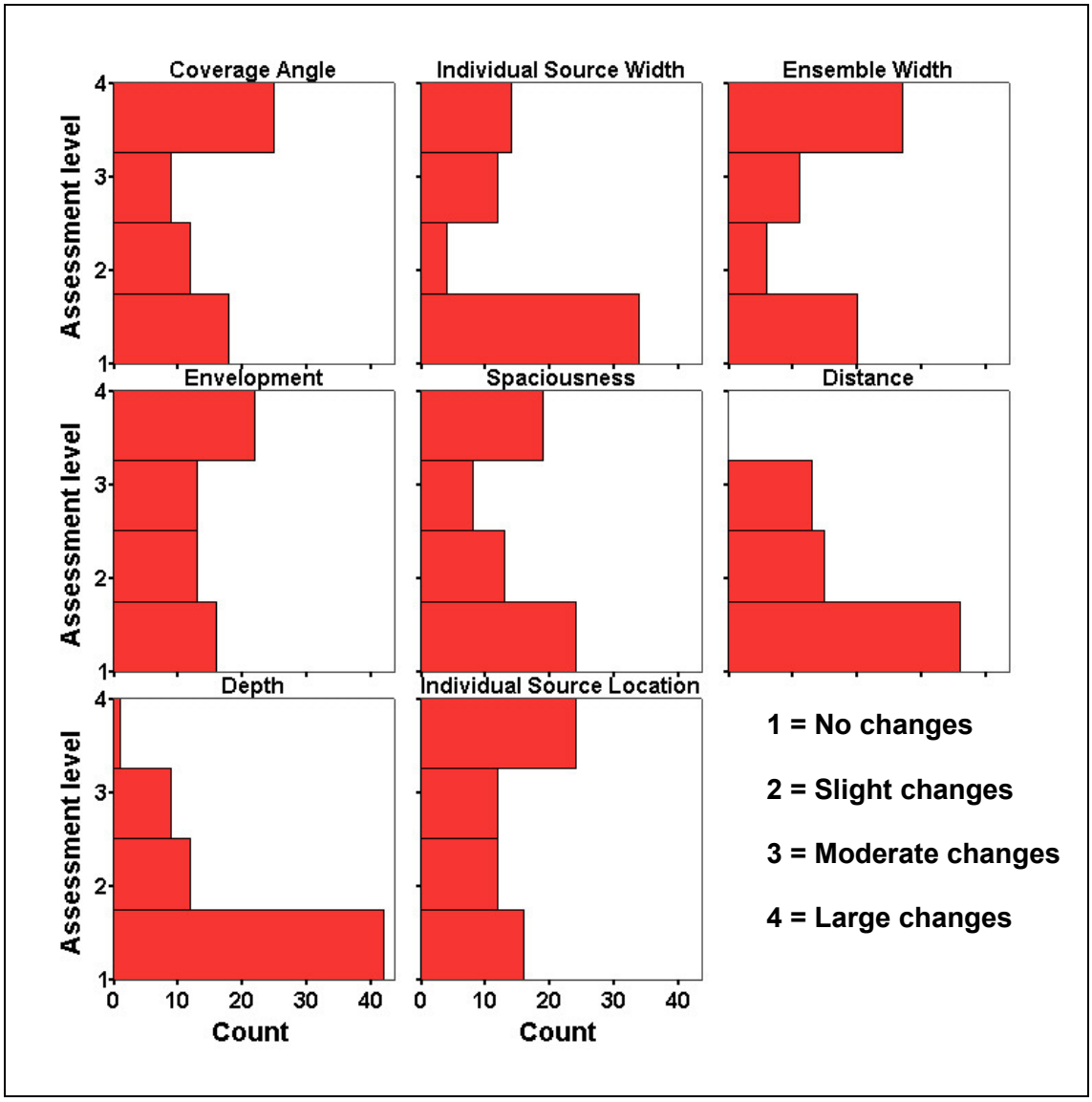


Fig 8.1 Histograms illustrating the numbers of large, moderate, slight and imperceptible impairments to each of 8 lower level spatial attributes reported in tests using the programme items and SAPs of listening tests 1 and 2.

These results show that the attributes suffering the highest number of large impairments were source location, envelopment, coverage angle, ensemble width and spaciousness. Hence metrics capable of measuring these attributes were selected. As identified in pilot study 4, a perceived change to timbral quality was created by a number of different SAP types. It was shown that the largest impairments to timbral quality were created from SAPs such as spectral filtering, multichannel audio coding and downmixing from 5-channel. Hence as in George [2009], a metric to measure changes in timbre was

included. All metrics were developed and created by the QESTRAL project research team. A discussion of their development is beyond the scope of this thesis and therefore, except in cases where this author was principally responsible, only an overview of each metric is given. Further information on the metrics and their implementation in the QESTRAL model evaluation scheme is described in Jackson *et al* [2008] and Dewhirst *et al* [2008].

8.2.2 Description and optimisation of the objective metrics

This section describes the objective metrics used in the QESTRAL model.

8.2.2.1 Metrics based upon IACC

Three metrics were based on measuring interaural cross-correlation (IACC) using a method developed by Mason [2006]. As discussed in chapter 3, IACC has been employed by a number of researchers, to measure perceived envelopment, ensemble width and spaciousness. It measures the similarity of the left and right channels of a binaural signal.

Two IACC metrics were calculated using a virtual dummy head at the listening position with two different head rotations: a 0° head rotation ('IACC0') and a 90° head rotation ('IACC90') using probe signal 2. A preliminary comparison of these metrics with the subjective spatial quality scores showed that 'IACC0' had a correlation (r) of 0.65 and 'IACC90' had a correlation (r) of 0.51. The product of both IACC calculations was used as an additional metric ('IACC0*IACC90'). This had been shown to work successfully in previous work conducted by this author [Conetta, 2007] and George [George, 2009]. 'IACC0*IACC90' showed a correlation (r) of 0.62 with spatial quality.

To optimise the IACC metrics, inspiration was drawn from concert hall acoustics research [Beranek, 1996] and George [2009], who employed a band limited (or octave band) measure of IACC where a mean value of IACC was calculated from three frequency bands; 500Hz, 1kHz and 2kHz. Beranek showed how this type of IACC measurement correlated well with a listener's spatial impression of a concert hall. George employed this method in his models predicting frontal spatial fidelity (FSF), surround spatial fidelity (SSF) and envelopment. However despite these previous findings, there was no guarantee that a band limited or octave band method of measuring IACC would have similar success for evaluating spatial quality. So an investigation of the metric IACC0 was undertaken to ascertain which of the 22 frequency bands had the highest correlation with the subjective scores (the results of this study are presented in figure 8.2). This revealed that 9 bands between 570Hz and 2160Hz produced the highest correlation to spatial quality. Based upon this a bandwidth-limited IACC metric was designed, which was calculated from the mean IACC value of the 9 bands. This metric ('IACC0_9band') had a higher correlation (r = 0.71) than the original broadband (22 band) IACC0 metric (r = 0.65). Interestingly the range of frequencies is similar to those used in concert hall acoustics. The same idea was also employed for 'IACC90' ('IACC90_9band') (r = 0.53)

and 'IACC0*IACC90' ('IACC0*IACC90_9band') (r = 0.66). Both optimised and original IACC metrics will be employed in the calibration process.

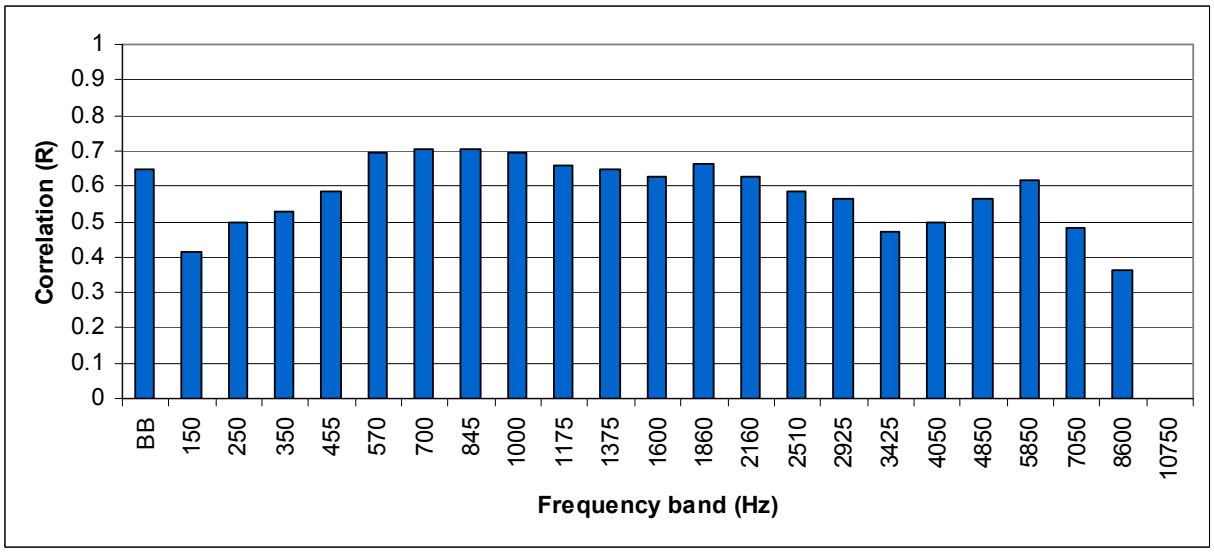


Fig. 8.2 IACC individual frequency band IACC correlation (r) with spatial quality, compared with broadband mean IACC (BB) correlation (r) with spatial quality.

8.2.2.2 Metrics based upon localisation

From his localisation model Dewhirst [2008] (see section 4.1.2) developed a metric ('Mean_Ang_Diff') which was capable of measuring the average degree of changes to source locations [discussed in Jackson *et al*, 2008]. The metric was developed alongside probe signal 1 and hence changes to source locations are calculated using this probe signal. 'Mean_Ang_Diff' is a measure of the mean absolute displacement of each noise burst from probe signal 1 created by the SAP when compared against their intended locations in the reference. 'Mean_Ang_Diff' was shown to have a good correlation (r) to spatial quality (r = 0.61).

A preliminary model employing 'Mean_Ang_Diff' [Conetta *et al*, 2008] revealed that this metric could not predict accurately the perceived spatial quality arising when certain SAPs were applied to audio recordings with an F-B scene type, such as classical recordings (e.g. programme item 2). This was a programme item dependent problem stemming from the difference between F-B and F-F scene types.

As described 'Mean_Ang_Diff' measures the change in location of 36 noise bursts in 360°. The measured source location changes created by a SAP such as a 3.0 downmix are quite large because the sources in the rear scene (rear loudspeakers) are re-positioned in the front scene (front loudspeakers). When this SAP was applied to programme items with an F-F scene type (i.e. programme items 1, 3 and 5) the change measured by 'Mean_Ang_Diff' related closely to the perceived response of the listeners, because they perceived the re-positioning of the sources from the rear scene to the front scene and scored it appropriately. However when applied to programme items

with an F-B scene type (i.e. programme items 2, 4 and 6), the change measured was not representative, because the rear channels contain ambient or reverberant energy and hence the repositioning of the rear sources was not perceived as overly degrading (NB. A discussion of the perceptual differences created by a 3.0 downmix is provided in section 7.4.3.5). As approximately half of the subjective data was collected using F-B scene type programme items it was decided that a more intelligent or generic metric, which could incorporate the subjective differences between these different scene types, should be developed.

Two additional metrics ('Mean_Ang_Diff_FrontWeighted' and 'Mean_Ang_Diff_Front60') were proposed which take greater account of the differences between scene types. A description of these metrics is given in table 8.2.

Metric	Correlation (r) to spatial quality	Description
Mean_Ang_Diff_FrontWeighted	0.73	The mean or maximum absolute change to localisation, compared to reference localisation for the 36 noise bursts, with a linear weighting of decreasing importance from 0° applied to each angle.
Mean_Ang_Diff_Front60	0.67	The mean or maximum absolute change to localisation, compared to reference localisation for 7 noise bursts between 0-30° and 330-350°.

Table 8.2 Descriptions of front biased angle difference metrics.

To demonstrate the performance of these new metrics, figure 8.3 compares them against 'Mean_Ang_Diff' for measuring SAPs that involve changes to the rear scene (e.g. 3/1 downmixes, 3.0 downmixes and altering the locations of rear loudspeakers) only. The subjective scores collected when these SAPs were applied to F-F scene type programme items are shown in red and F-B scene type programme items in blue. The three plots in figure 8.3 show that these types of SAPs create no perceived change in spatial quality when applied to F-B scene type material, as illustrated by the blue samples having the same subjective score as the reference recordings (square). However 'Mean_Ang_Diff' measured large differences between the SAPs and the reference, which is shown by the vertical stacking of the blue samples. These differences are reduced when measured using 'Mean_Ang_Diff_FrontWeighted', and disappear when measured using 'Mean_Ang_Diff_Front60'. The additional metrics have a superior correlation to spatial quality, so were included in the calibration process and 'Mean_Ang_Diff' was removed.

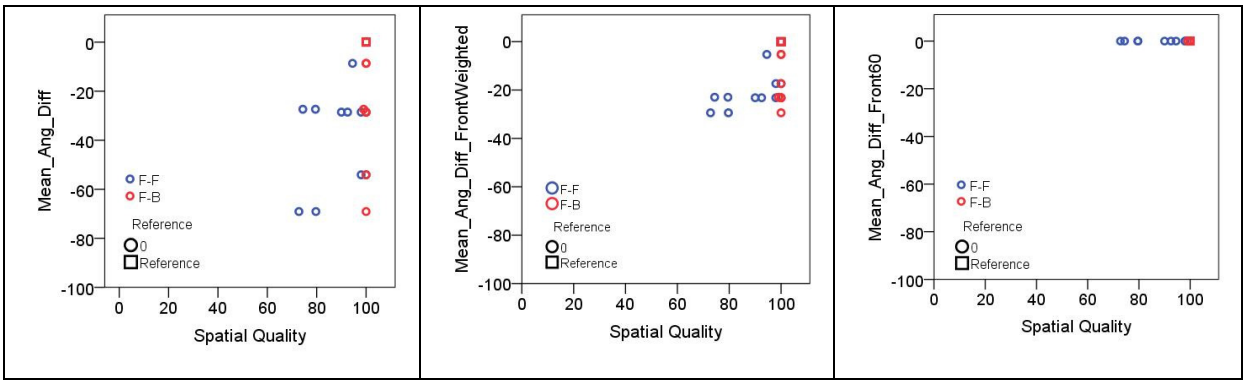


Fig 8.3 Comparison of the performance of Mean_Ang_Diff (left), Mean_Ang_Diff_FrontWeighted (centre) and Mean_Ang_Diff_Front60 (right).

8.2.2.3 Other metrics

'Hull' (named after the shape of the hull of a ship) is another metric created by Dewhirst [2008][discussed in Jackson *et al*, 2008] and could be considered as a measure of scene width. Measured from the listening position this metric uses the binaural signal from the directional localisation model to calculate the angular position in 360° for each of the 36 noise bursts of probe signal 1 after it has been processed by a SAP. The angles are then plotted on the circumference of a unit circle and from this the smallest polygon containing all these points (the convex hull) is determined. The final value of the metric is the area inside the convex hull. 'Hull' showed a negative correlation (r) of -0.56 with spatial quality

One metric was inspired by Karhunen-Lòeve Transform (KLT) analysis (for a detailed explanation see Jiao [2008]). KLT, an extension of principal component analysis (PCA), is a linear transform which can be used to statistically analyse the co-variance between audio channels in a multichannel recording. This is achieved by transposing the audio channels into eigen-channels each containing co-varying audio. The eigen-channels are ordered hierarchically; the first being the most statistically important and containing the largest portion of co-varying audio. The statistical contribution each makes to the original audio is indicated by its co-variance value, for example if all audio channels of a 5-channel recording are correlated this will be transposed to a single eigen-channel with a co-variance value of 1, alternatively if the channels are completely uncorrelated it will be transposed to five eigen-channels with a co-variance value of 0. In broadcast applications these eigen-channels are transmitted with several coefficients so that the receiver can then rebuild the audio accurately. The metric 'CardKLT' measures, in percent, the co-variance value of the first eigen-channel of a KLT decomposition of the signals from four coincident orthogonal cardioid capsules (facing 0°, -90°, 90° and 180°) at the listening position. This is calculated using probe signal 2. 'CardKLT' was originally employed during a previous study where it was used to predict perceived envelopment [Conetta, 2007] by measuring the correlation between the front, rear, left and right of the reproduced soundfield (as discussed in Chapter 3). A similar metric was also used successfully by George [2009] in the prediction of envelopment, using a method that directly analysed the loudspeaker

signals. The 'CardKLT' method is an adaptation of this principle to a system-independent metric. Its implementation in the QESTRAL model is described in Jackson *et al* [2008]. 'CardKLT' had a correlation (r) of 0.6 with the spatial quality scores.

The use of entropy was originally proposed by Jackson and Dewhirst and was also employed in a previous study conducted by this author [Conetta, 2007], discussed in section 4.1.3, where it contributed to a regression model predicting the perceived envelopment arising from speech signals. In that study it was shown that perceived envelopment was influenced by the density of the reproduced soundfield. The entropy was calculated from the left ear of a binaural signal using probe signal 2, as described in Jackson *et al* [2008]. However it was shown [Dewhirst *et al*, 2008] in a preliminary calibration of the QESTRAL model that the value of measured entropy was altered by filtering of the signal, created by the pinna and the shadowing of the head. Hence measuring entropy from the left signal only would not create a consistent measurement between the left and right sides of the soundfield. Therefore to account for this problem an improvement was made to the metric and a mean value of entropy was calculated from both left and right binaural signals ('Mean_Entropy'). For comparison entropy calculated from only the left ear signal had a correlation (r) of -0.38, whereas 'Mean_Entropy' had a correlation (r) of -0.58.

'TotEnergy' was also employed in this author's envelopment prediction model. This was because the perception of envelopment was shown to be altered when the loudness of the reproduced soundfield was changed [Conetta, 2007]. This metric is the calculated root mean square (RMS) sound pressure at the listening position using probe signal 2, captured using a simulated omni-directional microphone. The implementation of 'TotEnergy' in the QESTRAL model is described in Jackson *et al* [2008]. It had a negative correlation (r) of -0.27 to the subjective spatial quality scores. A second level difference metric was created which using the directional localisation model calculates and averages the mean RMS sound pressure difference, between the SAP and the reference, of each noise burst in probe signal 1 from the binaural signal of the virtual dummy head at the listening position. The implementation of this metric in the QESTRAL model is described in Jackson *et al* [2008]. 'Mean_RMS_Diff' had a correlation (r) of 0.55 to the spatial quality subjective scores.

As discussed in pilot study 4, many of the SAPs evaluated affected the perceived timbral quality of the programme items as well as the spatial quality. 'Mean_SpecRollOff' (or mean spectral roll-off)' was included to measure the changes to timbral quality. Similar metrics were used successfully by George [2009] where they were found to be useful for measuring degradations to frontal spatial fidelity (FSF) and surround spatial fidelity (SSF) created by bandwidth limitation filters. The metric was calculated as the mean magnitude of the fast Fourier transform (FFT) of both left and right binaural signals (from a simulated dummy head at the listening position with 0° head orientation) using probe signal 2. 'Mean_SpecRollOff' had a negative correlation (r) of -0.2 with the subjective spatial quality scores.

8.3 Summary of objective metrics

Table 8.3 summarises the 14 metrics used in the calibration of the QESTRAL model for the objective evaluation of spatial quality.

	Metric	Probe signal	Description	R
1	IACC0	1	The mean IACC value calculated across 22 frequency bands (150Hz-10kHz) calculated from a 0° head rotation.	0.64
2	IACC90	1	The mean IACC value calculated across 22 frequency bands (150Hz-10kHz) calculated from a 90° head rotation.	0.51
3	IACC0*IACC90	1	The product of IACC0 and IACC90.	0.62
4	IACC0_9band	1	The mean IACC 0 value calculated from 9 frequency bands (570Hz-2160Hz).	0.71
5	IACC90_9band	1	The mean IACC 90 value calculated from 9 frequency bands (570Hz-2160Hz).	0.53
6	IACC0*IACC90_9band	1	The product of IACC0_9Band and IACC90_9Band.	0.66
7	Mean_Ang_FrontWeighted	2	The mean absolute change to localisation, compared with the reference localisation for the 36 noise bursts, with a linear weighting of decreasing importance from 0° applied to each angle.	0.67
8	Mean_Ang_Front60	2	The mean absolute change to localisation, compared to reference localisation for 7 noise bursts between 0-30° and 330-350°.	0.73
9	Hull	1	The convex area of the localised 36 noise burst plotted on a unit circle	-0.56
10	CardKLT	1	The contribution in percent of the first eigenvector from a Karhunen-Loeve Transform (KLT) decomposition of four cardioid microphones placed at the listening position and facing in the following directions: 0˚, 90˚, 180˚ and 270˚.	0.60
11	Mean_Entropy	1	The mean Shannon entropy value measured from both binaural signals.	-0.58
12	TotEnergy	1	RMS of pressure value measured by a pressure microphone.	-0.27
13	Mean_RMS_diff	2	The mean absolute change to RMS compared with the reference RMS for the 36 noise bursts.	0.55
14	Mean_SpecRollOff	1	The mean magnitude of the FFT from both binaural signals.	-0.20

Table 8.3 Metrics employed for the calibration of the QESTRAL model.

8.4 Calibrating the QESTRAL model for the prediction of spatial quality

This section describes the calibration of the QESTRAL model using partial least squares (PLS) regression. As discussed in section 3.2 this method of regression analysis was chosen because it is adept at calibrating models using a large selection of metrics [Abdi, 2007] and gives the investigator freedom to experiment with different metric combinations.

A number of target specifications for the performance of the QESTRAL model were discussed in chapter 3, and are summarised here in table 8.4. The target value of RMS Error was calculated from the average intra-listener error in listening tests 1 and 2 (see Appendix I). It was also desirable to calibrate the QESTRAL model so that it is generalisable. Therefore to help the model

generalise to a wider selection of SAPs it will be calibrated using the minimum number of metrics and principle components (PCs) required to meet the target specifications. The generalisability will be checked statistically using a number of statistical tests recommended by Field [2005]. The calibration of the QESTRAL model will be terminated once the target specifications are met (NB. All metric measurements were standardised using the inverse of the standard deviation before being entered into The Unscrambler because they used different units of measurement).

Criteria	Target specification
Correlation (r)	≥ 0.86
Root Mean Square Error (RMSE) (%)	≈ 10%
Variance Inflation Factor (VIF)	Mean VIF ≈ 1

Table 8.4 QESTRAL model target specifications.

8.4.1 Calibration method

The aforementioned 14 metrics were entered as independent variables into The Unscrambler simultaneously. For the initial calculation of the model 14 PCs (i.e. 1 PC per metric) were employed. To interpret this calculation 4 graphs were used (Fig 8.4 and 8.5). Figure 8.4 shows the explained variance for calibration and cross-validation against the number of PCs, and shows how much variance in the dependent variable (spatial quality) is explained by the independent variables, as the number of PCs used in the calculation increased (as the model becomes more sophisticated). It can be seen that with all 14 metrics (and PCs) it is possible to explain approximately 81% of the total variance in the subjective scores which is equivalent to a correlation (r) of approximately 0.9. Unfortunately using 14 metrics in the model will not make it very practical to use and potentially not generalisable. However the plots show that it is still possible to achieve a total variance of approximately 74% (equivalent to 0.86 R) in calibration and cross-validation using just 2 PCs.

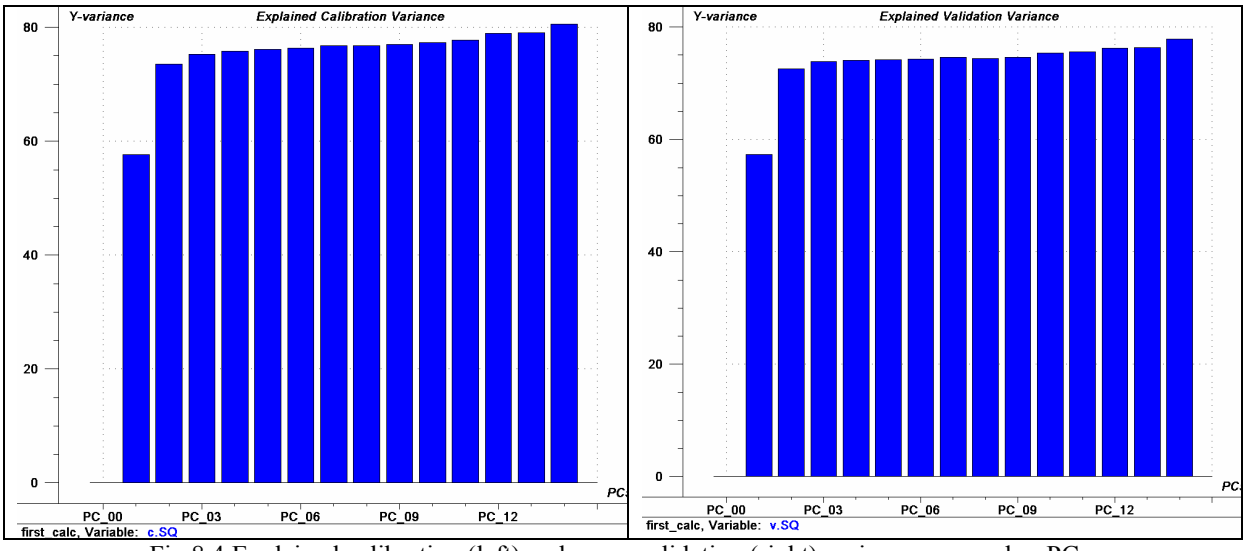


Fig 8.4 Explained calibration (left) and cross-validation (right) variance vs. number PCs.

Figure 8.5 shows the RMSE (%) for calibration and cross-validation against the number of PCs, and reveals how the RMSE (%) reduced as the number of PCs used in the model increased. Although it is not possible to achieve the desired error even with 14 PCs (Root Mean Square Error in Calibration (RMSEC) = 10.66%, Root Mean Square Error in Prediction (RMSEP) = 11.5%), figures 8.4 and 8.5 indicate that the model can be simplified further by reducing the number of PCs used in the calibration, showing that it is possible to achieve a similar value of RMSE (RMSEC = 12.5%, RMSEP = 12.8%) again using just 2 PCs.

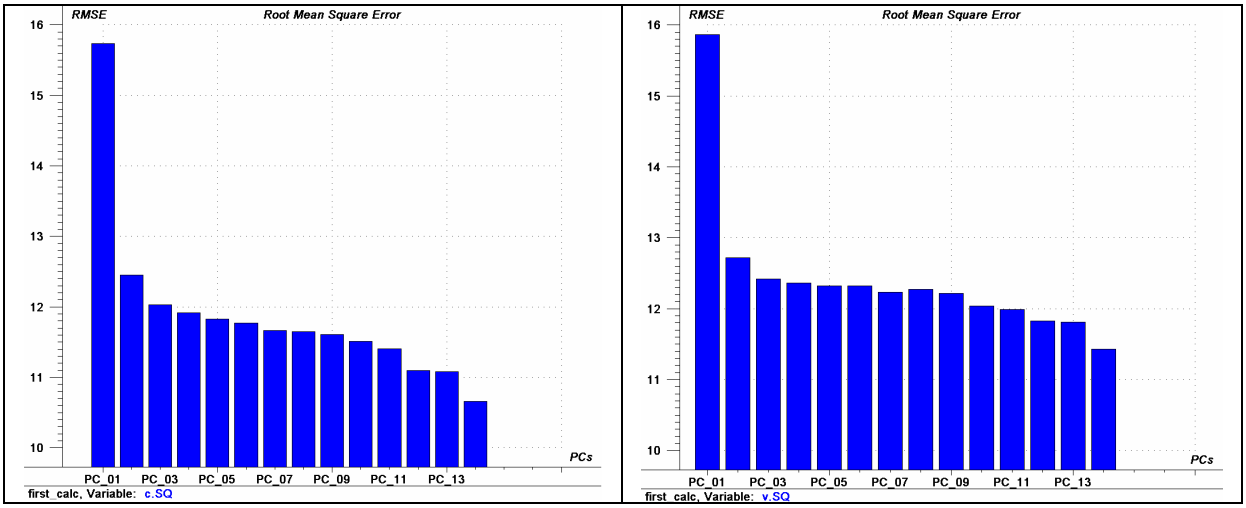


Fig 8.5 RMSE (%) in calibration (left) and validation (right) variance vs. number PCs.

Observing the scatter-plot (fig 8.6) of the subjective scores (measured) vs. predicted results shows the distribution of the subjective scores along the target line (y = x).

A limiting effect is observed at the top of the scale, where the highest quality SAPs (those subjectively scored at 100 or close) are not predicted any higher than ~90%. Therefore if this effect isn't removed with further iterations or recalculations of the model it might be necessary to apply a post-correction transformation to the whole model.

The observations above indicate that using all 14 metrics (and PCs) it is not possible to meet the target specifications, however it is possible to simplify the model to 2 PCs and still achieve a performance close to the target specifications. Based upon this the model was recalculated using 2 PCs. This re-calculation was the first of a series of iterations; The aims of which were to simplify the model by reducing the number of metrics used by the model while still achieving the desired target specifications. The removal of metrics during this process was determined primarily by analysing the weighted coefficient beta values and VIF values for each of them. The entire model iteration process is summarised in table 8.6 and described in detail in the sections which follow.

	No. of Metrics used in calc	PCs	Calibration (R)	RMSEC %	Observation	Action
Initial calculation	14	14	0.90	10.66	The model was over complicated. A model of similar acceptable performance can be achieved using 2 PCs.	Recalculate the model using 2 PCs.
Iteration 1	14	2	0.86	12.45	IACC90_9band, Hull and TotEnergy were found to be statistically insignificant.	Recalculate the model with IACC90_9band, Hull and TotEnergy removed.
Iteration 2	11	2	0.86	12.45	IACC90 was found to be statistically insignificant.	Recalculate the model with IACC90 removed.
Iteration 3	10	2	0.86	12.48	VIF for IACC0*IACC90 and IACC0*IACC90_9band was very high and importance (BW) very low.	Recalculate the model with these metrics removed.
Iteration 4	8	2	0.86	12.33	Model shows same performance but was simpler. VIF between IACC0_9band and IACC0 was high. IACC0 had lowest importance of the two. They were also very correlated.	Recalculate the model with IACC0 removed.
Iteration 5	7	2	0.86	12.32	IACC0_9band and CardKLT were highly correlated and also exhibit a VIF higher than desired. CardKLT had lowest importance.	Recalculate the model with CardKLT removed.
Iteration 6	6	2	0.86	12.16	The model was improved and simpler. Mean_Ang_Diff_FW and Mean_Ang_Diff_60 were both important metrics. Mean_Ang_Diff_FW had a high correlation with Mean_Ang_Diff_60 and IACC0_9band, and also a VIF higher than desired.	Recalculate the model with Mean_Ang_Diff_FW removed.
Iteration 7	5	2	0.87	12.12	The model was improved and simpler. There was a high correlation between Mean_Entropy and IACC0_9band. VIF values were acceptable. Mean_Entropy had the lowest importance of these.	To simplify the model further, recalculate the model with Mean_Entropy removed.
Iteration 8	4	2	0.86	12.39	The model was simpler but the performance is reduced.	Return to iteration 7 and terminate calibration.

Table 8.6 Overview of the QESTRAL model calibration process.

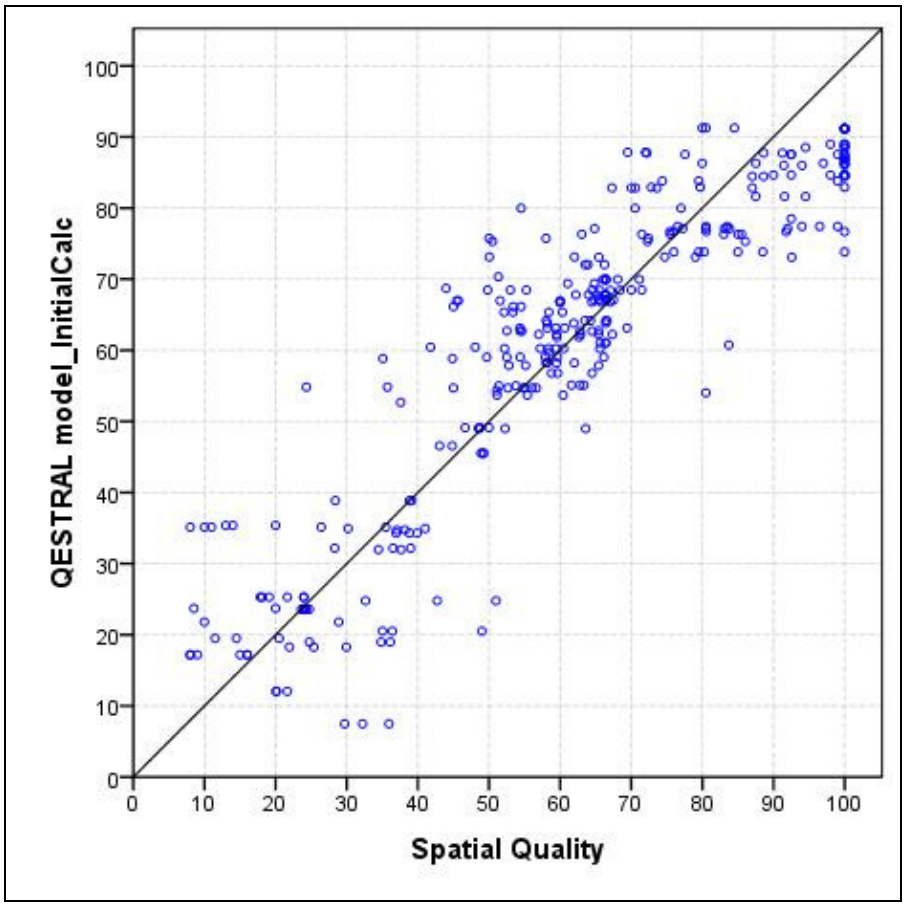


Fig 8.6 Initial calculation; Subjective scores (Spatial Quality) vs. Predicted scores (QESTRALmodel_InitialCalc).

8.4.1.1 Outcome of calibration iteration 1

After iteration 1 three metrics, 'IACC90_9band', 'Hull' and 'TotEnergy', were found to be statistically insignificant, as (highlighted in blue) in table 8.7. The confidence intervals of their weighted beta coefficient values crossed zero. The polarity of the weighted beta coefficient value represents each metric's relationship to the dependent (spatial quality) and hence if the confidence intervals cross zero it suggests that this relationship is uncertain. These metrics also had low statistical importance in the model so they were removed and the model was recalculated.

Metrics	BW
IACC0	0.067
IACC0_9band	0.114
IACC90	-0.0296
IACC90_9band	-0.01833
Mean_Entropy	-0.118
Mean_SpecRollOff	-0.173
CardKLT	0.03185
TotEnergy	-0.01901
Hull	-0.03295
Mean_Ang_Diff_FrontWeighted	0.199
Mean_Ang_Diff_Front60	0.284
Mean_RMS_Diff	0.176
IACC0*IACC90	0.02964
IACC0*IACC90_9band	0.06215

Table 8.7 Weighted beta coefficient values (BW) of the metrics after iteration 1.

8.4.1.2 Outcome of calibration iteration 2

After recalculation the performance of the model was unchanged but the metric 'IACC90' (highlighted in blue in table 8.8) was found to be statistically insignificant and had the lowest importance so it was removed and the model recalculated.

Metrics	BW
IACC0	0.06854
IACC0_9band	0.113
IACC90	-0.02015
Mean_Entropy	-0.119
Mean_SpecRollOff	-0.174
CardKLT	0.03825
Mean_Ang_Diff_FrontWeighted	0.201
Mean_Ang_Diff_Front60	0.279
Mean_RMS_Diff	0.175
IACC0*IACC90	0.03429
IACC0*IACC90_9band	0.06616

Table 8.8 Weighted beta coefficient values (BW) of the metrics after iteration 2.

8.4.1.3 Outcome of calibration iteration 3

All of the metrics were found to make a significant contribution to the model after iteration 3 however there was still a large number of metrics so to reduce them and simplify the model the methods of analysis discussed above were employed. The VIF and weighted beta coefficient values for each metric were examined. The VIF values were very high for the metrics 'IACC0*IACC90' and 'IACC0*IACC90_9band' (Table 8.9); also the weighted beta coefficients for these metrics (highlighted in blue in table 8.10) indicated that they had low importance in the model. Therefore they were removed and the model was recalculated.

Metrics	VIF
IACC0	75.521
IACC0_9band	86.486
Mean_Entropy	2.317
Mean_SpecRollOff	1.081
CardKLT	10.659
Mean_Ang_Diff_FrontWeighted	7.991
Mean_Ang_Diff_Front60	6.418
Mean_RMS_Diff	1.646
IACC0*IACC90	167.924
IACC0*IACC90_9band	156.052

Table 8.9 VIF values after iteration 3.

Metrics	BW
IACC0	0.06604
IACC0_9band	0.111
Mean_Entropy	-0.118
Mean_SpecRollOff	-0.176
Mean_Ang_Diff_FrontWeighted	0.196
Mean_Ang_Diff_Front60	0.274
Mean_RMS_Diff	0.170
IACC0*IACC90	0.03549
IACC0*IACC90_9band	0.06635

Table 8.10 Weighted beta coefficient values (BW) of the metrics after iteration 3.

8.4.1.4 Outcome of calibration iteration 4

After iteration 4 the performance of the model remained unchanged however the model was slightly simpler. Therefore it was decided to continue with the approach and try to simplify the model further. The VIF values for 'IACC0' and 'IACC0_9band' were very high (Table 8.11), because these metrics

perform very similar roles in the model. 'IACC0' (Table 8.12 highlighted in blue) had the lowest importance so this metric was removed and the model recalculated.

Metrics	VIF
IACC0	35.289
IACC0_9band	26.572
Mean_Entropy	2.296
Mean_SpecRollOff	1.054
CardKLT	5.997
Mean_Ang_Diff_FrontWeighted	7.662
Mean_Ang_Diff_Front60	6.200
Mean_RMS_Diff	1.641

Table 8.11 VIF values after iteration 4.

Metrics	BW
IACC0	0.102
IACC0_9band	0.150
Mean_Entropy	-0.137
Mean_SpecRollOff	-0.195
CardKLT	0.06546
Mean_Ang_Diff_FrontWeighted	0.198
Mean_Ang_Diff_Front60	0.265
Mean_RMS_Diff	0.163

Table 8.12 Weighted beta coefficient values (BW) of the metrics after iteration 4.

8.4.1.5 Outcome of calibration iteration 5

Removing 'IACC0' and recalculating the model did not lower its performance, however it did reduce the VIF of 'IACC0_9Band'. The VIF values for the metrics had reduced substantially although they were not as low as desired. The metric 'CardKLT' had the lowest importance in the model (Table 8.13 highlighted in blue). It also exhibited a relatively high VIF (Table 8.14) and was closely correlated to 'IACC0_9band' and 'Mean_Entropy' (Table 8.15). 'CardKLT' was removed and the model was recalculated.

Metrics	BW
IACC0_9band	0.204
Mean_Entropy	-0.162
Mean_SpecRollOff	-0.213
CardKLT	0.101
Mean_Ang_Diff_FrontWeighted	0.201
Mean_Ang_Diff_Front60	0.249
Mean_RMS_Diff	0.157

Table 8.13 Weighted beta coefficient values (BW) of the metrics after iteration 5.

Metrics	VIF
IACC0_9band	4.957
Mean_Entropy	2.253
Mean_SpecRollOff	1.053
CardKLT	5.297
Mean_Ang_Diff_FrontWeighted	6.069
Mean_Ang_Diff_Front60	4.056
Mean_RMS_Diff	1.640

Table 8.14 VIF values after iteration 5.

Correlation (r)	CardKLT
IACC0_9band	0.872
Mean_Entropy	-0.680

Table 8.15 Correlation (r) of CardKLT with IACC0_9band and Mean_Entropy.

8.4.1.6 Outcome of calibration iteration 6

After iteration 6 the model was slightly simiplified, but not at the expense of performance. 'Mean_Ang_Diff_FrontWeighted' and 'Mean_Ang_Diff_Front60' had the highest VIF values (Table 8.16). These metrics were also highly correlated (Table 8.17). 'Mean_Ang_Diff_FrontWeighted' also exhibited a high correlation with 'IACC0_9band' and had a low weighted beta coefficient value (Table 8.18 highlighted in blue), so it was removed and the model recalculated.

Metrics	VIF
IACC0_9band	3.113
Mean_Entropy	1.936
Mean_SpecRollOff	1.037
Mean_Ang_Diff_FrontWeighted	5.403
Mean_Ang_Diff_Front60	3.614
Mean_RMS_Diff	1.640

Table 8.16 VIF values after iteration 6.

Correlation (r)	Mean_Ang_Diff_FrontWeighted
IACC0_9band	0.637
Mean_Ang_Diff_Front60	0.81

Table 8.17 Correlation (r) of Mean_Ang_Diff_FrontWeighted with IACC0_9band and Mean_Ang_Diff_Front60.

Metrics	BW
IACC0_9band	0.276
Mean_Entropy	-0.215
Mean_SpecRollOff	-0.203
Mean_Ang_Diff_FrontWeighted	0.212
Mean_Ang_Diff_Front60	0.224
Mean_RMS_Diff	0.151

Table 8.18 Weighted beta coefficient values (BW) of the metrics after iteration 6.

8.4.1.7 Outcome of calibration iteration 7

After iteration 7 the model performance improved. The VIF values were also more acceptable suggesting that the model exhibited a tolerable level of multi-colinearity (Table 8.19). However Mean_Entropy and IACC0_9band were highly correlated (Table 8.20). As Mean_Entropy had the lowest importance (Table 8.21 highlighted) it was removed and the model recalculated.

Metrics	VIF
IACC0_9band	2.039
Mean_Entropy	1.826
Mean_SpecRollOff	1.032
Mean_Ang_Diff_Front60	1.499
Mean_RMS_Diff	1.549

Table 8.19 VIF values after iteration 7.

Correlation (r)	Mean_Entropy
IACC0_9band	-0.662

Table 8.20 Correlation (r) of Mean_Entropy with IACC0_9band.

Metrics	BW
IACC0_9band	0.336
Mean_Entropy	-0.215
Mean_SpecRollOff	-0.211
Mean_Ang_Diff_Front60	0.339
Mean_RMS_Diff	0.213

Table 8.21 Weighted beta coefficient values (BW) of the metrics after iteration 7.

8.4.1.8 Outcome of calibration iteration 8

After iteration 8 the performance of the model worsened (r = 0.86, RMSEC = 12.39%). Although the recalculated model was simpler and its performance still within the target specifications, a comparison of the correlation of iteration 7 with 6 and 8 within individual SAP groups suggested that iteration 7 had optimal performance (Table 8.22). Hence it was decided to return to iteration 7 and terminate the calibration.

	Process type	Iteration 6	Iteration 7	Iteration 8
1	Down-mixing from 5 CH	0.84	0.82	0.84
2	Multichannel audio coding	0.81	0.81	0.77
3	Altered loudspeaker locations	0.80	0.84	0.83
4	Channel rearrangements	0.57	0.60	0.61
5	Inter-channel level miss-alignment	0.94	0.93	0.88
6	Inter-channel out-of-phase errors	0.93	0.94	0.94
7	Channel removal	0.64	0.66	0.69
8	Spectral filtering	0.81	0.82	0.80
9	Inter-channel crosstalk	0.71	0.64	0.65
10	Virtual surround algorithms	-0.92	-0.92	-0.92
11	Combinations of 1-10	0.82	0.82	0.81
12	Scale anchors	0.95	0.96	0.92

Table 8.22 Comparison of the correlation (r) to individual SAPs of iterations 6, 7 and 8.

8.4.2 Calibrated QESTRAL model

The performance of the calibrated QESTRAL model (after iteration 7) is similar to the initial model which had 14 metrics and used 14 PCs. However the calibrated model uses just 5 metrics and 2 PCs, meets the target specifications for correlation and has a suitable RMSE (%) and VIF value for calibration. The performance of this model in cross-validation was also close to the target specifications (Table 8.23). The generalisability of the model is also tested using a series of statistical tests suggested by Field [2005]. The results of these tests are presented in Appendix J; they show that the QESTRAL model passes these tests indicating that it is generalisable (NB. To run these tests the model had to recalculated in SPSS using PCR regression with same five objective metrics).

	Correlation (r)	RMSE (%)
Calibration	0.87	12.12
Cross-validation	0.86	12.34

Table 8.23 Calibration and cross-validation correlation (r) and RMSE (%) of the calibrated QESTRAL model.

Figure 8.7 shows the distribution of the subjective scores along the target line (y = x) and illustrates the ability of the calibrated QESTRAL model to predict the effect of a wide range of different SAPs. However the limiting effect noticed after the first calculation still remains at the top of the scale. This causes the prediction of highest quality stimuli to be limited to ~90% (eg. hidden reference recordings are predicted at 91% rather than 100%). It is desirable to remove this limiting effect so that the model performs closer to the subjective response of the listeners (i.e. so that the SAPs perceived at the top of the scale are predicted at the top of the scale).

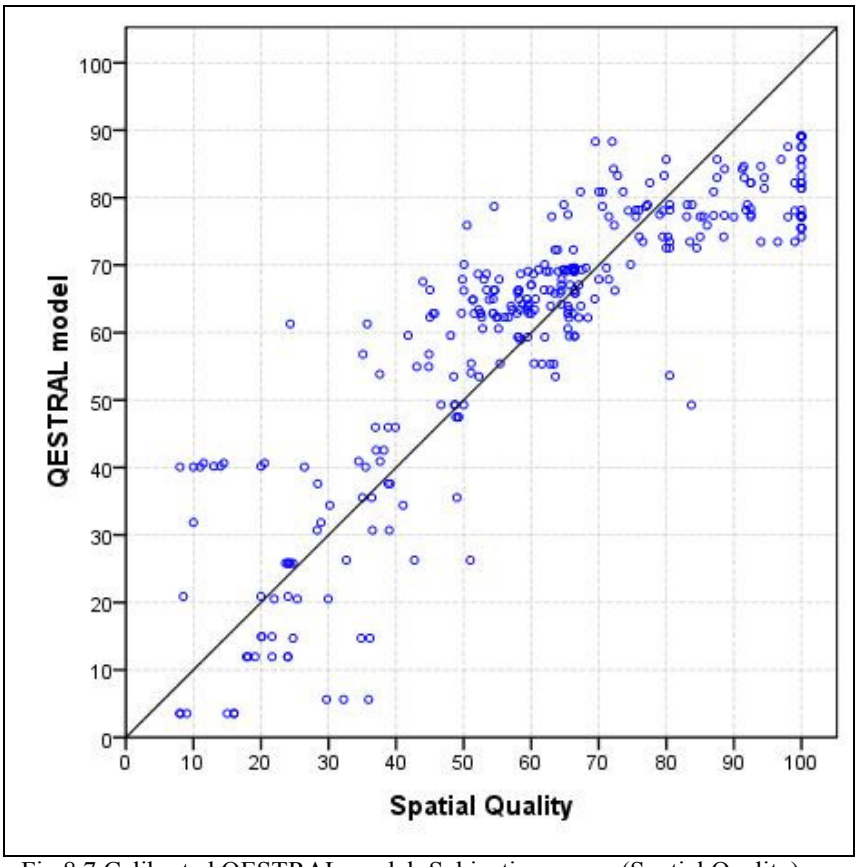


Fig 8.7 Calibrated QESTRAL model; Subjective scores (Spatial Quality) vs. predicted scores (QESTRAL model).

The regression equation for the calibrated QESTRAL model is shown in equation 8.1.

$$\begin{aligned} QESTRAL &= 61.887(IACC0\_9Band) \\ &+ 0.352(Mean\_Ang\_Front60) \\ &- 23.017(Mean\_Entropy) \\ &- 0.002153(Mean\_SpecRollOff) \\ &+ 695.407(Mean\_RMS\_diff) \\ &+ 89.069916 \end{aligned} \qquad \text{(eq. 8.1)}$$

Observing the weighted beta coefficient values of each metric (Table 8.21) it is shown that 'IACC0_9band' and 'Mean_Ang_Diff_Front60' are the most statistically important metrics in the model. It is possible to identify the role of each PC in the model using the correlation loading plot (Fig 8.8).

Figure 8.8 shows that the metrics 'IACC0_9band', 'Mean_Ang_Diff_Front60', 'Mean_Entropy' and 'Mean_RMS_Diff' are distributed along lie along PC1 while Mean_Spec_RollOff clearly lies along PC2. The distribution of the metrics suggests that PC1 (x axis) represents spatial quality and PC2 (y- axis) timbral quality. These are the two domains of audio quality as discussed in section 2.1. The y-explained variance shows that PC1 explains 73% of the dependent

variable and PC2 explains 2%. If the meaning of the PCs is interpreted correctly then it indicates that the prediction of spatial quality predominantly relies upon the measurement of changes to the spatial characteristics created by the SAPs. However a small contribution is made from measuring the changes to the timbral characteristics. This supports the hypothesis that changes to timbral quality might have a small influence on the perceived spatial quality, as discussed in pilot study 4, and that the domains are interlinked as discussed in section 2.1.1.

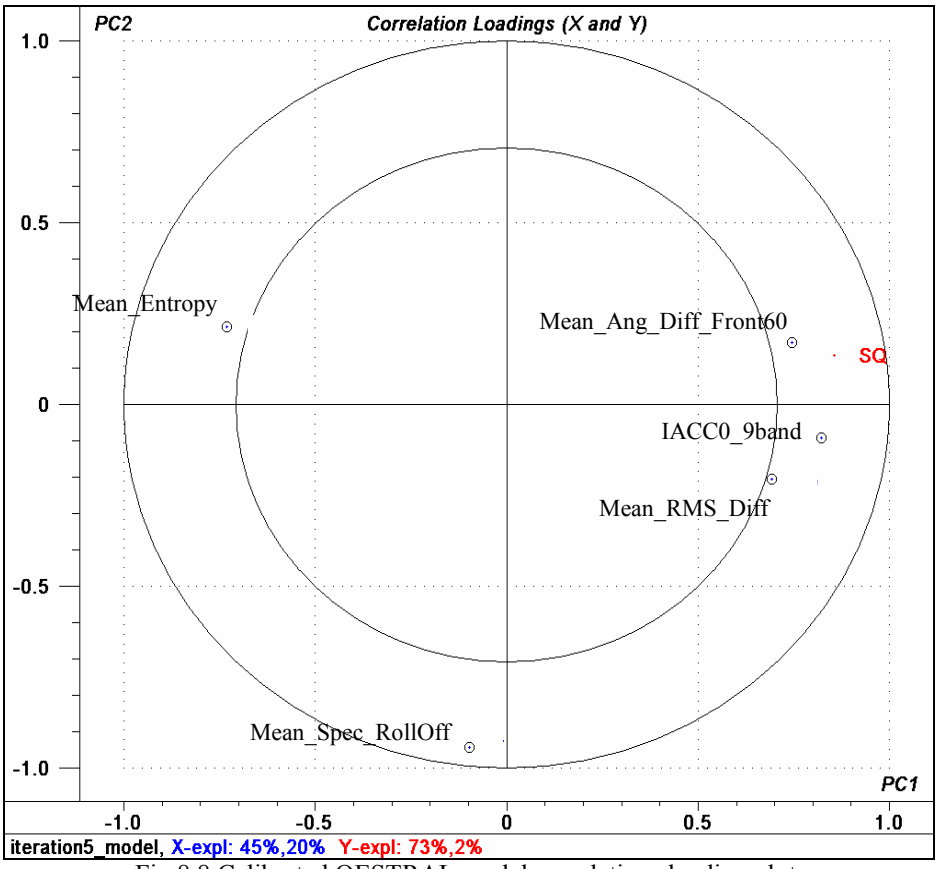


Fig 8.8 Calibrated QESTRAL model correlations loading plot.

8.5 Corrected QESTRAL model

The correction procedure followed two stages, the first stage was to correct for the limiting effect to straighten the fit of the model. This was done by determining the trend of the current fit by calculating the equation of best-fit. This revealed that an exponential correction was required to improve the performance of the model. As shown by figure 8.9 this removes the compression effect. Unfortunately the scores for the high anchor recording which should have been predicted at 100% are slightly over predicted at 100.069%. Therefore so that the model represented the paradigm employed for collecting the subjective data correctly a simple linear adjustment was required: 0.069 (2sf) was removed from each score. These corrections resulted in a statistically significantly different ($p < 0.05$) model with an improved performance, producing a correlation (r) of 0.89 and an RMSEC of 11.06%. Although it

could be argued that the correction potentially limits the QESTRAL models validity and generalisability it is believed that any negative effects are mitigated by the large number and varied range of SAPs used in the calibration (NB. It was not possible to re-run the statistical tests suggested by Field after the model's performance had been corrected). The corrected QESTRAL model is given by equation 8.3

$$QESTRAL_{corrected} = 14.102e^{0.022QESTRAL_{model}} - 0.069 \quad \text{(eq 8.3)}$$

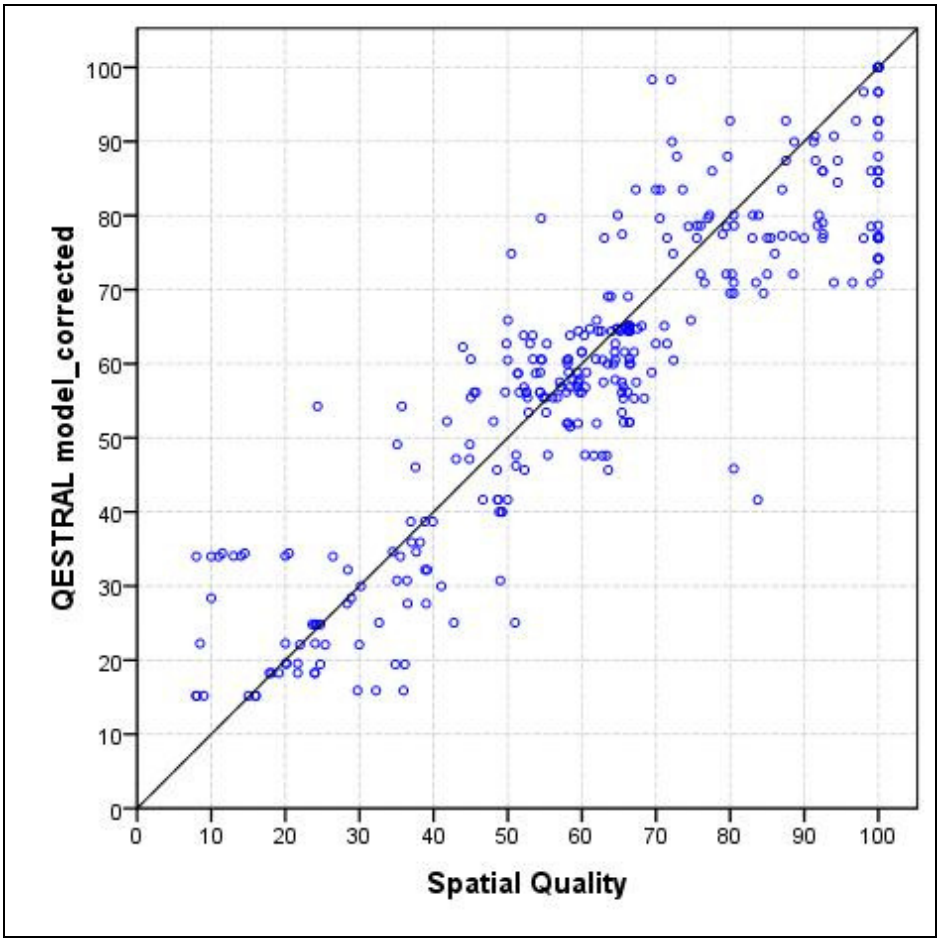


Fig 8.9 QESTRAL model corrected; Subjective scores (Spatial Quality) vs. predicted scores (QESTRAL model_corrected).

8.6 Discussion of the performance of the QESTRAL model after correction

This section evaluates the QESTRAL model's performance by calculating the correlation of the calibrated model to the subjective scores for the twelve SAPs types, six different types of programme items and two different listening positions. Appendix K presents a comparison between the listener scores and QESTRAL model prediction for each stimulus in the calibration data set.

8.6.1 Calibration correlation and RMSE of the QESTRAL model to individual SAPs

Table 8.24 summarises the correlation (R) and RMSE (%) between the calibrated QESTRAL model and the subjective scores for individual SAPs.

Group	Process type	n	R	RMSE (%)
1	Down-mixing from 5 CH	35	0.86	12.68
2	Multichannel audio coding	37	0.86	8.68
3	Altered loudspeaker locations	29	0.85	9.28
4	Channel rearrangements	19	0.63	13.87
5	Inter-channel level miss-alignment	16	0.93	17.50
6	Inter-channel out-of-phase errors	16	0.94	5.25
7	Channel removal	22	0.66	11.57
8	Spectral filtering	13	0.86	13.36
9	Inter-channel crosstalk	11	0.67	15.82
10	Virtual surround algorithms	4	-0.92	23.17
11	Combinations of 1-10	70	0.88	9.83
12	Scale anchors	36	0.99	4.83

Table 8.24 Calibration correlation (r) and RMSE (%) of the QESTRAL model with each SAPs (n = number of samples).

The correlation of the QESTRAL model is acceptable for all SAPs and meets the model target specifications for five of the twelve. The QESTRAL model performs best in the prediction of the scale anchor processes and worst in the prediction of channel rearrangement SAPs. As the test scale was calibrated using the anchors the high correlation (r = 0.99) and low RMSE (4.83%) to the scale anchors indicate that the model is a good representation of the subjective experiments. Figure 8.10 shows the distribution of these scores along the model regression target line. The subjective scores for the anchors vary along the x-axis (except for the hidden reference) because the anchors were scored differently depending upon the programme item they were applied to. However the QESTRAL model only produces one value for each anchor recording, causing a discrepancy between the predicted scores and subjective scores.

The model also has a high correlation (r = 0.88) and low RMSE (9.83%) with group 11. Figure 8.11 shows how the samples for this group are spread quite closely to, and along the length the of, the target line. This is promising as this group contains combinations of all of the other SAPs, it can be seen as a representation of the model's generalisability.

The calibrated QESTRAL model has a negative correlation (r = -0.92) and the highest RSME (23.17%) for virtual surround algorithms. However the number of samples (n) used to calculate these values are very small (less than the number of metrics used in the calibrated model) so the validity of the model's capability to predict this type of SAP is questionable.

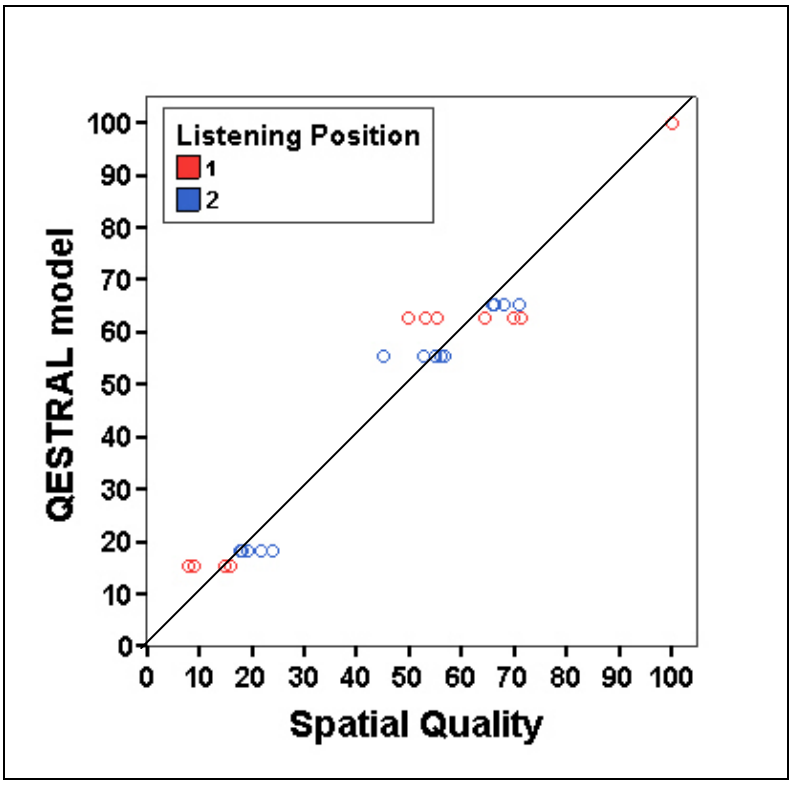


Fig 8.10 Spatial quality (subjective scores) vs. QESTRAL model (predicted scores) for scale anchor SAPs at listening position 1 and 2.

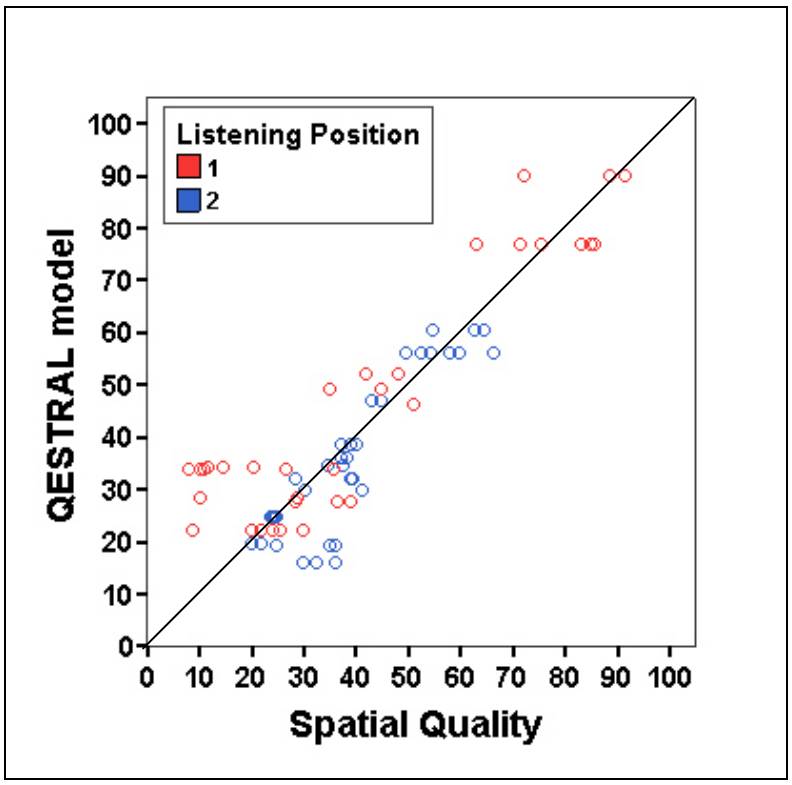


Fig 8.11 Spatial quality (subjective scores) vs. QESTRAL model (predicted scores) for SAP group 11 at listening position 1 and 2.

8.6.2 Calibration correlation and RMSE of the QESTRAL model to individual programme items

In chapters 6 and 7 it was determined that a listener's opinion of the spatial quality of a particular SAP was influenced by the programme item it was applied to and hence it was decided that programme item type should be included as an independent variable in the calibration of the QESTRAL model. It was hypothesised that by this approach the QESTRAL model would be sensitive to the perceptual differences created when a SAP is applied to different programme items. Table 8.25 shows the correlation (r) and RMSE (%) between the calibrated QESTRAL model and the subjective scores collected with each different programme item evaluated in the study. The scatterplots in figure 8.12 illustrate the distribution of the subjective scores along the model regression target line for each programme item. For all six programme items the performance values are within, or close to, the model target specifications. The calibration of the QESTRAL model was most highly correlated to programme item 3, and least correlated to programme items 1 and 2.

No.	Genre Type	Scene Type	Description	n	R	RMSE (%)
1	TV Sport	F-F	Wimbledon. Commentators and applause. Commentators panned mid-way between L, C and R. Audience applause in 360°.	73	0.88	11.05
2	Classical Music	F-B	Music. Wide continuous front stage including localisable instrument groups. Ambient surrounds with reverb from front stage.	69	0.86	13.01
3	Rock/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.	72	0.93	8.81
4	Jazz/Pop Music	F-B	Live music performance. Wide front stage, ambience from room and/or audience in the rear loudspeakers.	33	0.92	10.94
5	Abstract	F-F	Abstract or synthetic scene. Very immersive. Source positioned all around the listener. Some sources are moving.	31	0.92	11.23
6	Film	F-B	Dialogue in C. Ambience, SFX and Music in L, R, Ls, and Rs.	30	0.92	9.10

Table 8.25 Calibration correlation (r) and RMSE (%) of the QESTRAL model for each programme item.

Two types of error were identified in the calibrated QESTRAL model. The errors were caused because the calibrated model is not sensitive to the perceptual differences in spatial quality created when SAPs are applied to different programme items.

The first error occurs when SAPs designed/selected to create different changes to the spatial content of an audio recording are perceived as creating no impairment to the spatial quality (listeners giving them a score of 100%). The calibrated model is not sensitive to this perceptual phenomenon, as it bases its responses on the metric analysis of one set of probe signals. Instead it over estimates the impairment to spatial quality, resulting visually in a stacking of the predicted scores vertically along the y-axis. After this error was investigated it was established that it occurred for the prediction of SAPs which altered the rear channels (e.g. 3.0 downmix, Ls removed) when these were applied to

programme items with an F-B scene type, which only contain background, ambient or reverberant content in the rear channels. This was because these SAPs create an impairment to the programme items that the listeners did not perceive as degrading to the spatial quality of the recording. See also chapters 5 and 6 where the perceptual reasons for this phenomenon were discussed. To illustrate this, a comparison of programme items 2 (F-B scene type) and 3 (F-F scene type) is shown in figure 8.13. SAPs were selected which were perceived as subjectively identical when applied to programme item 2. The scores for programme item 2 are represented by circles while the scores for programme item 3 are represented by triangles. It can be seen that when the SAPs were applied to programme item 2, a mean score of 100, equal to the hidden reference, was given by the listeners. However the QESTRAL model predicts that they each create a different and greater impairment to spatial quality than had been perceived. As can be seen the model prediction is closer to the perceived impairment to spatial quality created when the SAPs are applied to programme item 3.

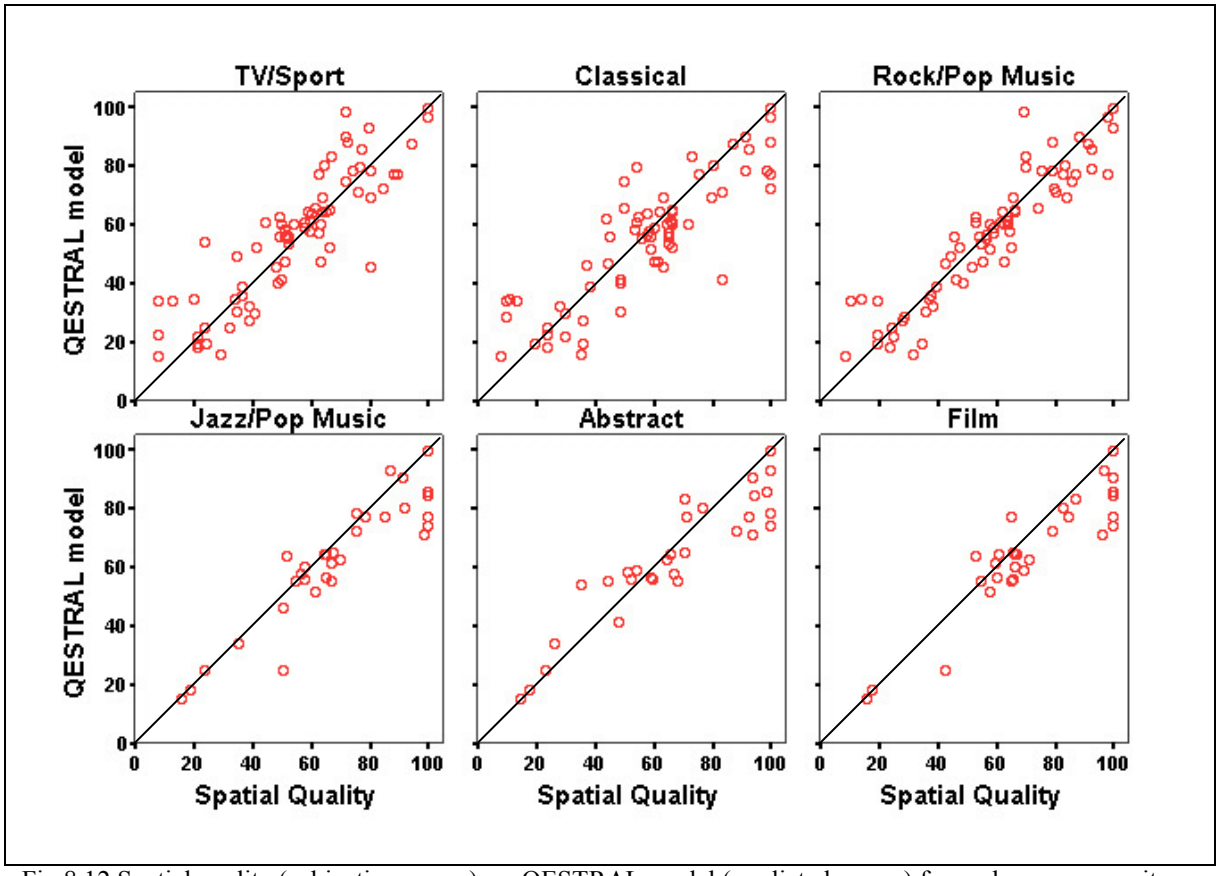


Fig 8.12 Spatial quality (subjective scores) vs. QESTRAL model (predicted scores) for each programme item.

The second error occurs when the same SAP applied to different programme items created perceptually different impairments to spatial quality. Again the calibrated model is not sensitive to this perceptual phenomenon because it bases its responses on the analysis of one set of probe signals. Instead, it under-estimates the subjective scores predicting that the SAP creates an identical impairment to spatial quality when it is applied to the different programme items. Visually this results

in a stacking of the predicted scores horizontally along the x-axis. After this error was investigated it was established that it occurred most significantly between programme items of F-F and F-B scene type. Figure 8.14 illustrates a particular example of this error.

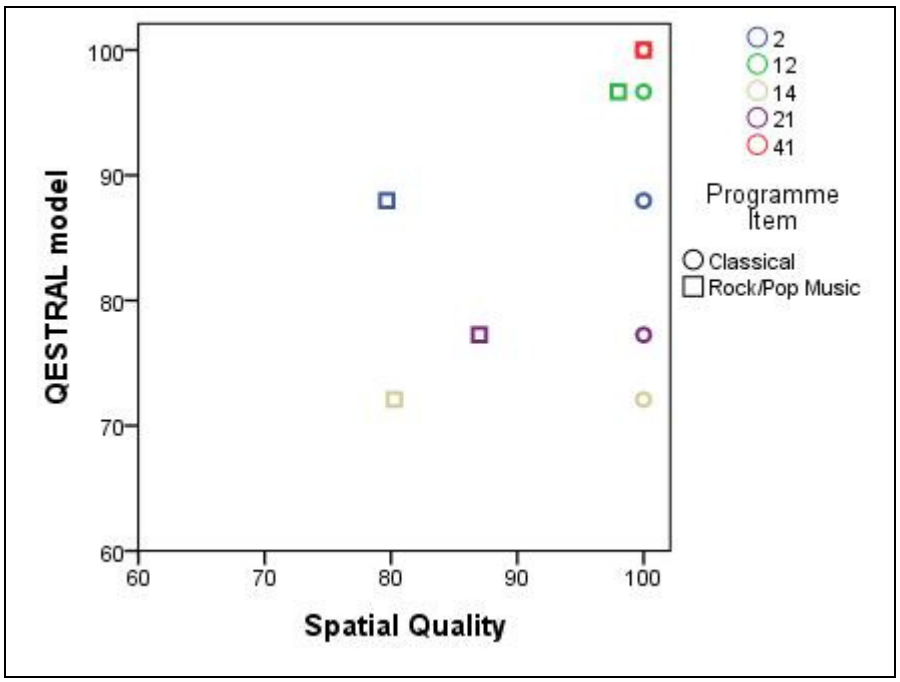


Fig 8.13 Vertical error - comparison of the subjective scores (Spatial Quality) vs. predicted scores (QESTRAL model) for different SAPs applied to programme items 2 (Classical)(F-B) and 3 (Rock/Pop Music)(F-F).

The QESTRAL model is not sensitive to two perceptual phenomena created when SAPs are applied to different programme items types. This creates error in the model contributing to its RMSE (%). The error occurs because the QESTRAL model bases its prediction of spatial quality on the metric analysis of one set of probe signals and produces only one prediction response for each SAP. Analysis of the error suggests that the model is not capable of predicting the perceptual effects observed when SAPs that alter the rear channels of the programme items (e.g. 3.0 downmix, Ls removed) are applied to programme items with an F-B scene type (programme items 2, 5 and 6). This suggests that the model is biased towards the assessment of the effects of SAPs on F-F programme item material. Although this is not supported by the model's performance, as it performs similarly in the prediction of both scene types (see table 8.19), a larger proportion of the dataset used for calibration consisted of scores collected from SAPs applied to the F-F scene type programme material (programme items 1, 3 and 4).

There are two possible ways of removing this insensitivity. The first is to remove programme material as a variable in the model by averaging the subjective scores to single mean value for each SAP. However this would mean that the QESTRAL model would not be capable of predicting the perceived differences created by different programme items and therefore not be as informative to a user. The second is to calibrate the model for different types of programme items, for example the two broad classes of scene types evaluated in this project, F-F and F-B. Although this could yield more

accurate predictions, from a practical viewpoint it would not be ideal to create a number of different calibrations, as a user might find this confusing.

At this point it is reiterated that although these errors are a major contribution to the RMSE (%) in the calibrated QESTRAL model, it performs close to the target specifications in the prediction of each of the independent variables, and the error itself is similar to the average listener error in listening tests 1 and 2.

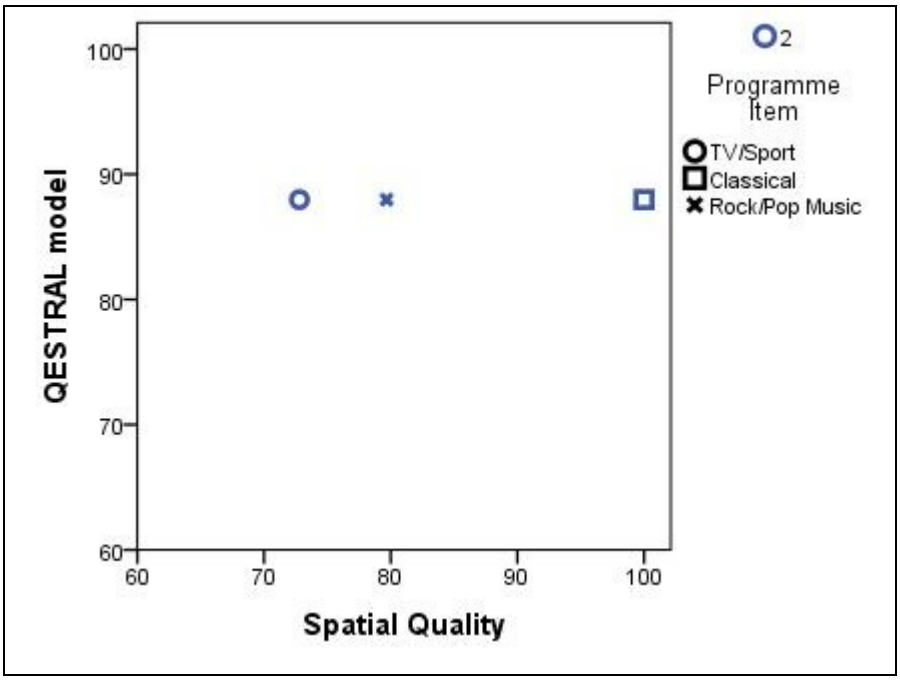


Fig 8.14 Horizontal error - comparison of the subjective scores (Spatial Quality) vs. predicted scores (QESTRAL model) for identical SAPs applied to programme item 1 (TV/Sport)(F-F), 2 (Classical)(F-B) and 3 (Rock/Pop Music)(F-F).

8.6.3 Calibration correlation and RMSE of the QESTRAL model to individual listening positions

As shown in table 8.26 the correlation (r) of the calibrated QESTRAL model with the scores collected at both listening positions exceeds the target specifications for the model. This indicates that the model is very capable of predicting the subjective scores at both listening position 1 and 2.

Listening position	Location	n	Correlation (r)	RMSE (%)
1	Centre	157	0.89	13.44
2	1m to the right of centre	151	0.88	7.86

Table 8.26 Calibration correlation (r) and RMSE (%) of the QESTRAL model for each listening position.

Figure 8.15 illustrates the distribution of the scores for both listening position 1 (in red) and 2 (in blue). The scores at listening position 1 are distributed between 100% and 15% and the scores at listening position 2 are distributed between 70% and 20%.

Listening position is not a source of error in the calibrated model because the QESTRAL model is designed to take measurements at different positions across the listening area. The correlation (r) and RMSE (%) values for each listening position indicate that the model is very capable of predicting the subjective scores at both positions.

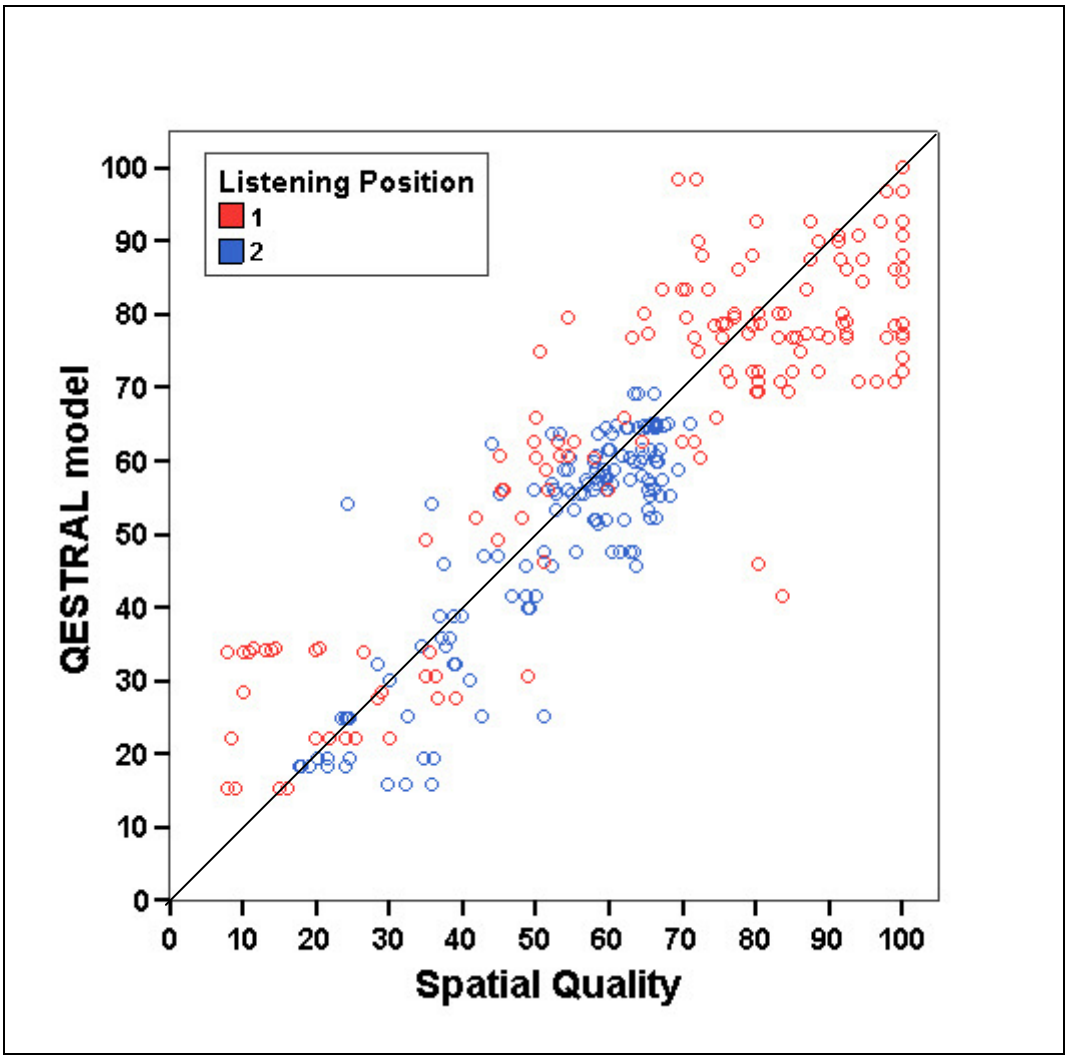


Fig 8.15 Scatterplot of spatial quality (subjective scores) vs. QESTRAL model (predicted scores) for listening position 1 (in red) and 2 (in blue).

8.6.4 Performance after correction: conclusions

The QESTRAL model's performance was investigated after correction for its prediction of the test variables incorporated into the calibration process. This investigation revealed that it performs well over the evaluated SAPs applied to six ecologically valid programme items at both listening position 1 and 2.

Using programme items and listening position dependent subjective scores to calibrate the QESTRAL model allows it to incorporate the perceived differences created by the variables. Closer inspection of the predicted scores revealed that using programme item dependent subjective scores is a source of error in the model. This showed that the QESTRAL model is not sensitive to subjective

differences created when SAPs are perceived as subjectively identical (error along the vertical y-axis) and when identical SAPs are applied to different programme items (error along the horizontal x-axis) (for examples see figures 8.13 and 8.14). Two ways of removing this insensitivity are suggested. The first is to remove programme material as a variable in the model by averaging the subjective scores to single mean value for each SAP. The drawback to this would be that the QESTRAL model would not capable of predicting the perceived differences created by programme items with different scene types. The second is to calibrate the model for different types of programme items, for example the two broad classes of scene types evaluated in this project, F-F and F-B. However from a practical viewpoint it would not be ideal to create a number of different calibrations, as a user might find this confusing. However although these errors are a major contribution to the RMSE (%) of the calibrated QESTRAL model, the model performs close to the target specifications in the prediction of each of the independent variables, and the error itself is similar to the average listener error in listening tests 1 and 2. Hence it is believed that the current method of calibration is the most suitable.

8.7 Summary and Conclusions

This chapter described the calibration and subsequent performance of the QESTRAL model for the automatic evaluation of spatial quality using the data collected in the listening tests discussed in chapter 7. The aims of chapter 8 were to:

i) Establish if the probe signals and objective metrics developed by the QESTRAL project team can be used to build a system that, after calibration against the listening test data from chapter 7, meets the target specifications proposed in section 3.6.

ii) Determine if the calibrated QESTRAL model is generalisable and performs within target specifications for the prediction of spatial quality for each of the independent test variables (SAPs, programme items and listening positions).

Two probe signals based upon pink noise signals were developed by the QESTRAL project team for the calibration of the QESTRAL model (Table 8.1). These were designed to allow the measurement of changes in the foreground and background audio streams.

A number of metrics were developed to predict the spatial characteristics tested during the listening tests. The development of the metrics was informed by the results of a study which determined the extent to which the SAPs evaluated during the listening tests changed a selection of different spatial attributes. An additional metric to measure changes to timbral characteristics was also included. Each metric was assessed for its correlation with the subjective scores (Table 8.3).

The calibrated QESTRAL model performed close to the target specifications, meeting them for correlation and VIF, but not for RMSE (%). It also passed all of statistical tests designed to measure its potential generalisability. Therefore the calibrated model fulfilled both aims of this chapter. However a limiting effect was observed for SAPs at the top of the scale and a method of correcting the model's

performance was required. The correction procedure followed two stages, the first stage was to correct for the limiting effect to straighten the fit of the model. An exponential correction was required to remove this. Unfortunately the scores for the high anchor recording which should equal 100 were slightly over predicted and a simple linear adjustment was required: 0.069 (2sf) was removed from each score. This resulted in a statistically significantly different model with an improved performance, producing a correlation (r) of 0.89 and an RMSEC of 11.06% (Table 8.27). Although it could be argued that the correction potentially limits the QESTRAL models validity and generalisability it is believed that any negative effects are mitigated by the large number and varied range of SAPs used for calibration.

QESTRAL model	
Correlation (r)	0.89
RMSE (%)	11.06%
No. of metrics	5
No. of PCs	2
VIF (max)	2

Table 8.27 QESTRAL model performance results.

The QESTRAL model performance was investigated after correction for its prediction of the test variables incorporated into the calibration process. This investigation revealed that it performs well over the evaluated SAPs applied to six ecologically valid programme items at both listening position 1 and 2. However closer inspection of the predicted scores revealed that using programme item dependent subjective scores was a source of error in the model. This showed that the QESTRAL model was not sensitive to subjective differences created when SAPs are perceived as identical (error along the vertical y-axis) and when identical SAPs are applied to different programme items (error along the horizontal x-axis) (for examples see figures 8.13 and 8.14). These errors contributed to the RMSE (%) of the calibrated QESTRAL model. However, the model's performance is close to the target specifications in the prediction of each of the independent variables, and the error itself is similar to the average listener error in listening tests 1 and 2.

Two ways of removing this insensitivity were suggested: (i) by removing programme item as a variable in the model; and (ii) to calibrate different versions of the model for programme items with different scene types. However either suggestion would limit the practical usage of the QESTRAL model. Hence the current method of calibration is believed to be the most suitable from both a practical and performance standpoint.

Chapter 9 – Summary and conclusions

As a contribution to the development of the QESTRAL model the main aim of this thesis was to establish a method for the prediction of spatial quality. The work presented has shown that using the QESTRAL model architecture, spatial quality can be predicted using a set of objective metrics, each of which relates to a low-level spatial attribute, and probe signals together with a polynomial weighting function derived from regression analysis of data from listening tests which employ SAPs proven to stress those low-level attributes.

This chapter summarises and draws conclusions from the research presented in each chapter. This is followed by a discussion of the limitations of the QESTRAL model alongside suggestions for future work and a discussion of this research project's novel contributions to knowledge. Finally a list of publications contributed to by this research project is presented.

9.1 Chapter summaries and conclusions

This section summarises the contents and findings of each chapter.

9.1.1 Chapter 1 – Introduction

Chapter 1 described the motivation and background of the research described in this thesis, detailing its aims. The development of the QESTRAL model was motivated by the increasing importance of spatial audio and the lack of a perceptually-representative objective measure. The QESTRAL project aimed to provide a model capable of predicting perceived spatial quality. The contributions made by this author to the development of the QESTRAL model were identified as including:

(i) defining spatial quality for this research project,
(ii) defining suitable performance criteria for the QESTRAL model,
(iii) identifying a suitable method for the development of the QESTRAL model,
(iv) identifying a suitable test environment (i.e. reference reproduction system),
(v) identifying appropriate objective metrics for spatial quality,
(vi) designing a listening test method to obtain the required subjective data,
(vii) collating subjective data,
(viii) calibrating the QESTRAL model for the prediction of spatial quality.

9.1.2 Chapter 2 – Sound quality and spatial quality in reproduced sound

To answer aim (i) chapter 2 concentrated on investigating spatial quality. Current objective models for sound quality were also reviewed in order to identify novel areas for investigation and also to help answer aim (ii).

A definition for spatial quality was established for the reproduced sound environment based upon Letowski's idea that spatial quality is a global assessment of lower level spatial attributes. It was defined for this research project as the attribute that describes any and all differences between the reference and impaired items, but only in the spatial characteristics of the recording. Hence spatial quality can be considered as a higher level assessment of the lower level spatial attributes, such as those identified in section 2.2.

Studies conducted by Zielinski *et al* found that the audio processes they investigated degraded both timbral fidelity and spatial fidelity. Letowski suggested that it might be possible for listeners to become confused in situations where a SAP causes a change in the quality across both domains, as their opinion of the spatial quality may be influenced by the perceived timbral quality. It is not possible to completely separate the two domains in the context of this research project. So it is important to establish if changes, created by different SAPs, to the timbral quality of an audio recording (programme item) have an affect on a listener's perception of the spatial quality.

Of the sound quality models reviewed only the recent models created by Choi *et al* and George *et al* could be considered as incorporating spatial quality to some degree. Both of these showed good performance, however they were both calibrated using a limited selection of audio process types (e.g. multichannel audio coding, bandwidth limitation and downmixes). Therefore it was decided that the QESTRAL model would be calibrated to measure a greater range of SAPs.

George *et al* specified performance criteria for the development of his models. The target specifications for the models were for them to achieve a correlation (r) equal to or greater than 0.9 and RMSE of less than 10%. This was based upon the performance of PEAQ and PESQ and achieving an RMSE (%) similar or better than the reported listener error from the listening tests. Similar criteria were considered for the QESTRAL model and discussed further in chapter 3.

9.1.3 Chapter 3 – Methods for the development of the QESTRAL model.

In chapter 3 topics relating to how the QESTRAL model was to be created were discussed in order to answer aims (ii), (iii) and (iv).

The contribution of lower level spatial attributes to sound quality or spatial quality has not been quantified and achieving this would require a substantial amount of time and research, which would not be possible during this research project and so a direct prediction method, as defined by Bech, was selected for the development of the QESTRAL model. In a direct method of model development, subjective data is collected on a global assessment of audio quality and objective metrics

are selected to measure the global and/or lower level attributes that comprise it. A potential risk of using this approach is that the experimenter mistakenly limits the validity of the model by only collecting data on some of the component attributes that contribute to the global attribute (e.g. if the stimuli tested do not exhibit traits of all of the lower level spatial attributes). This meant that the SAPs used to calibrate the QESTRAL model should stress the lower level spatial attributes (see section 2.2).

Partial least squares (PLS) regression was chosen as the best regression analysis method to calculate the QESTRAL model because it is adept at calibrating models using a large selection of independent variables and gives the investigator freedom to experiment with the use of different metrics. Table 9.1 summarises the target specifications for the calibrated QESTRAL model; these were based upon the performance criteria of similar models created by George and Choi *et al*. It was also decided that to facilitate the models generalisability it would be calibrated using the minimum number of principal components required to meet the target specifications. This would be checked using a number of statistical tests suggested by Field.

Criteria	Target specification
Correlation (r)	≥ 0.86
Root Mean Square Error (RMSE) (%)	≈ average intra-listener error
Variance Inflation Factor (VIF)	Mean VIF ≈ 1

Table 9.1 QESTRAL model target specifications.

The considerations for the selection of a suitable reference audio system were the ability of the system to reproduce spatial attributes and its widespread use. After a study of current commercial reproduction systems it was decided that 3/2 stereo was the most suitable system for this research. This system is also capable of replaying mono and 2-channel stereo material and so allows these systems to be investigated simultaneously.

9.1.4 Chapter 4 – Review of objective metrics that could be used in the QESTRAL model

As discussed in chapter 3, in using a direct method of model development objective metrics were required to measure the global and/or lower level attributes that comprise spatial quality. Current objective metrics for the measurement of individual spatial attributes and those used in current spatial sound quality models were reviewed in chapter 4 in order to answer aim (v).

In their study Choisel and Wickelmaier described the correlation of a number of different metrics designed to measure both timbral and spatial characteristics. In particular this identified that metrics based upon the measurement of IACC show good correlation with spatial attributes in the reproduced sound environment such as perceived width, envelopment and spaciousness. This was also supported by the work of other researchers.

A number of models for localisation have been developed and these predominantly rely upon measuring the interaural time difference and interarual level difference.

A number of metrics have been shown to correlate well with perceived changes to envelopment. Soulodre *et al* proposed a metric which combined measurements of the relative level and the angular distribution of late energy. Conetta, Dewhirst and George used metrics based upon measurements of the IACC, KLT, Entropy, ITD, ILD and also included metrics to measure scene or ensemble width and the timbral characteristics. In these models multiplicative metrics were also used to good effect.

Choi *et al* employed metrics to measure both timbral and spatial characteristics to predict degradations of BAQ created by low bit-rate multichannel audio codecs to a selection of 5.1 multichannel recordings. The metrics ILDD, ITDD and IACCD measuring spatial characteristics were shown to have the highest independent correlation to the subjective scores.

George *et al* created models for the prediction of frontal spatial fidelity (FSF) and surround spatial fidelity (SSF) in which he employed a wide selection of metrics for both spatial characteristics and timbral characteristics. George found that IACC based metrics were the most useful metrics in both models which indicates their potential importance for a model predicting spatial quality. George also found that a metric designed to measure changes to timbral quality made a significant contribution to the prediction of FSF and SSF which suggested that a similar metric could be useful for the objective evaluation of spatial quality.

The metrics discussed in this chapter formed the basis of objective metrics used for calibrating the QESTRAL model for the objective evaluation of spatial quality.

9.1.5 Chapter 5 – Identifying a listening test method for the evaluation of spatial quality

Formal subjective testing is currently regarded as the most reliable method for the evaluation of audio quality. To answer aim (vi) a suitable method for reliably investigating spatial quality was required, and so existing standards for the subjective assessment of audio quality were studied. Listening test standards, BS.1116-1, BS.1534 were developed by the ITU. BS.1116-1 was designed for the detection of small impairments between stimuli and is therefore limited to the evaluation of a single test condition per test page. A large amount of data was required for the calibration of the QESTRAL model and therefore using BS.1116-1 would have been inefficient and very time consuming. By comparison BS.1534 (MUSHRA), a multistimulus test, allows several stimuli to be compared simultaneously and it is therefore a much more efficient way of collecting the amount of subjective data required for this project. However it has been observed that results collected from experiments employing the MUSHRA method suffer from biasing. Biases are systematic errors which influence the mapping process that listeners use to transfer their perception of a stimulus to the test scale. Using

biased data to calibrate a model would limit its validity and generalisability and so it was desirable to remove or reduce the appearance of bias in the data collected for this project. Therefore a new listening test method and graphical user interface (GUI) were developed that incorporate methods of reducing bias. Table 9.2 summarises the methods used to reduce each bias. The GUI is depicted in figure 9.1.

Biases known to affect audio quality listening tests	Examples of bias reduction	Method of reduction
Stimulus spacing bias	Select stimuli that are perceptually equally spaced. Randomise the presentation of stimuli	Stimuli will be carefully selected and their presentation order will be randomised.
Range equalising bias – "Rubber ruler" effect.	Use direct or indirect anchoring	Indirect anchoring
Bias due to perceptually non-linear scale	Use a label-free scale or only label the top and bottom of the scale.	GUI labels removed
Interface bias	Remove labels, numbers or markings from the interface. Use a large population of listeners.	GUI labels and markings are removed
Stimulus frequency bias	Use a balanced design (avoid presenting perceptually similar or identical stimuli more often than other stimuli).	The presentation order of stimuli will be randomised
Centring bias	Use direct or indirect anchoring.	Indirect anchoring
Recency effect bias (halo bias)	Use short looped recordings with consistent characteristics. Randomise the stimuli. Synchronously loop the stimuli.	Stimuli will be synchronously looped and their presentation order will be randomised.
Equipment bias, Listener expectation bias	Use blind listening tests. Use a large population of listeners from different backgrounds	An acoustically transparent curtain will be used to disguise the test equipment
Unfamiliarity with magnitude/stimuli	Familiarise or train the listeners before the test.	Listeners will be given test instructions and a familiarisation session.

Table 9.2 Summary of biases affecting audio quality tests (adapted from Zielinski et al [2008]) and methods of reducing them employed in the new listening test method.

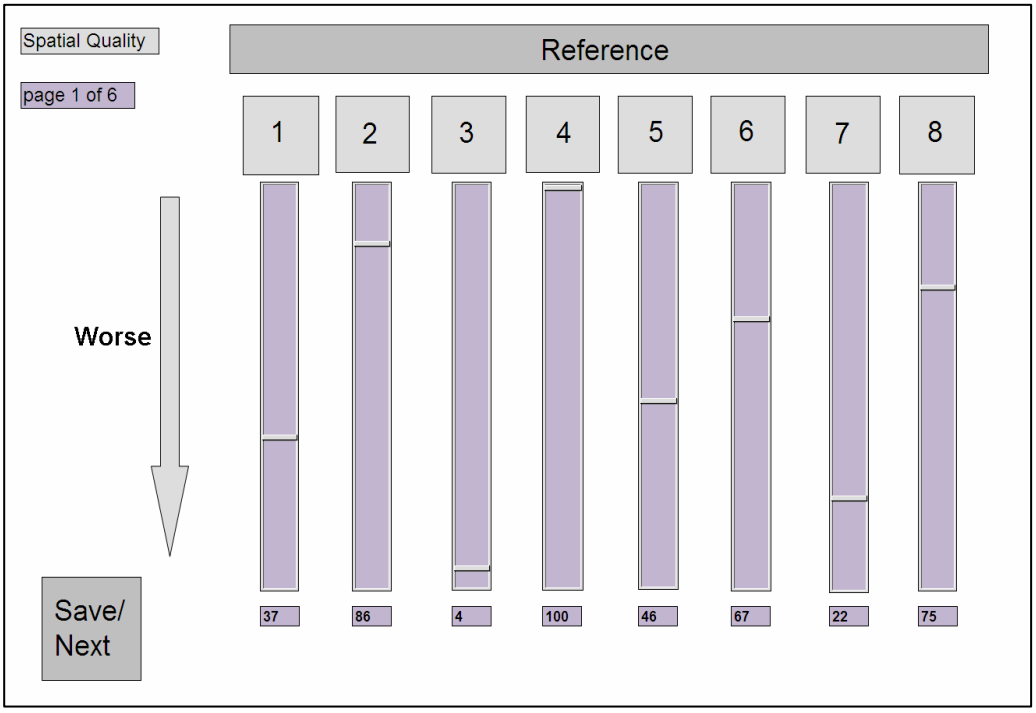


Fig 9.1 Screenshot of the proposed GUI.

9.1.6 Chapter 6 – Pilot studies

Following the development of a new listening test method, chapter 6 described and discussed four listening tests conducted as pilot studies prior to the large scale listening tests, 1 and 2, discussed in chapter 7.

To fulfil aim (vi) pilot studies 1 and 2 tested the suitability of the listening test method and GUI design proposed in chapter 5. It was tested with a wide range of different SAPs applied to a varied selection of different programme items at two listening positions. Analysing the listeners' performance in both studies showed that consistency levels similar to other listening tests were achieved using the proposed listening test method and GUI. This suggested that the listeners could use the GUI to consistently assess the spatial quality of the stimuli investigated. Therefore the test method and GUI was deemed suitable for the reliable assessment of SAPs in a large scale listening test.

The SAPs evaluated, created impairments to spatial quality across the entire range of the test scale. Suitable SAPs for use as indirect anchors were indentified in pilot study 2 (Table 9.3).

Anchor	Anchor description
Anchor recording A	High Anchor - Unprocessed reference
Anchor recording B	Mid Anchor - Audio codec (80kbs)
Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only

Table 9.3 Description of indirect anchor recordings.

A univariate ANOVA of the collected data showed that, in addition to SAP, listener, listening position and programme item influenced the perception of spatial quality. The interaction of listener with SAP had the second largest effect (after SAP) on perceived spatial quality and suggests that there was a difference in opinion or lack of consensus between listeners for certain stimuli. These stimuli were deemed to have unreliable score averages which was particularly important for calibrating the QESTRAL model, as score averages would be used to describe the spatial quality of each SAP predicted by the model. Therefore it was decided that stimuli where this effect was observed should be considered for removal.

The interactions of both listening position and programme item with SAP were also shown to have a large effect on perceived spatial quality. This suggested that certain SAPs created an impairment to spatial quality that was different when listening off-centre (LP2) than on-centre (LP1) and also different between programme items. These SAPs will also have unreliable means and therefore it was suggested that listening position and programme item should be included as separate variables in the QESTRAL model. This could be achieved either by creating different calibrations for each or by using a calibration dataset that incorporates scores collected from both listening positions separately.

In pilot study 3 a method was trialled to aid in the selection of suitable SAPs prior to listening tests 1 and 2. This was important to avoid a risk associated with using a direct method of model

development, discussed in chapter 3 and was achieved by asking listeners to evaluate the extent to which the selected SAPs exhibit changes to a range of 8 lower level spatial attributes. The method was successful as the extent to which the low-level spatial attributes were stressed was determined and in addition the distribution of the results when the stimuli were assessed in this manner was indicative of the results obtained in pilot test 1, suggesting that this method could be used to select suitable SAPs for the main listening tests.

The aims of pilot study 4 were to identify whether the types of SAPs investigated in this project affect spatial and timbral quality together or separately and whether listeners can assess timbral and spatial quality separately when the two domains are separately affected. The results showed that when a SAP affects one domain more than the other, listeners do assess them differently. However in the majority of SAPs investigated in pilot study 4 there was no significant difference in the listeners' scores between the two domains suggesting that these SAPs impaired spatial quality and timbral quality similarly and therefore it was possible that the perceived spatial quality of a stimulus would be influenced by its timbral quality. Hence as the QESTRAL model aimed to be a perceptual model, it was decided that an objective metric designed to measure changes to the timbral quality could be useful to predict the subjective scores collected from the SAPs, as George had previously indicated and would be included in the QESTRAL model calibration process.

Questionnaires conducted during pilot studies 1, 2 and 4 established that listeners found the task of scaling spatial quality easy to moderately difficult at both listening positions and found assessing spatial quality slightly more difficult than timbral quality. However as discussed above, analysing the listeners' responses has shown that it is possible for them to make reliable and consistent assessments of spatial quality using the proposed listening test method.

9.1.7 Chapter 7 – Subjective assessment of spatial quality

As chapter 6 had proven the suitability of the listening test method, and GUI, to answer aim (vii), chapter 7 described and discussed the results of two large scale listening tests which used the developed listening test method and GUI to collect a reliable database of listener scores characterising the effects of a large and varied range of SAPs on perceived spatial quality, for calibrating the QESTRAL model.

Over the two listening tests the effects on spatial quality of 48 SAPs were evaluated using six different ecologically valid programme items at two listening positions. The SAPs were chosen using the selection method discussed in pilot study 3 and created impairments to spatial quality across the whole range of the test scale. In listening test 1 listener responses were collected at an on-centre listening position (LP1) and an off-centre listening position (LP2) independently. In listening test 2 the effect of off-centre listening on spatial quality was examined and compared with on-centre listening; this lead to the development of a transform function which allowed the responses collected in listening test 1 at LP2 to be converted, allowing the data to be included in the subjective database.

The effects of these SAPs on spatial quality were examined and a number of examples were discussed. Analysing the results of the listening tests using ANOVA it was identified that differences in listener opinion, listening position and programme item type influenced the perception of spatial quality. This had also been observed previously in the results of pilot studies 1 and 2 in chapter 6. As the QESTRAL model was to be calibrated as a perceptual model it was decided that it should be sensitive to the changes in perceived spatial quality created by listening position and programme item type. Therefore these variables were incorporated into the calibration process by including the stimulus score averages collected at both listening positions for all six programme items separately. Any stimuli which elicited a large difference in opinion or lack of consensus between listeners had unreliable score averages, and so stimuli where this effect was observed were removed from the subjective database.

The entire database was analysed and the most reliable data were identified, leading to 308 scores which could be used for calibrating the QESTRAL model. The results of this data screening are summarised in tables 7.6 and 7.11 and presented in full in Appendix C.

9.1.8 Chapter 8 – Calibrating the QESTRAL model for the objective evaluation of spatial quality

To answer aim (viii) chapter 8 described the calibration and subsequent performance of the QESTRAL model for the objective evaluation of spatial quality using the subjective database collected from the listening tests described in chapter 7.

Two probe signals based upon pink noise signals were developed by the QESTRAL project team for the calibration of the QESTRAL model (Table 7.1). These were designed to allow the measurement of changes in the foreground and background audio streams.

Fourteen metrics were developed to predict the spatial characteristics tested during the listening tests. The development of the metrics was informed by identifying which lower level spatial attributes had been stressed by the SAPs evaluated during the listening tests in chapter 7. As suggested by the conclusions of pilot study 4, an additional metric to measure changes to timbral characteristics was also included. These metrics are described in table 9.4

After calibration and correction the QESTRAL model performed close to the target specifications (Table 9.5). It also passed a number of statistical tests designed to measure its potential generalisability (see Appendix J). The objective metrics used in the final model are described alongside their regression coefficients in table 9.6. The corrected QESTRAL model equation is given in equation 9.1.

	Metric	Probe signal	Description	R
1	IACC0	1	The mean IACC value calculated across 22 frequency bands (150Hz-10kHz) calculated from a 0° head rotation.	0.64
2	IACC90	1	The mean IACC value calculated across 22 frequency bands (150Hz-10kHz) calculated from a 90° head rotation.	0.51
3	IACC0*IACC90	1	The product of IACC0 and IACC90.	0.62
4	IACC0_9band	1	The mean IACC 0 value calculated from 9 frequency bands (570Hz-2160Hz).	0.71
5	IACC90_9band	1	The mean IACC 90 value calculated from 9 frequency bands (570Hz-2160Hz).	0.53
6	IACC0*IACC90_9band	1	The product of IACC0_9Band and IACC90_9Band.	0.66
7	Mean_Ang_FrontWeighted	2	The mean absolute change to localisation, compared with the reference localisation for the 36 noise bursts, with a linear weighting of decreasing importance from 0° applied to each angle.	0.67
8	Mean_Ang_Front60	2	The mean absolute change to localisation, compared to reference localisation for 7 noise bursts between 0-30° and 330-350°.	0.73
9	Hull	1	The convex area of the localised 36 noise burst plotted on a unit circle	-0.56
10	CardKLT	1	The contribution in percent of the first eigenvector from a Karhunen-Loeve Transform (KLT) decomposition of four cardioid microphones placed at the listening position and facing in the following directions: 0˚, 90˚, 180˚ and 270˚.	0.60
11	Mean_Entropy	1	The mean Shannon entropy value measured from both binaural signals.	-0.58
12	TotEnergy	1	RMS of pressure value measured by a pressure microphone.	-0.27
13	Mean_RMS_diff	2	The mean absolute change to RMS compared with the reference RMS for the 36 noise bursts.	0.55
14	Mean_SpecRollOff	1	The mean magnitude of the FFT from both binaural signals.	-0.20

Table 9.4 Metrics employed for the calibration of the QESTRAL model.

	QESTRAL model	Target specifications
Correlation (r)	0.89	≥ 0.86
RMSE (%)	11.06%	≈ 10%
No. of metrics	5	Low
No. of PCs	2	Low
VIF (max)	2	Mean VIF ≈ 1

Table 9.5 QESTRAL model performance results.

Metric	Probe signal	Description	Regression coefficient
IACC0_9band	1	The mean IACC 0 value calculated from 9 frequency bands (570Hz-2160Hz).	61.887
Mean_Ang_Front60	2	The mean absolute change to localisation, compared to reference localisation for 7 noise bursts between 0-30° and 330-350°.	0.352
Mean_Entropy	1	The mean Shannon entropy value measured from both binaural signals.	-23.017
Mean_RMS_diff	2	The mean absolute change to RMS compared with the reference RMS for the 36 noise bursts.	695.407
Mean_SpecRollOff	1	The mean magnitude of the FFT from both binaural signals.	-0.002153
		Constant	89.069916

Table 9.6 QESTRAL model objective metrics and regression coefficients.

$$QESTRAL_{corrected} = 14.102e^{0.022QESTRAL_{model}} - 0.069 \quad \text{(eq 9.1)}$$

Investigating the QESTRAL model's performance in the prediction of different test variables (i.e. SAPs, programme items and listening positions) revealed that it performs well over the evaluated SAPs applied to the six programme items at both listening positions. However closer inspection of the predicted scores revealed that using programme item dependent subjective scores (to make the model more sensitive to the changes created by different programme item types) was a source of error in the model. In particular this showed that the QESTRAL model was not accurate when SAPs were perceived as identical (error along the vertical y-axis) nor when identical SAPs were applied to different programme items (error along the horizontal x-axis) (for examples see figures 8.13 and 8.14). These errors contributed to the RMSE (%) of the calibrated QESTRAL model. However, the model's performance was close to the target specifications in the prediction of each of the independent variables, and the error itself was similar to the average listener error of listening tests 1 and 2.

Two ways of removing this inaccuracy were suggested either by removing programme material as a variable in the model or by calibrating different versions of the model for programme items with different scene types. However both suggestions would limit the practical usage of the QESTRAL model. Hence the current method of calibration was believed to be the most suitable from both a practical and performance standpoint.

9.2 Limitations of the QESTRAL model and future work

This section discusses the limitations of the current calibration of the QESTRAL model in order to suggest ideas for future work.

9.2.1 Expanding the generalisability of the QESTRAL model

There are a number of ways in which the generalisability of the QESTRAL model could be expanded. The QESTRAL model was calibrated for use with a 3/2 stereo reproduction system. Although this system is currently the most commercially successful spatial reproduction system, as discussed in section 3.4 there are other systems that are currently gaining popularity (e.g. Ambisonics, 7.1, 10.2 [Rumsey, 2001] and Wave Field Synthesis (WFS) [De Vries, 2007]). Audio reproduction in automobiles is also currently of commercial interest, with manufacturers such as Bang & Olufsen and Bose attempting to deliver high sound quality to automobile users [Bang & Olufsen, 2010][Bose, 2010]. The evaluation scheme employed by the QESTRAL model allows it to be reproduction independent. Using the methods employed in this research project should enable the calibration of the QESTRAL model for other reproduction systems. A research project is currently underway which

aims to develop the QESTRAL model for the evaluation of car audio systems [University of Surrey, 2010].

The generalisability of the model is also limited by the population of listeners used to collect subjective scores. Listeners were employed exclusively from post-graduate and under-graduate Tonmeisters at the IoSR, University of Surrey. It is accepted that the opinions of this group of experienced listeners may differ from those of non-experienced listeners. Rumsey *et al* [2005] have shown that the opinions of experienced and non-experienced listeners can be similar. Nevertheless to expand the generalisability of the model subjective scores could be collected from other populations of listeners.

Although the programme items used in this research project were chosen as representative examples of the commercially available 5-channel audio recordings. The model could be improved by evaluating a larger number of programme items (as George [2009] has done). In this respect it may also be worth investigating whether calibrating the model for different scene types may yield better results.

As preliminary work has shown [Jackson *et al*, 2010] it has been possible, using an evaluation model developed by Dewhirst [Dewhirst *et al*, 2005], to estimate the impairment to spatial quality, created by different SAPs across the listening area. These estimates could be improved by collecting subjective scores from a greater number of listening positions.

9.2.2 Improving the performance of the QESTRAL model

The probe signals used in the QESTRAL model are based upon pink noise signals. More complex signals could be designed that may yield better results. These could be more programme-like and exhibit typical properties of programme items such as scene type, or be optimised for the measurement of individual attributes. Mason [2006] provides some examples of such probe signals.

Examples of how objective metrics such as IACC could be optimised for the prediction of spatial quality were given in section 8.2.2. It may be possible to optimise the other metrics used in the model or develop others (such as those discussed in chapter 4) for the prediction of the spatial quality.

9.3 Contributions to knowledge

The completion of this research has yielded a number of distinct contributions to knowledge. These contributions are outlined below.

A novel and repeatable listening test method for the subjective assessment of spatial quality

A new multistimulus listening test method was developed that incorporates methods of reducing bias in audio quality listening tests.

Development of a method to determine the suitability of SAPs used to calibrate the QESTRAL model

A method was developed in pilot study 3 which allowed an optimal or balanced selection of SAPs to be chosen for evaluation in large scale experiment, ensuring that developing the QESTRAL model using a direct method did not limit its generalisability.

The identification and analysis of spatial audio processes (SAPs) that result in diverse judgements of spatial quality

In listening test 1 and 2 a database of subjective scores, describing the perceived impairment to spatial quality arising from a wide range of different SAPs commonly encountered by consumers, was collected. The effects of these SAPs on spatial quality were detailed in figures 7.7 and 7.13 and appendix K, showing that they create impairments to spatial quality that span the whole range of the test scale.

The identification and analysis of test variables that influence the perception of spatial quality

A univariate ANOVA of the subjective data collected from listening tests 1 and 2 showed that in addition to SAP, listener, listening position and programme item type influenced the perception of spatial quality.

Identification, creation and development of appropriate objective metrics for the objective evaluation of perceived spatial quality

Fourteen different metrics were developed by the QESTRAL project team to measure the changes in spatial quality. Each was designed to analyse either probe signal 1, or probe signal 2, as received by a virtual binaural simulator or other virtual microphone receivers at the listening position simulated in the QESTRAL model.

The calibration of a perceptual model (QESTRAL model) for the objective evaluation of perceived spatial quality

The QESTRAL model is an objective evaluation model that, using five objective metrics, is capable of accurately predicting changes to perceived spatial quality created by a large range SAPs applied to six ecologically valid programme items at both listening position 1 and 2.

9.4 Publications contributed to by this research project

This research has contributed to six published papers, two conference abstracts, one poster presentation and one piece of software; these are listed below in chronological order. The various publications give an overview of the QESTRAL project and present the results of preliminary studies. The software is an online public-use version of the QESTRAL model created in this thesis.

9.4.1 Conference & Convention papers

Rumsey, F., Zielinski, S., Jackson, P.J.B, Dewhirst, M., Conetta, R., George, S., Bech, S. & Meares D. (2008) "QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener" presented at the Audio Engineering Society 125th Convention, Oct 2 – 5, San Francisco, Preprint 7595.

Conetta, R., Rumsey, F., Zielinski, S., Jackson, P.J.B, Dewhirst, M., Bech, S., Meares D. & George, S (2008). "QESTRAL (Part 2): Calibrating the QESTRAL spatial quality model using listening test data" presented at the Audio Engineering Society 125th Convention, Oct 2 – 5, San Francisco, Preprint 7596.

Jackson, P.J.B, Dewhirst, M., Conetta, R., Rumsey, F., Zielinski, 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.

Dewhirst, M., Conetta, R., Rumsey, F., Jackson, P.J.B, Zielinski, S., Bech, S., Meares D. & George, S (2008) "QESTRAL (Part 4): Test signals, combining metrics and the prediction of overall spatial quality" presented at the Audio Engineering Society 125th Convention, Oct 2 – 5, San Francisco, USA, Preprint 7598.

Conetta, R., Rumsey, F., Zielinski, S., Jackson, P.J.B, Dewhirst, M., Bech, S., Meares D. & George, S (2008). "Calibration of the QESTRAL model for the prediction of spatial quality" proceedings of the Institute of Acoustics 24th Reproduced Sound Conference, Nov 20-21, Brighton, UK.

Jackson, P.J.B., Dewhirst, M., Conetta, R. & Zielinski, S. (2010) "Estimates of perceived spatial quality across the listening area" presented at the Audio Engineering Society 31st International Conference, Jun 13 – 15, Pitea, Sweden.

9.4.2 Conference abstracts

Rumsey, F., Zielinski, S., Jackson, P.J.B, Dewhirst, M., Conetta, R., Bech, S. & Meares D. (2008) "Measuring perceived spatial quality changes in surround sound reproduction" presented at the 155th meeting of the Acoustical Society of America, Acoustics 2008, June 30, Paris, France, p2280, (invited).

Jackson, P.J.B, Rumsey, F., Zielinski, S., Dewhirst, M., Conetta, R., Bech, S. & Meares D. (2008) "Prediction of spatial perceptual attributes of reproduced sound across the listening area" presented at the 155th meeting of the Acoustical Society of America, Acoustics 2008, June 30, Paris, France, p2279.

9.4.3 Posters

Conetta, R., Jackson, P.J.B., Zielinski, S. & Rumsey, F., (2007) "Envelopment: What is it? A definition for multichannel audio" presented at the 1st SpACE-Net Workshop, Jan 25, University of York, UK.

9.4.4 Software

George, S., Dewhirst, M., Conetta, R., Zielinski, S., Rumsey, F., Jackson, P.J.B., Bech, S., Meares D. & Supper, B (2009) "QESTRAL demonstrator", Online, version 1.0.

Appendix A - Listener instructions for listening tests

A.1 Listener instructions for pilot study 1 and 3 and listening tests 1 and 2

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.

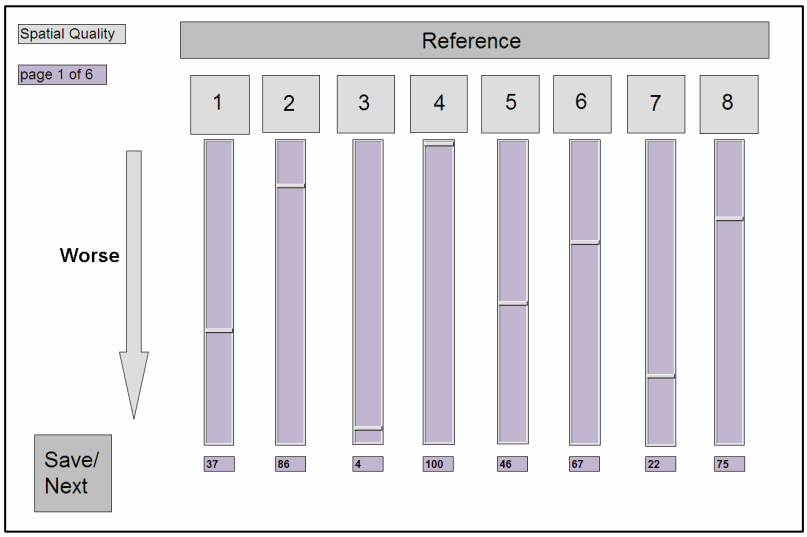
A spatial quality scale is a hybrid scale that is primarily a fidelity evaluation (one measuring the degree of similarity to the reference). However it also enables you to give an opinion about the extent to which any differences are inappropriate, unpleasant or annoying. In other words, which affect 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 (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 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.

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**.

A.2 Listener instructions for pilot study 4

Thank you for participating in this training experiment.

Please read the instructions below.

Description of the task and scale
You are asked to compare a number of sound recordings, which have been processed in various ways, with an unprocessed original reference recording. You are asked to firstly rate the spatial quality of the processed recordings and then on the following page rate the timbral quality (or vice versa). Both tasks will be completed using the same test scale.

What follows is a description of how you should use the test scale for each task.

Spatial Quality
A spatial quality scale is a hybrid scale that is primarily used for 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)

Timbral Quality
A timbral quality scale is a hybrid scale that is primarily used for 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 timbral reproduction compared with the reference. So, for example, if you can hear a change in the timbral reproduction compared with the reference but it doesn't make much difference to your overall opinion about the timbral quality, you should rate it towards the top of the scale. On the other hand, if the timbral 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 timbral 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 timbral 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 timbral.

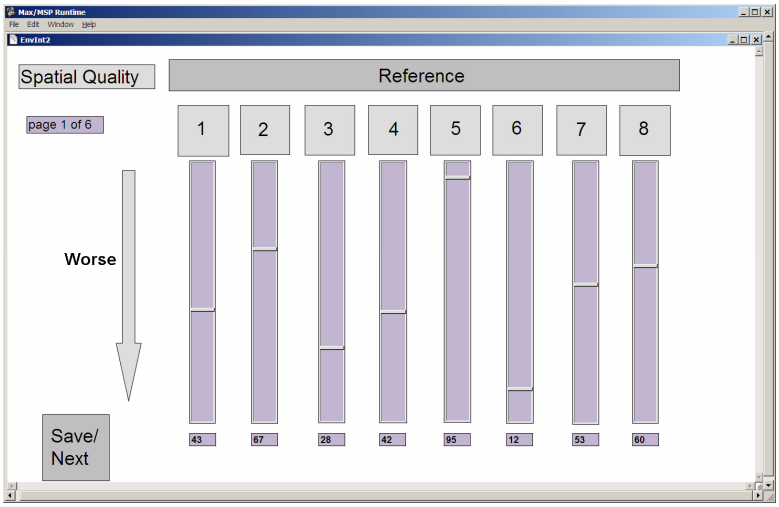
"Timbre enables the listener to judge that two sounds which have, but do not have to have, the same spaciousness, loudness, pitch, and duration are dissimilar." Letowski (1989)

The following are examples of changes to timbral quality 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 brightness
- Changes in sharpness or clarity
- Changes in colouration
- Changes in powerfulness

User Interface

Each page contains 8 test recordings to be evaluated alternately for **timbral quality** or **spatial quality** against a reference recording. The required evaluation for a page is given in the top left hand corner.



There are 12 pages in the test. These are split over two parts, 'Training 1' and 'Training 2' (6 pages in each).

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).

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

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.**

*****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**.

Appendix B – Univariate ANOVA structure

$$Y_{A,B,X,\Delta,E} = \pi + \alpha_A + \beta_B + \chi_X + \delta_\Delta + \varepsilon_E + \phi_{A,B} + \varphi_{A,X} + \gamma_{A,\Delta} + \eta_{A,E} + \iota_{B,X} + \kappa_{B,\Delta} + \lambda_{B,E} + \mu_{X,\Delta} + \nu_{X,E} + o_{\Delta,E} + \varpi_{A,B,X,\Delta,E}$$

(eq. B1)

Where:
π = overall mean,
α_A = SAP 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 SAP,
$\varphi_{A,X}$ = interaction of programme item with SAP,
$\gamma_{A,\Delta}$ = interaction of session with SAP,
$\eta_{A,E}$ = interaction of listener with SAP,
$\iota_{B,X}$ = 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 $\varpi_{A,B,X,\Delta,E}$ = the error.

Appendix C – Analysing screening and removing data influenced by listener

ANOVA revealed that the variable "listener" had a statistically significant ($p < 0.05$) influence on the scoring of spatial quality in pilot studies 1 and 2 and listening tests 1 and 2. This suggested that there was a difference in opinion or lack of consensus between listeners in their scoring of the spatial quality for certain stimuli. Any stimulus which exhibits a large difference in opinion or lack of consensus will have unreliable score averages. This is particularly important for the development of the QESTRAL model where score average values will be used to describe the spatial quality score for each stimulus in the model, and therefore these stimuli should be considered for removal from the calibration data set.

In order to screen data influenced by "listener" the distribution of the subjective scores for each stimulus was analysed using a combination of statistical and visual analysis techniques.

C.1 Normality

The normality of the distribution of the listener scores for each stimulus was assessed using a kolmogorov-smirnov analysis. Stimuli which did not have a normal distribution of listener scores were not automatically removed.

C.2 Modality

To assess whether the distribution had more than one mode, each stimulus was assessed statistically (using SPSS) and visually. If the distribution had more than two predominant peaks or modes it was considered for removal.

C.3 Spread or range

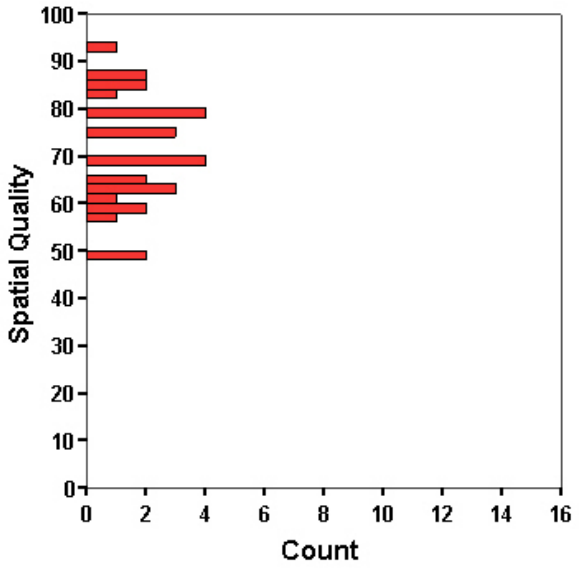
To assess the spread and flatness of the distributions the following statistical analyses were used:

- Standard deviation > 20.
- Range < 75 – The range of the scale which the listener's scores cover.
- Kurtosis (z-score) > -1 – A statistical measure of the flatness of the distribution.

Any stimulus failing each test was automatically removed. However as a rule the results of statistical analysis were used as a guide, visual assessment was always used to make the final decision. If a stimulus passed these tests and was found to have a statistically normal distribution the mean value of the distribution was used. Whereas if the stimulus did not have a statistically normal distribution the median value of the distribution was used (NB. In cases where the distribution was assessed as being

both normally distributed and multimodal the most suitable value (mean/median) was selected via a visual assessment).

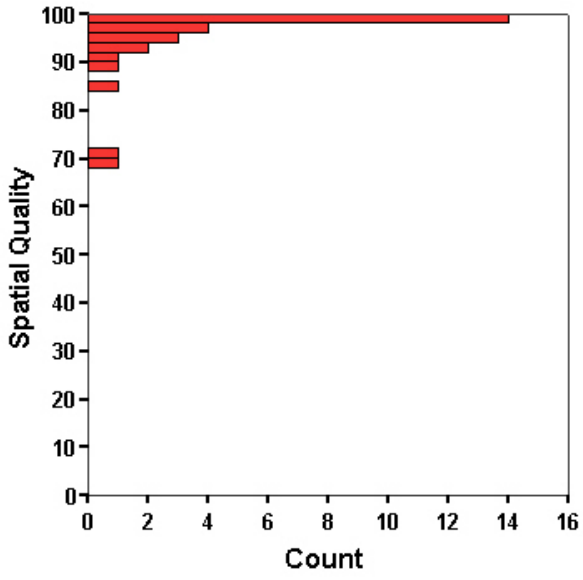
Figures C1 -3 illustrate examples of data distributions, analysis results and decision outcomes for three different stimuli.



Stats:-
Normality: p = 0.200 (lower bound of true significance)
Modality (stats): No
Modality (visual): No
Std. Deviation: 11.559
Range: 44
Kurtosis (z score): -0.82

Result:- Mean value reported (72)

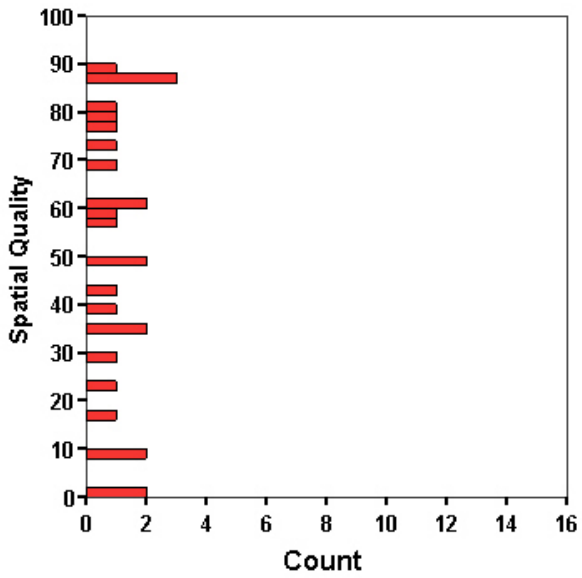
Fig C1 Example of a data distribution where the mean value was reported.



Stats:-
Normality: p = 0.640
Modality (stats): No
Modality (visual): No
Std. Deviation: 7.861
Range: 30
Kurtosis (z score): 6.39

Result:- Median value reported (99)

Fig C2 Example of a data distribution where the median value was reported.



Stats:-
Normality: p = 0.200 (lower bound of true significance)
Modality (stats): Yes
Modality (visual): Yes
Std. Deviation: 29.022
Range: 90
Kurtosis (z score): -1.23

Result:- Removed from data set

Fig C3 Example of a data distribution which was removed from the data set.

C.4 Results

Tables C1 – 23 summarise the results of the analysis technique for each stimulus, primarily identifying which stimuli should be removed, but also establishing for each stimulus whether mean or median values should be used.

C.4.1 Pilot study 1

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2	X		X						X		
3		X								X	
4		X	X							X	
5	X		X						X		
6		X								X	
7	X			X					X		
8	X								X		

Table C1. Stimulus analysis results for pilot study 1, listening position 1, programme item 1.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2	X		X	X			X				X
3		X					X			X	
4	X								X		
5		X								X	
6		X								X	
7		X								X	
8		X								X	

Table C2. Stimulus analysis results for pilot study 1, listening position 1, programme item 2.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2		X	X							X	
3	X		X	X			X				X
4	X				X	X					X
5		X								X	
6		X								X	
7	X		X							X	
8		X								X	

Table C3. Stimulus analysis results for pilot study 1, listening position 1, programme item 3.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X	X							X	
2	X								X		
3	X			X	X						X
4	X						X		X		
5	X								X		
6		X								X	
7	X		X	X			X				X
8		X								X	

Table C4. Stimulus analysis results for pilot study 1, listening position 1, programme item 4.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1	X								X		
2	X		X						X		
3	X						X		X		
4	X			X			X		X		
5		X	X	X			X			X	
6		X								X	
7	X		X				X		X		
8		X								X	

Table C5. Stimulus analysis results for pilot study 1, listening position 2, programme item 1.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1	X								X		
2	X		X	X			X		X		
3		X								X	
4	X				X		X		X		
5	X		X						X		
6		X								X	
7	X								X		
8		X								X	

Table C6. Stimulus analysis results for pilot study 1, listening position 2, programme item 2.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2	X						X		X		
3	X								X		
4	X		X	X	X						X
5		X								X	
6		X								X	
7	X		X						X		
8		X								X	

Table C7. Stimulus analysis results for pilot study 1, listening position 2, programme item 3.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1	X								X		
2	X								X		
3	X								X		
4	X		X	X			X				X
5	X								X		
6		X								X	
7	X								X		
8		X								X	

Table C8. Stimulus analysis results for pilot study 1, listening position 2, programme item 4.

C.4.2 Pilot study 2

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1	X								X		
2	X								X		
3	X		X	X	X		X				X
4	X		X	X	X						X
5	X			X					X		
6	X		X				X		X		
7		X	X							X	
8	X				X	X			X		
9	X		X						X		
10		X	X							X	
11		X								X	
12		X								X	
13		X								X	

Table C9. Stimulus analysis results for pilot study 2, programme item 1.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2		X								X	
3		X	X							X	
4	X				X	X					X
5	X		X						X		
6	X		X		X	X					X
7	X		X						X		
8	X		X						X		
9	X		X	X	X						X
10	X								X		
11		X			X	X				X	
12		X								X	
13		X								X	

Table C10. Stimulus analysis results for pilot study 2, programme item 2.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1	X								X		
2		X								X	
3		X								X	
4	X								X		
5	X		X						X		
6	X		X	X	X	X	X				X
7	X		X	X	X		X				X
8	X						X				X
9		X			X						X
10	X		X						X		
11		X								X	
12		X	X							X	
13		X								X	

Table C11. Stimulus analysis results for pilot study 2, programme item 3.

C.4.3 Listening test 1

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X	X							X	
2	X								X		
3	X								X		
4		X	X							X	
5	X		X						X		
6	X								X		
7		X		X	X		X				X
8		X								X	
9		X					X			X	
10	X								X		
11		X								X	
12		X								X	
13	X		X				X		X		
14		X								X	
15	X					X			X		
16	X		X						X		
17	X		X	X		X				X	
18	X								X		
19		X								X	
20		X								X	
21	X								X		
22	X		X						X		
23		X	X	X	X					X	
24	X		X						X		
25	X								X		
26		X								X	
27	X		X						X		
28	X		X		X	X					X
29	X			X	X		X				X
30	X								X		
31		X								X	
32	X						X		X		
33		X								X	
34	X		X		X	X				X	
35	X								X		
36	X		X						X		
37	X		X						X		
38		X								X	
39		X								X	
40	X		X				X		X		
41		X								X	
42	X						X		X		
43		X								X	

Table C12. Programme item 1, Listening position 1. Summary of subjective score distribution analysis.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2		X								X	
3	X		X						X		
4		X								X	
5		X								X	
6	X								X		
7		X	X	X			X				X
8	X						X		X		
9	X								X		
10	X		X						X		
11		X								X	
12		X								X	
13		X								X	
14		X								X	
15	X		X	X	X	X	X				X
16	X		X						X		
17	X		X	X	X		X				X
18	X		X	X						X	
19		X		X							X
20	X		X	X	X	X				X	
21		X								X	
22	X								X		
23	X		X	X							X
24	X								X		
25	X		X		X	X				X	
26		X								X	
27	X		X				X		X		
28	X								X		
29		X	X							X	
30	X			X	X		X				X
31		X								X	
32	X			X	X		X				X
33		X	X							X	
34	X		X				X				X
35	X		X						X		
36	X						X		X		
37	X								X		
38		X			X					X	
39		X								X	
40	X			X	X	X					X
41		X								X	
42	X					X	X		X		
43		X								X	

Table C13. Programme item 2, Listening position 1. Summary of subjective score distribution analysis.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2	X								X		
3	X		X						X		
4		X								X	
5		X								X	
6	X		X	X		X					X
7		X		X							X
8	X								X		
9		X	X	X						X	
10	X								X		
11	X								X		
12		X								X	
13		X								X	
14	X								X		
15		X				X				X	
16	X		X						X		
17	X		X		X						X
18	X			X						X	
19		X								X	
20	X		X						X		
21		X					X			X	
22	X		X		X	X			X		
23		X		X						X	
24	X						X		X		
25	X					X			X		
26		X				X				X	
27	X								X		
28	X		X	X		X	X				X
29	X			X	X				X		
30	X		X						X		
31		X								X	
32	X		X						X		
33		X								X	
34	X		X			X			X		
35		X	X							X	
36	X								X		
37	X						X		X		
38	X		X						X		
39		X								X	
40	X		X	X	X	X	X				X
41		X								X	
42		X				X				X	
43		X								X	

Table C14. Programme item 3, Listening position 1. Summary of subjective score distribution analysis.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2	X		X						X		
3	X					X			X		
4	X		X						X		
5	X								X		
6		X	X							X	
7	X		X						X		
8	X		X						X		
9	X		X						X		
10	X		X						X		
11	X		X						X		
12	X		X						X		
13	X								X		
14		X	X							X	
15	X		X						X		
16		X				X	X				X
17	X		X				X				X
18		X	X			X					X
19	X								X		
20	X								X		
21		X								X	
22	X								X		
23		X								X	
24		X								X	
25		X	X				X			X	
26		X								X	
27		X					X			X	
28	X								X		
29		X								X	
30	X								X		
31	X		X						X		
32		X								X	
33		X								X	
34		X					X			X	
35	X								X		
36		X								X	
37	X		X						X		
38	X								X		
39		X								X	
40	X		X						X		
41		X								X	
42	X								X		
43		X								X	

Table C15. Programme item 1, Listening position 2. Summary of subjective score distribution analysis.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2		X								X	
3	X		X		X	X			X		
4		X		X	X		X				X
5		X								X	
6	X								X		
7	X		X						X		
8	X						X		X		
9	X		X						X		
10	X		X		X				X		
11	X								X		
12		X								X	
13		X								X	
14		X								X	
15	X		X		X				X		
16	X		X		X		X		X		
17	X		X	X							X
18		X								X	
19		X								X	
20	X	X	X		X				X		
21		X								X	
22		X								X	
23	X			X	X						X
24	X		X						X		
25		X	X	X	X		X				X
26		X								X	
27		X								X	
28	X								X		
29		X								X	
30		X			X		X			X	
31		X								X	
32		X	X				X			X	
33		X	X							X	
34	X				X	X			X		
35	X								X		
36		X								X	
37	X		X						X		
38		X								X	
39	X								X		
40	X				X		X		X		
41		X								X	
42		X								X	
43		X								X	

Table C16. Programme item 2, Listening position 2. Summary of subjective score distribution analysis.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2	X		X						X		
3	X								X		
4		X		X			X				X
5	X								X		
6	X								X		
7	X								X		
8	X		X		X	X					X
9	X								X		
10	X		X						X		
11		X								X	
12	X		X						X		
13	X		X						X		
14		X								X	
15	X		X							X	
16	X		X						X		
17	X				X	X					X
18	X		X						X		
19		X								X	
20		X	X							X	
21		X								X	
22		X			X					X	
23	X		X	X	X						X
24	X								X		
25		X	X		X		X				X
26	X		X						X		
27		X								X	
28	X		X						X		
29		X	X							X	
30	X		X				X		X		
31	X								X		
32	X						X		X		
33		X	X							X	
34	X								X		
35		X								X	
36		X								X	
37	X		X						X		
38	X		X						X		
39	X								X		
40	X		X						X		
41		X								X	
42		X								X	
43		X								X	

Table C17. Programme item 3, Listening position 2. Summary of subjective score distribution analysis.

C.4.4 Listening test 2

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2		X								X	
3	X		X		X						X
4		X								X	
5	X			X	X	X					X
6	X		X	X		X					X
7	X		X						X		
8		X								X	
9	X								X		
10		X								X	
11		X								X	
12		X								X	
13		X								X	
14	X		X							X	
15	X		X	X	X	X					X
16		X		X		X	X				X
17		X			X		X				X
18	X		X				X			X	
19	X						X		X		
20	X				X	X					X
21		X								X	
22		X				X				X	
23		X								X	

Table C18. Programme item 4, Listening position 1. Summary of subjective score distribution analysis.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2	X		X						X		
3	X								X		
4		X								X	
5	X		X		X	X			X		
6	X				X	X			X		
7	X								X		
8		X								X	
9	X				X	X					
10		X								X	
11		X								X	
12		X								X	
13		X								X	
14		X								X	
15		X		X	X	X					X
16		X	X		X	X				X	
17	X		X	X	X	X					X
18	X								X		
19		X	X			X				X	
20	X		X	X		X	X				X
21		X								X	
22		X				X				X	
23		X								X	

Table C19. Programme item 5, Listening position 1. Summary of subjective score distribution analysis.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1		X								X	
2	X								X		
3	X		X	X		X					X
4		X								X	
5	X		X							X	
6	X		X		X	X			X		
7	X								X		
8		X								X	
9		X				X				X	
10		X								X	
11		X								X	
12		X								X	
13		X								X	
14		X	X	X		X	X				X
15		X	X	X	X		X				X
16		X	X	X		X	X				X
17	X			X	X	X					X
18	X		X	X		X					X
19	X		X				X			X	
20	X		X	X		X			X		
21		X								X	
22		X								X	
23		X								X	

Table C20. Programme item 6, Listening position 1. Summary of subjective score distribution analysis.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1	X		X						X		
2	X		X						X		
3		X	X	X	X		X				X
4	X								X		
5	X		X			X			X		
6	X		X						X		
9	X		X				X		X		
10		X								X	
11	X								X		
12	X		X	X					X		
13	X								X		
14	X		X						X		
15	X		X						X		
16	X		X		X		X				X
17	X					X			X		
18	X								X		
19	X								X		
21	X		X						X		
22	X					X			X		
23	X						X		X		

Table C21. Programme item 4, Listening position 2. Summary of subjective score distribution analysis.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1	X		X	X	X		X		X		
2	X		X						X		
3	X		X				X				X
4	X		X	X			X		X		
5	X								X		
6	X				X		X		X		
9	X		X		X		X				X
10	X		X		X	X					X
11	X		X						X		
12	X		X						X		
13	X		X				X		X		
14	X						X		X		
15	X		X						X		
16	X			X	X	X					X
17	X		X						X		
18	X			X					X		
19		X				X				X	
21	X		X						X		
22		X				X				X	
23	X								X		

Table C22. Programme item 5, Listening position 2. Summary of subjective score distribution analysis.

SAP	Norm (stat)	Not Norm (stat)	MM (stat)	MM (vis) >2	Std > 20	Range >75	Kurtosis Z score <-1		Mean	Median	Remove
1	X								X		
2		X	X		X	X				X	
3	X						X		X		
4	X								X		
5	X		X						X		
6	X		X	X			X		X		
9	X								X		
10	X								X		
11	X		X						X		
12	X		X						X		
13	X								X		
14		X		X		X	X				X
15	X		X						X		
16	X					X			X		
17	X		X	X	X		X				X
18	X			X	X		X				X
19	X								X		
21		X	X							X	
22	X		X						X		
23	X								X		

Table C23. Programme item 6, Listening position 2. Summary of subjective score distribution analysis.

Appendix D – Means and 95% confidence intervals for SAPs whose subjective scores were influenced by listening position in pilot study 1 and listening test 1

D.1 Pilot study 1

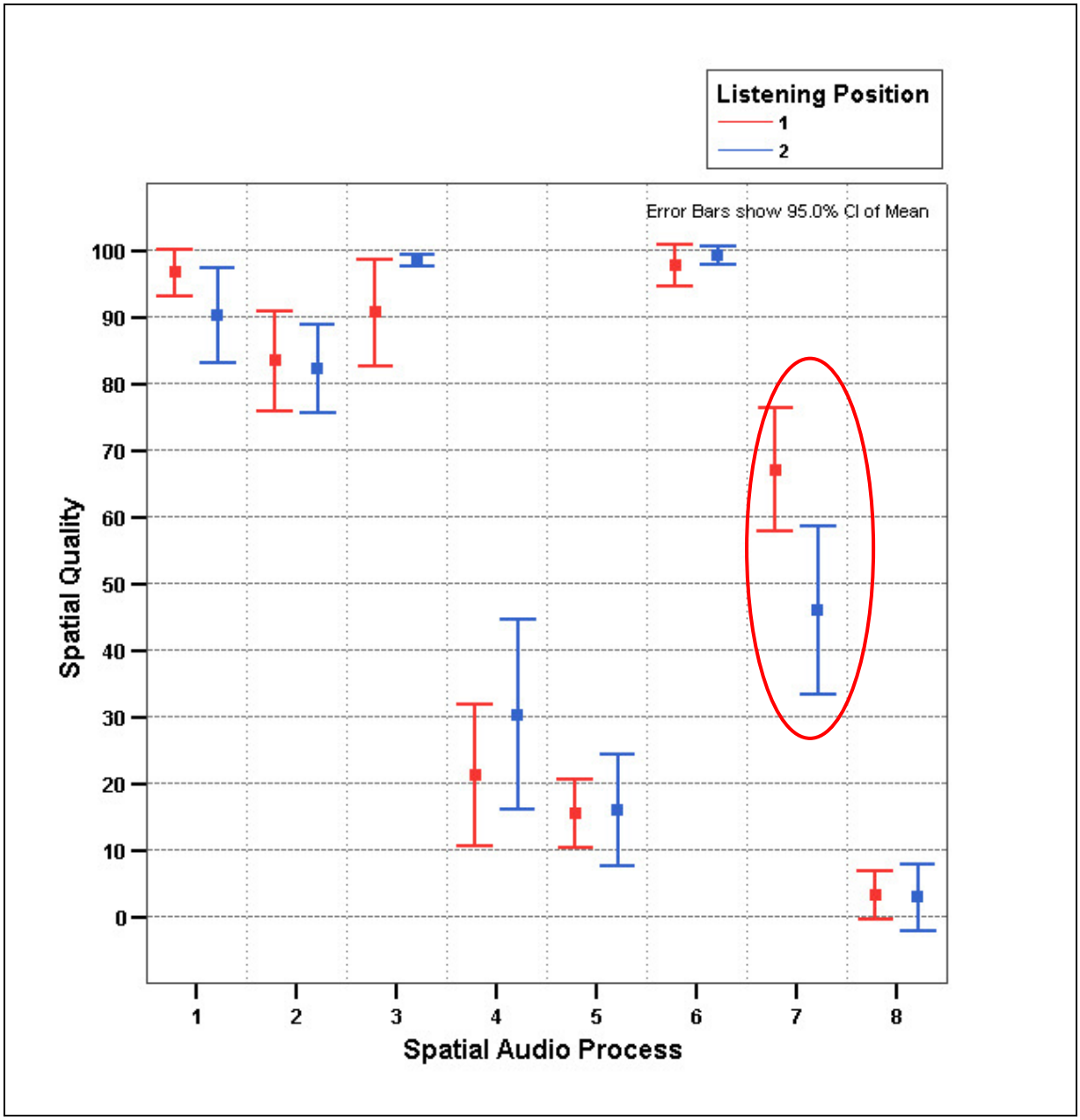


Fig D1. SAPs (circled in red) which create a difference in perceived spatial quality between listening positions with programme item 1 in pilot study 1.

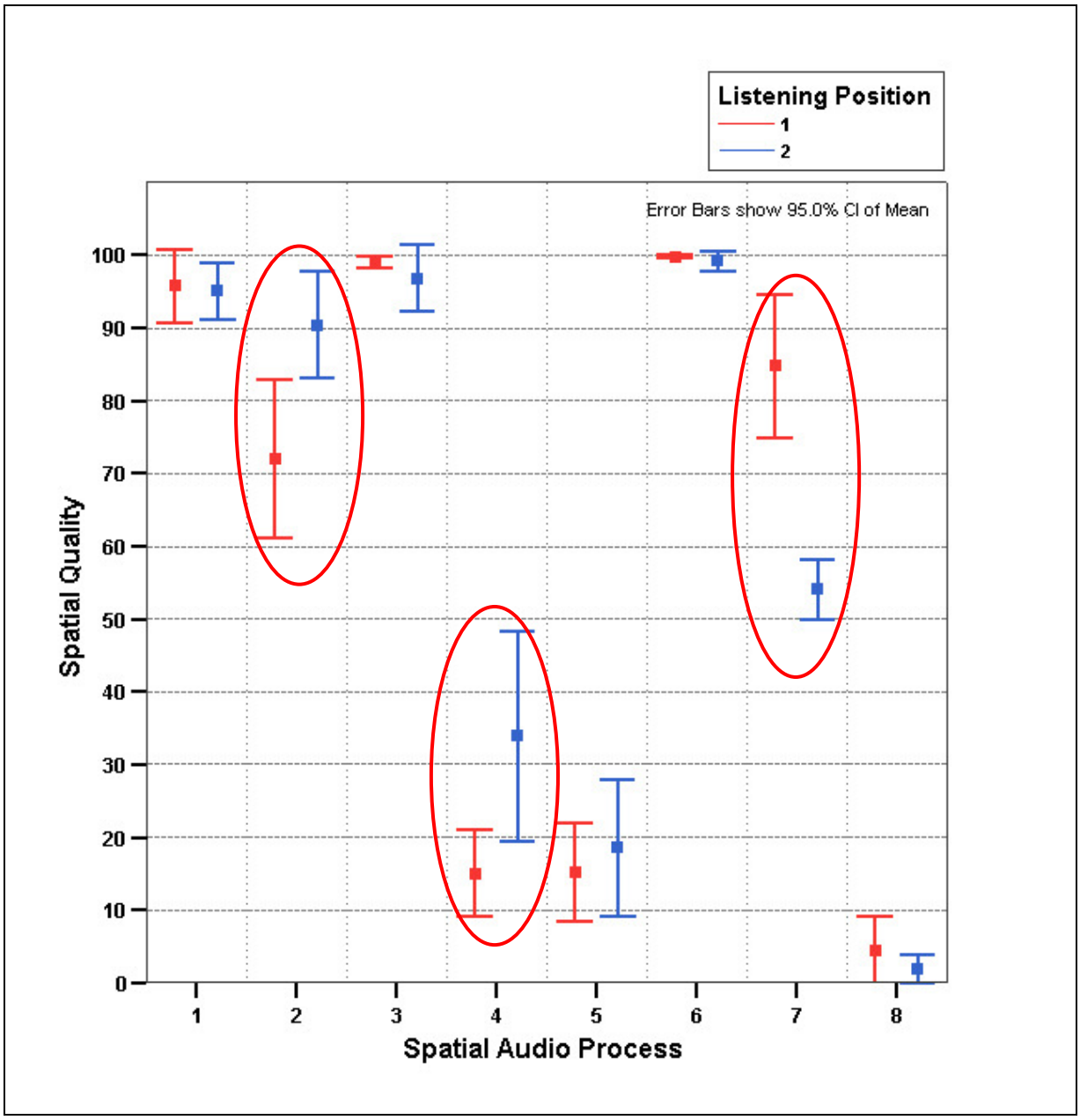


Fig D2. SAPs (circled in red) which create a difference in perceived spatial quality between listening positions with programme item 2 in pilot study 1.

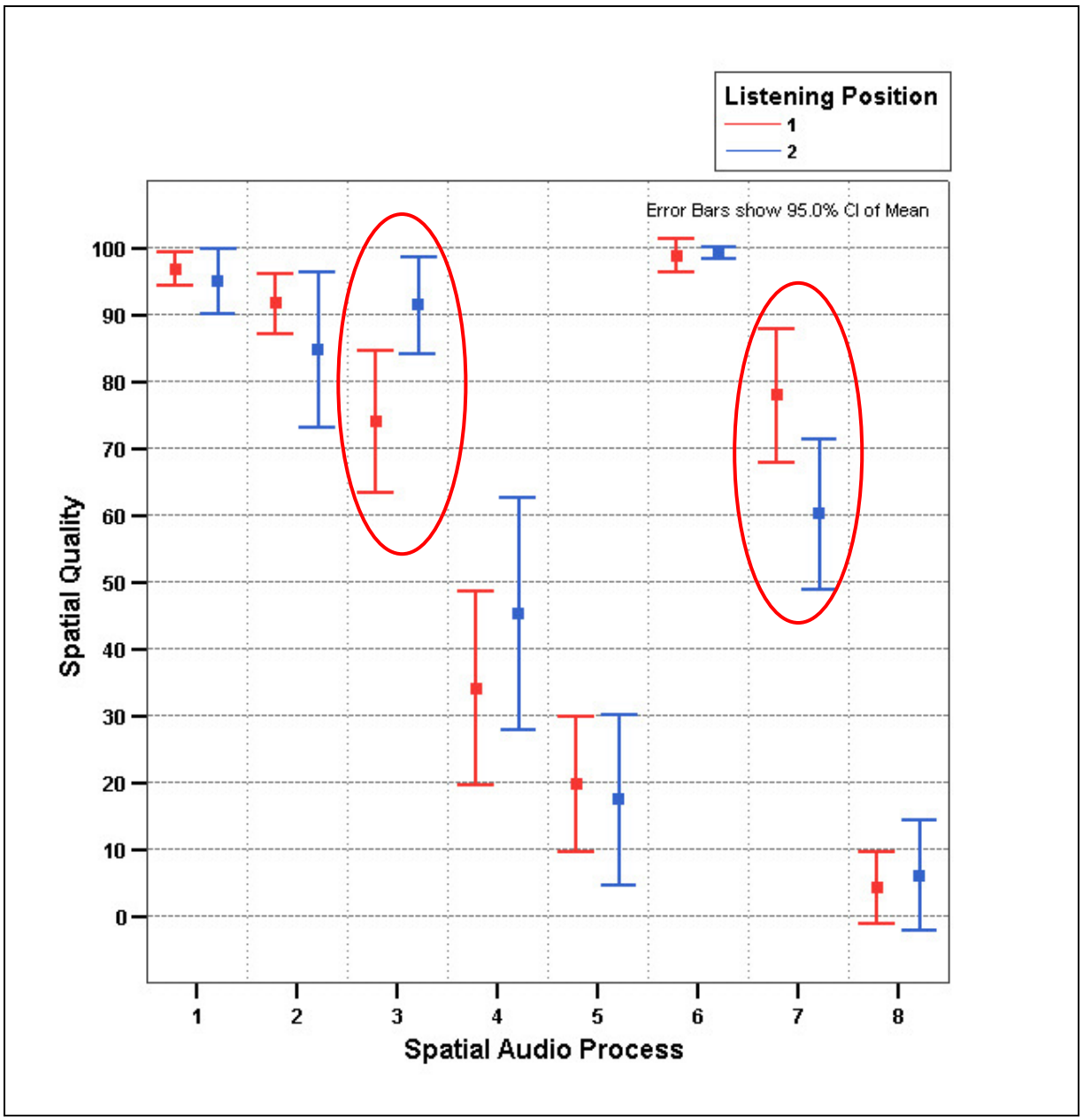


Fig D3. SAPs (circled in red) which create a difference in perceived spatial quality between listening positions with programme item 3 in pilot study 1.

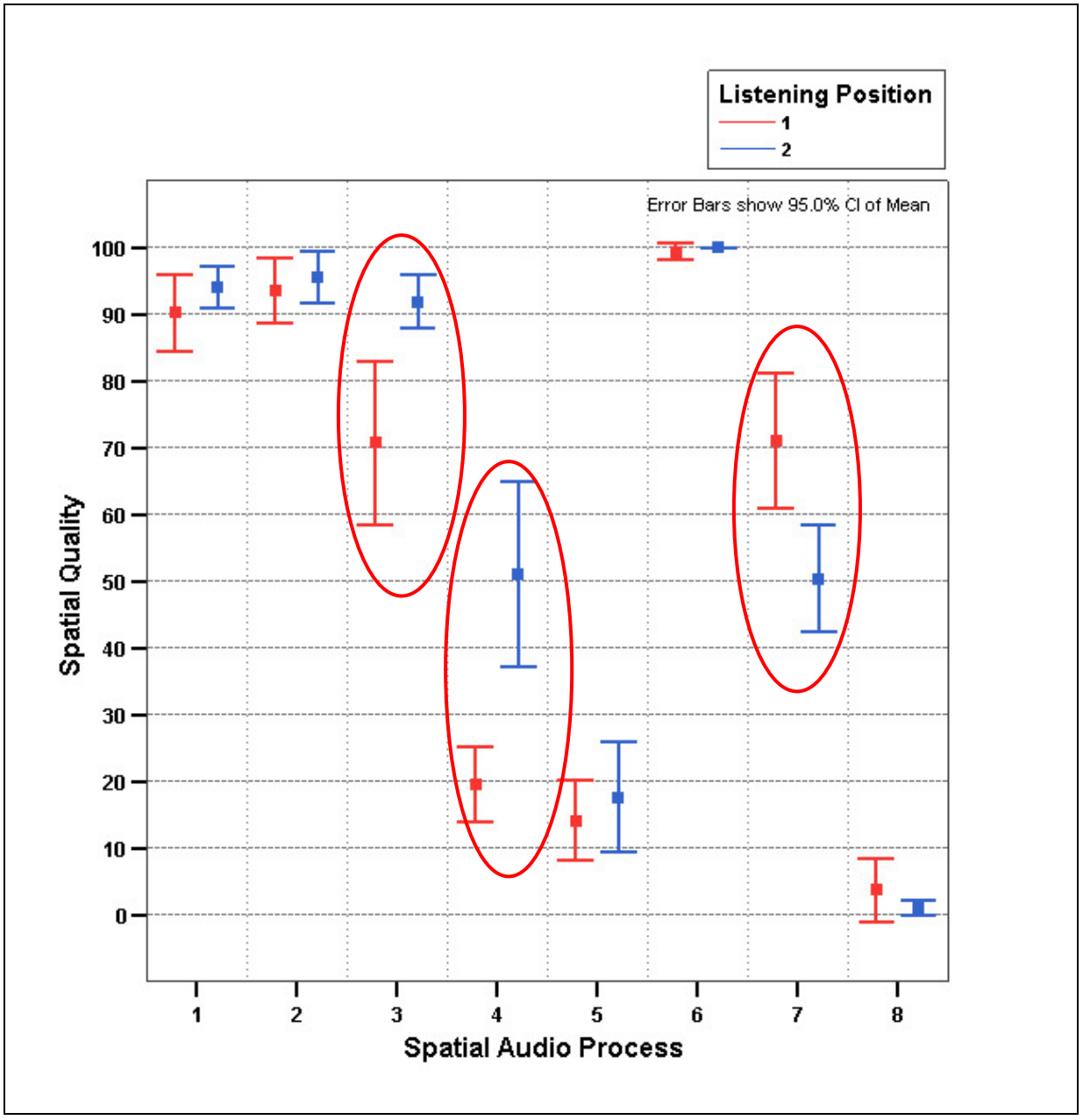


Fig D4. SAPs (circled in red) which create a difference in perceived spatial quality between listening positions with programme item 4 in pilot study 1.

D.2 Listening test 1

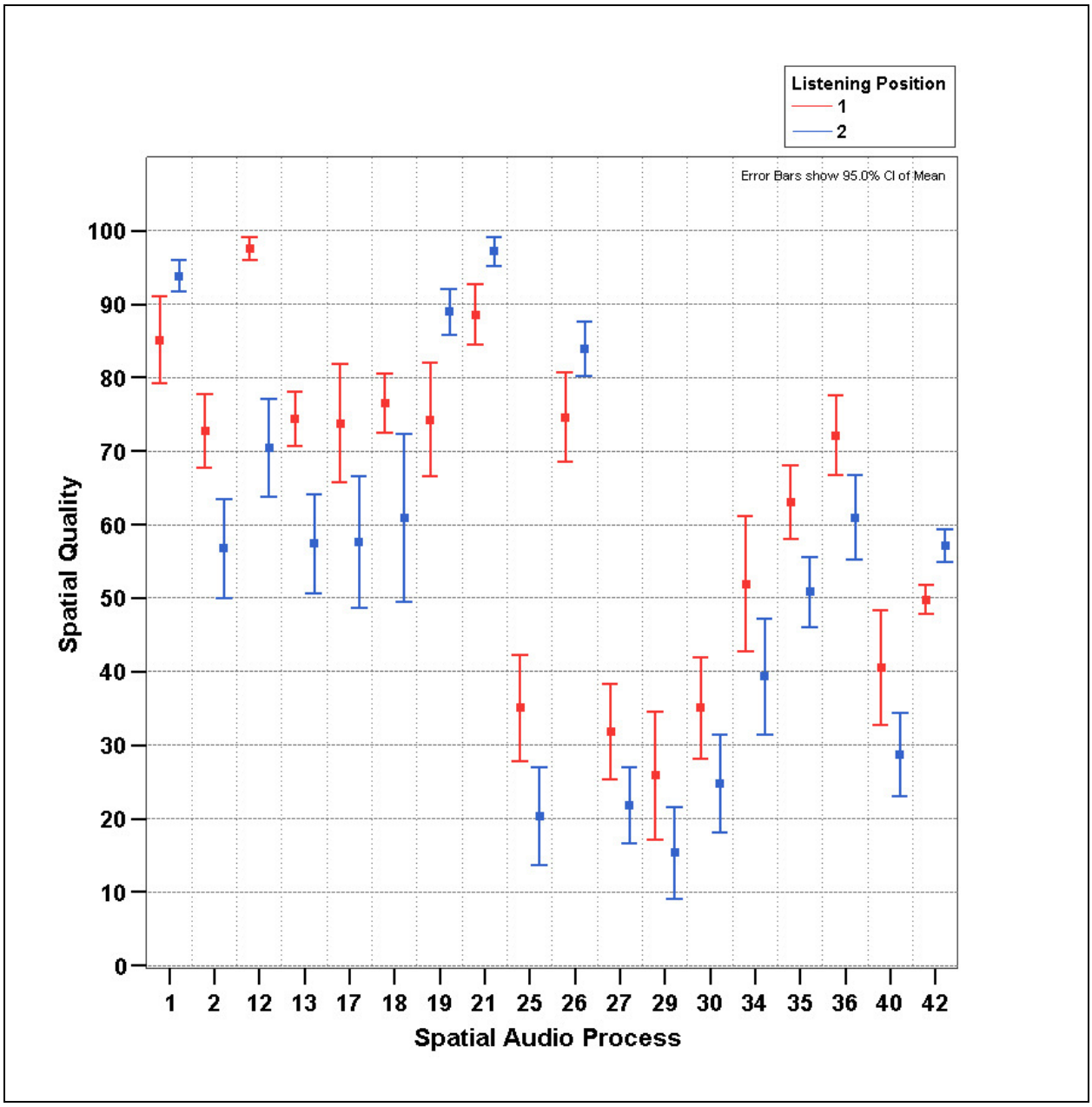


Fig D5. SAPs which create a difference in perceived spatial quality between listening positions with programme item 1 in listening test 1.

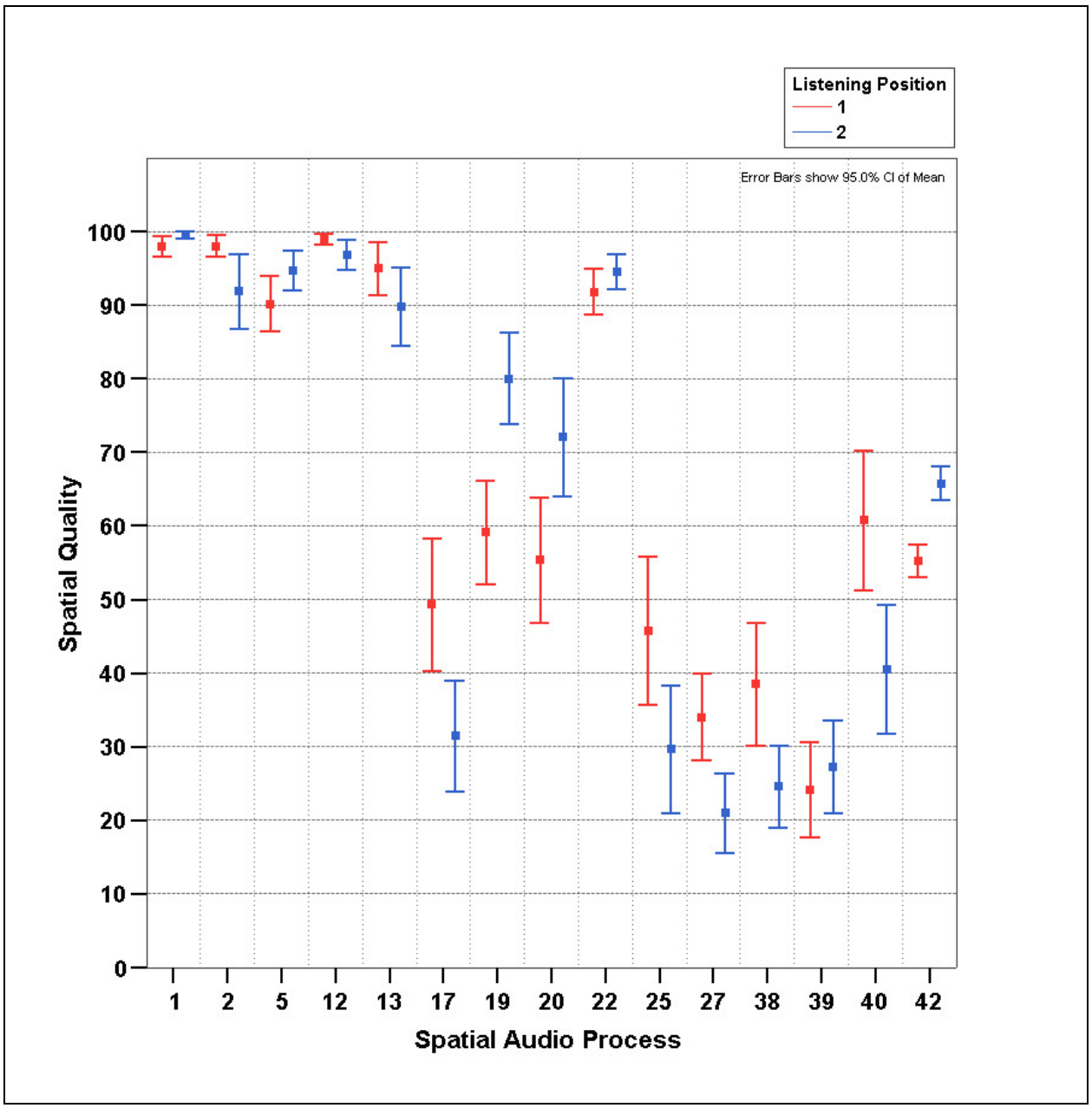


Fig D6. SAPs which create a difference in perceived spatial quality between listening positions with programme item 2 in listening test 1.

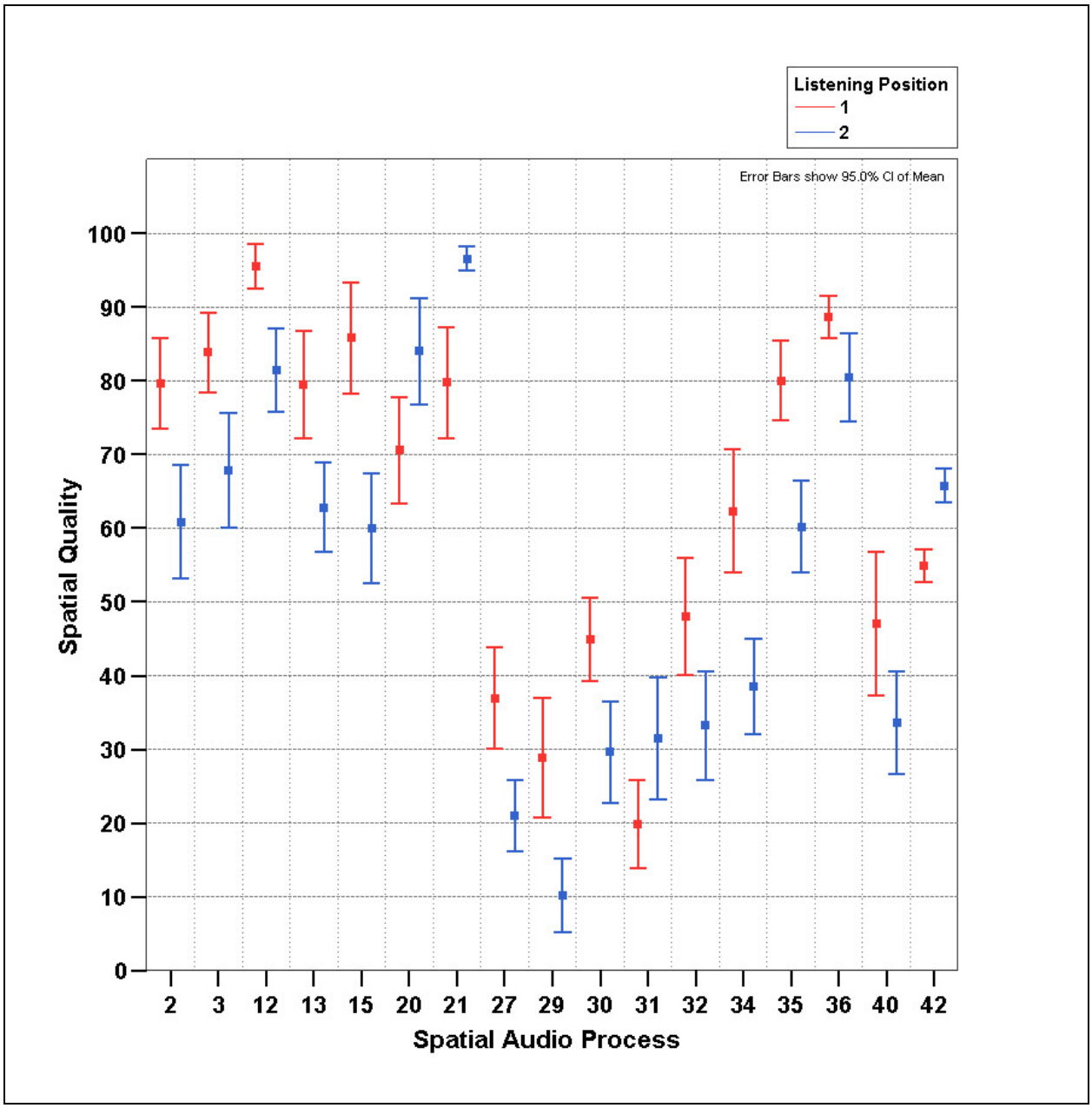


Fig D7. SAPs which create a difference in perceived spatial quality between listening positions with programme item 3 in listening test 1.

Appendix E – Means and 95% confidence intervals for SAPs whose subjective scores were influenced by programme item type in pilot study 1 and 2 and listening test 1 and 2

E.1 Pilot study 1

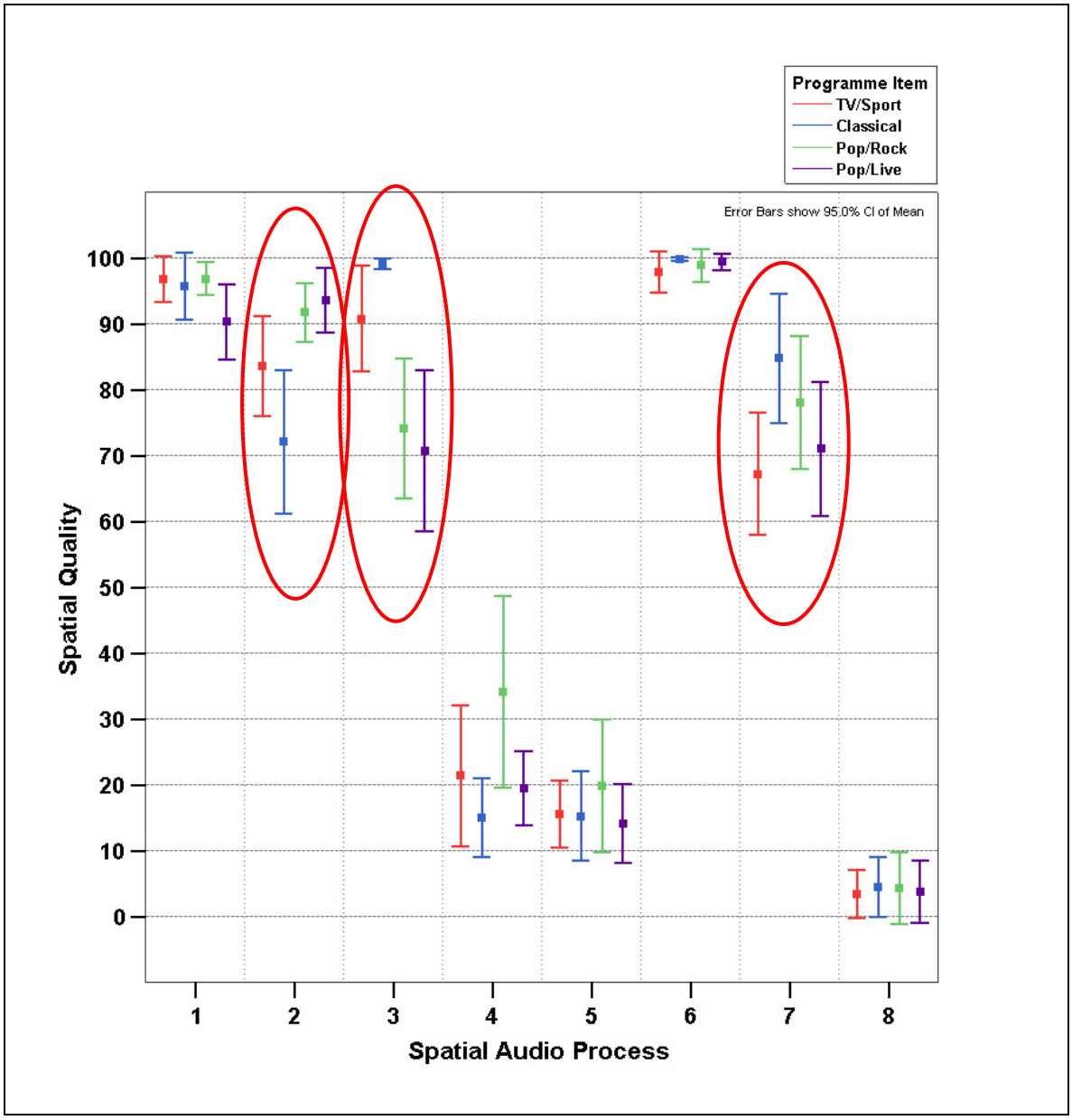


Fig E1. SAPs (circled in red) which create a difference in perceived spatial quality between programme item types at listening position 1 in pilot study 1.

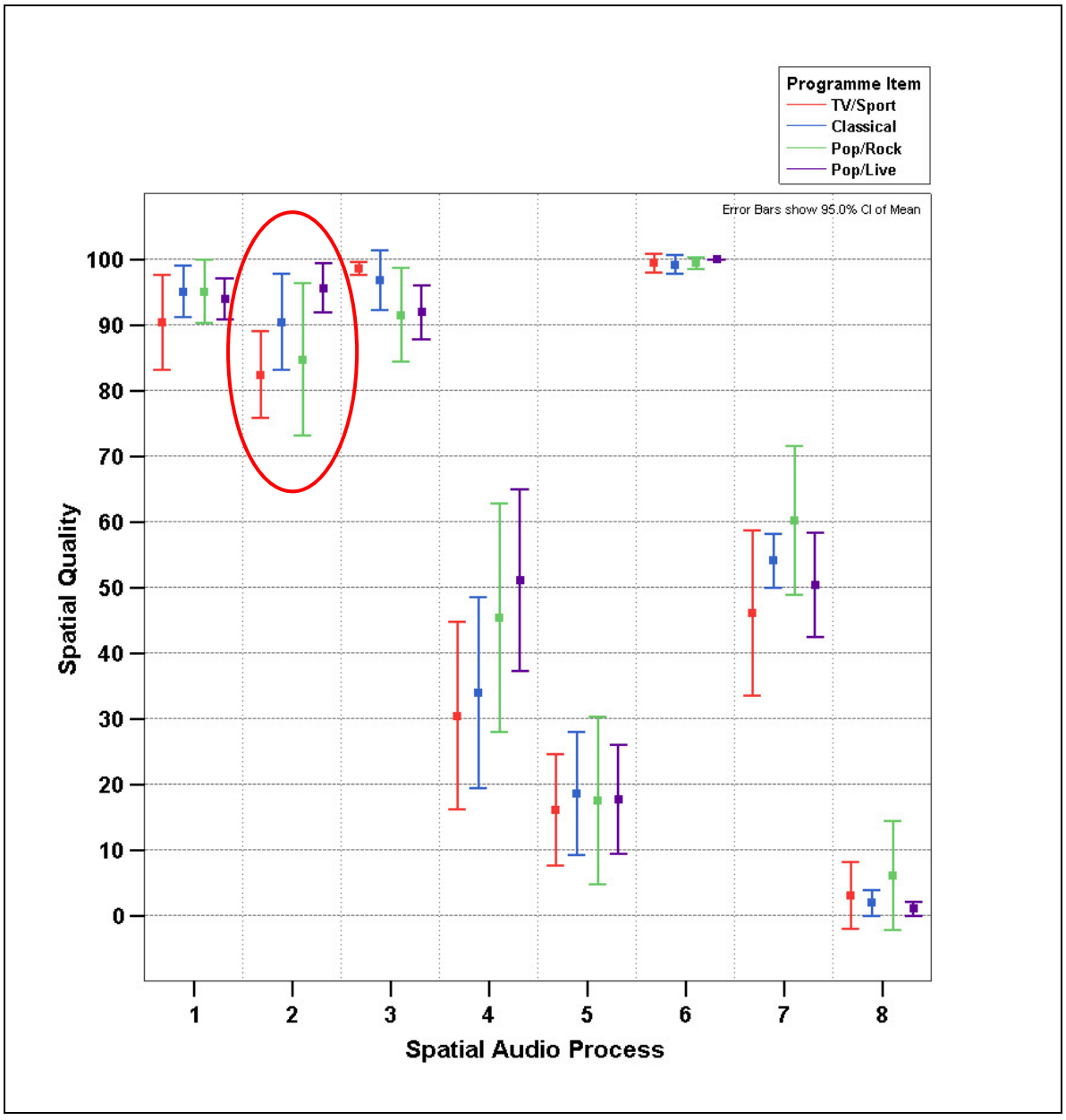


Fig E2. SAPs (circled in red) which create a difference in perceived spatial quality between programme item types at listening position 2 in pilot study 1.

E.2 Pilot study 2

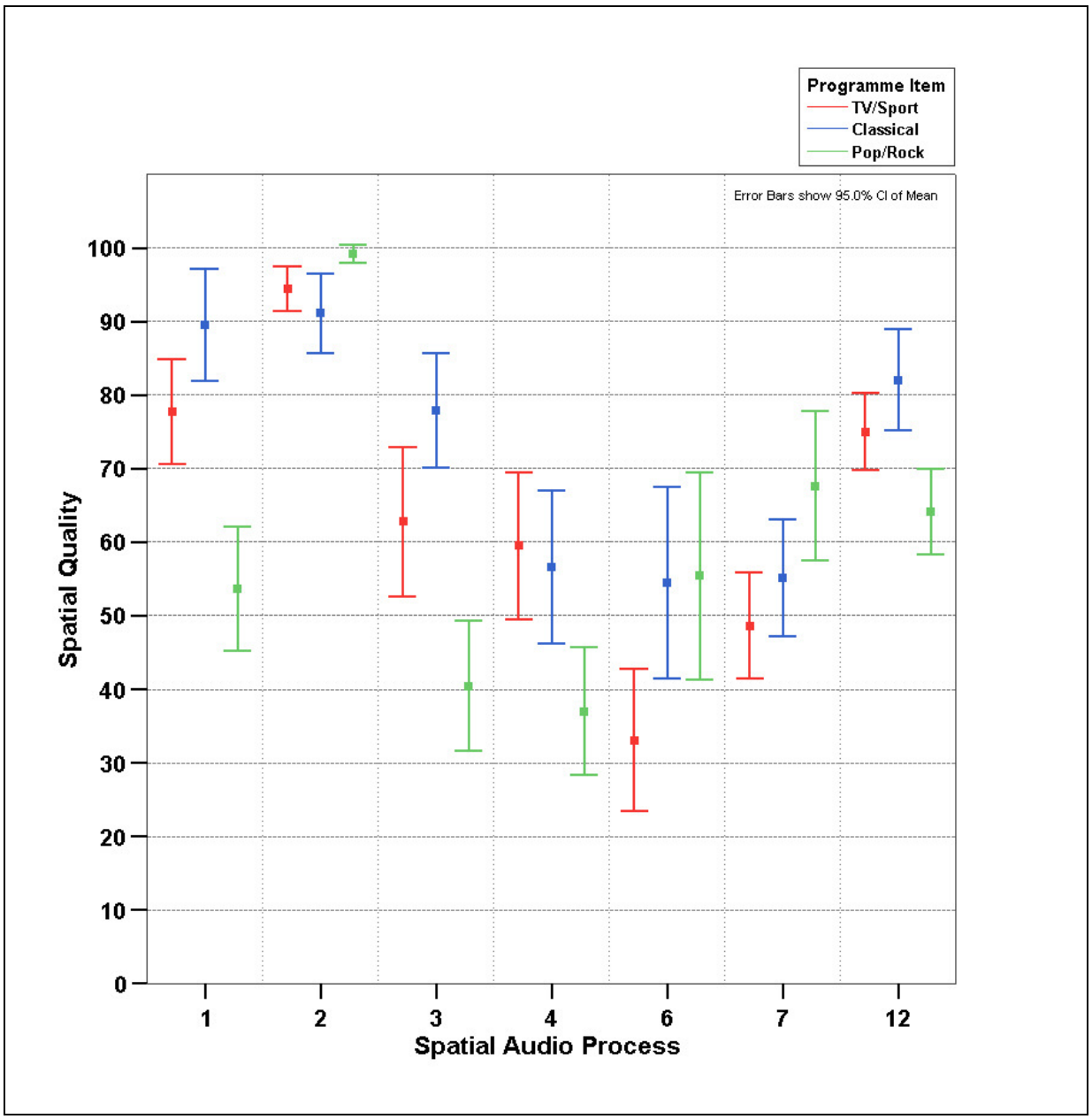


Fig E3. SAPs which create a difference in perceived spatial quality between programme item types in pilot study 2.

E.3 Listening test 1

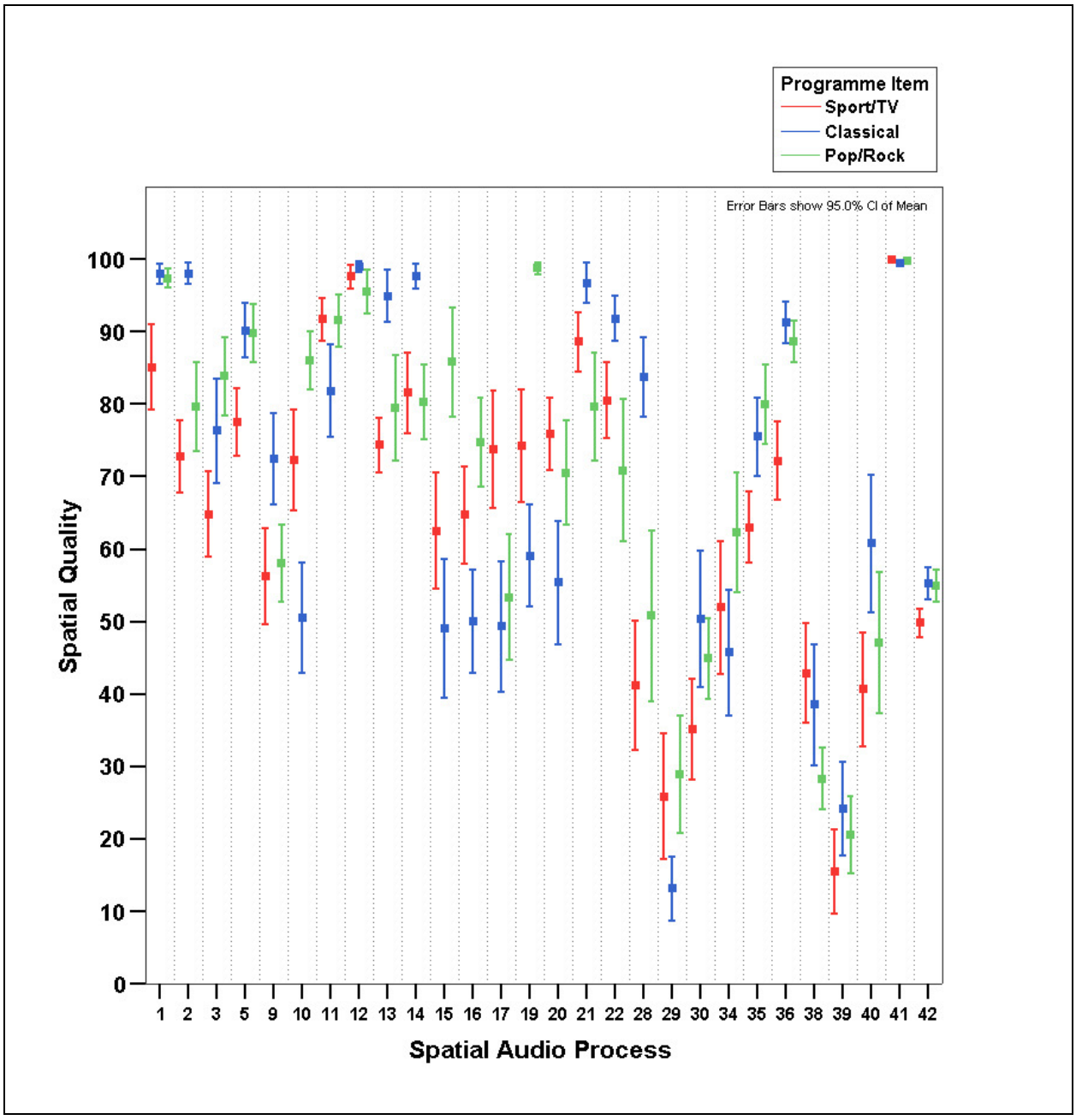


Fig E4. SAPs which create a difference in perceived spatial quality between programme item types at listening position 1 in listening test 1.

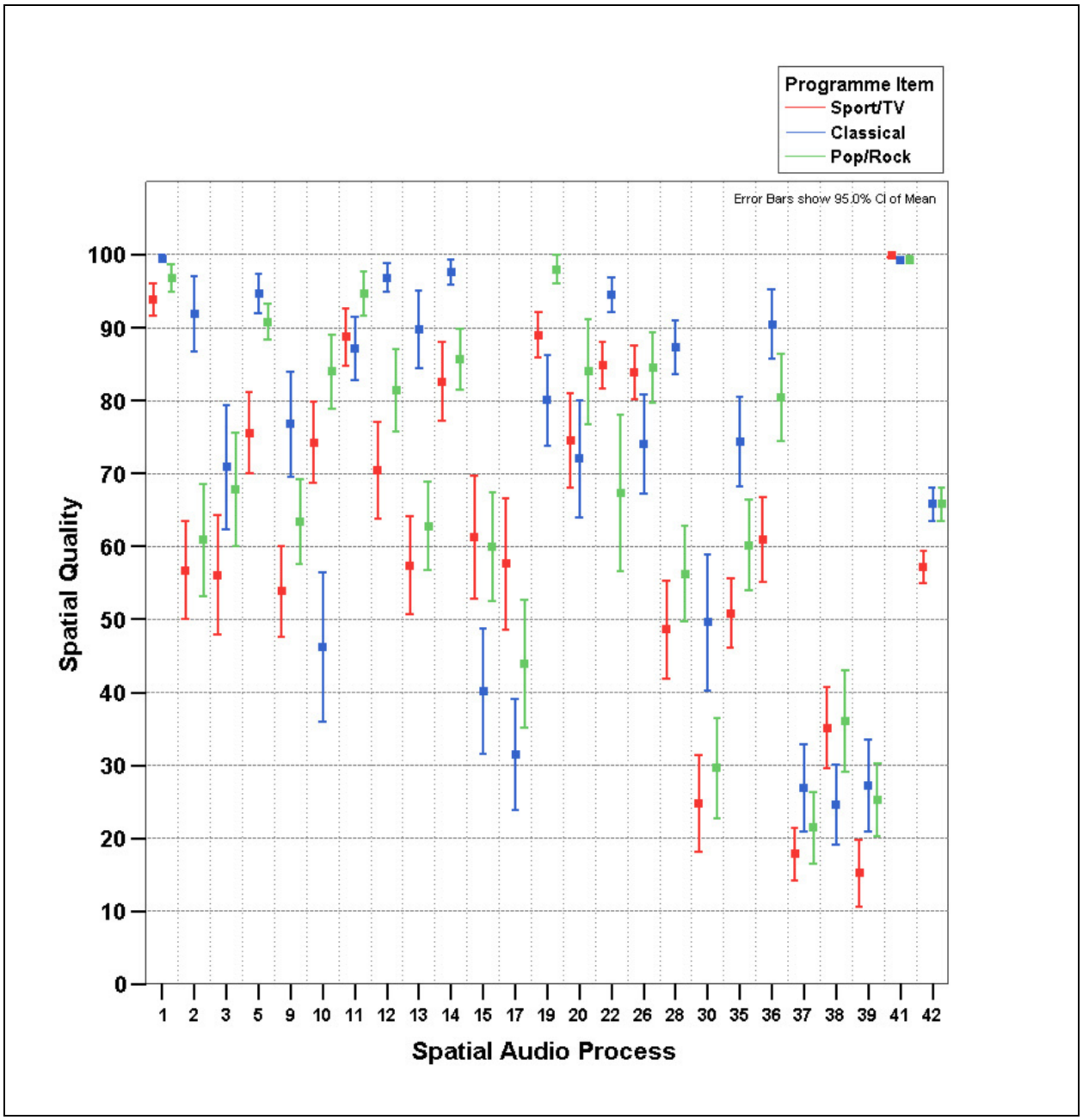


Fig E5. SAPs which create a difference in perceived spatial quality between programme item types at listening position 2 in listening test 1.

E.4 Listening test 2

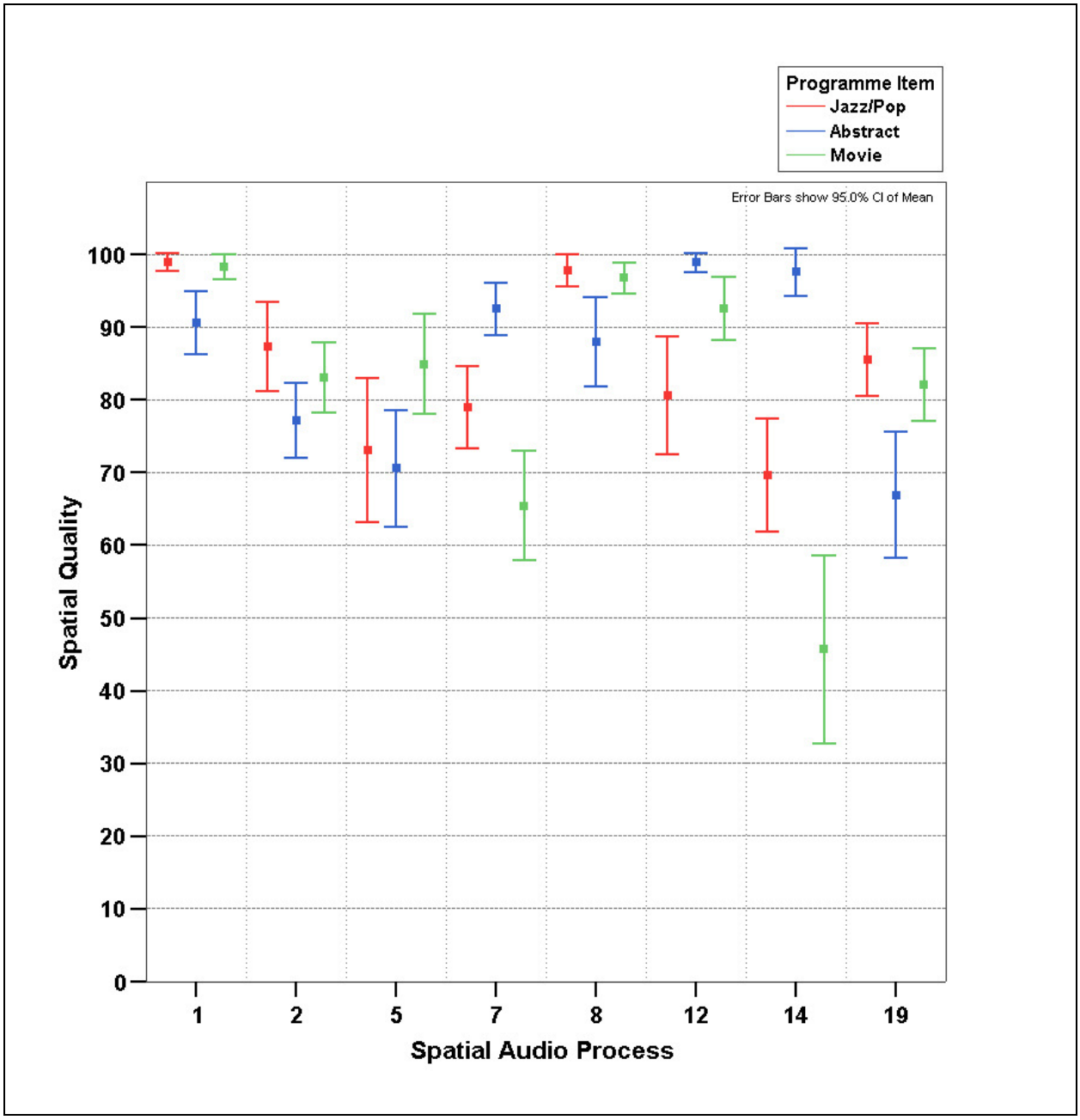


Fig E6. SAPs which create a difference in perceived spatial quality between programme item types at listening position 1 in listening test 2.

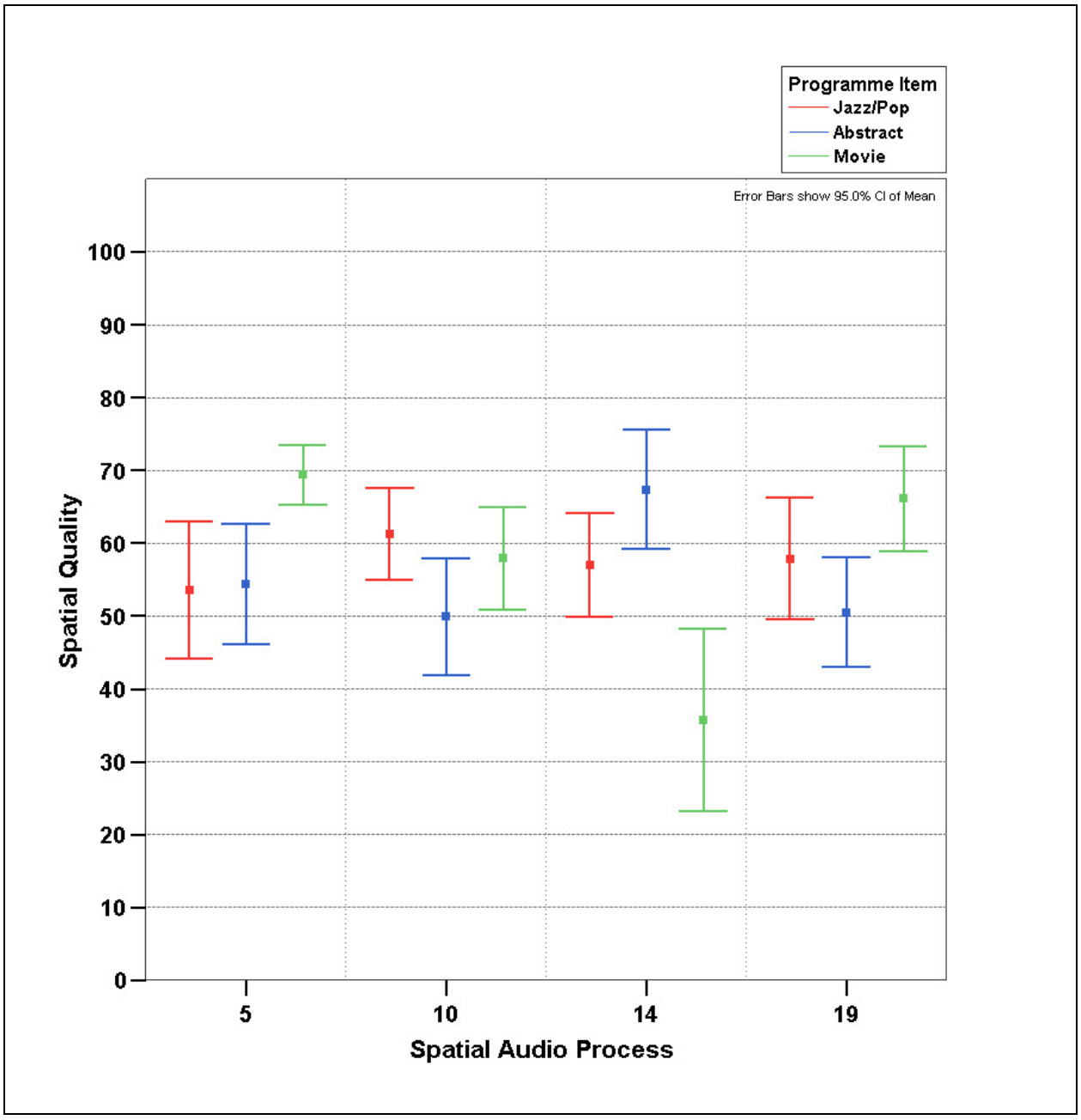


Fig E7. SAPs which create a difference in perceived spatial quality between programme item types at listening position 2 in listening test 2.

Appendix F - Results of spatial attribute analysis for SAPs used in listening tests 1 and 2

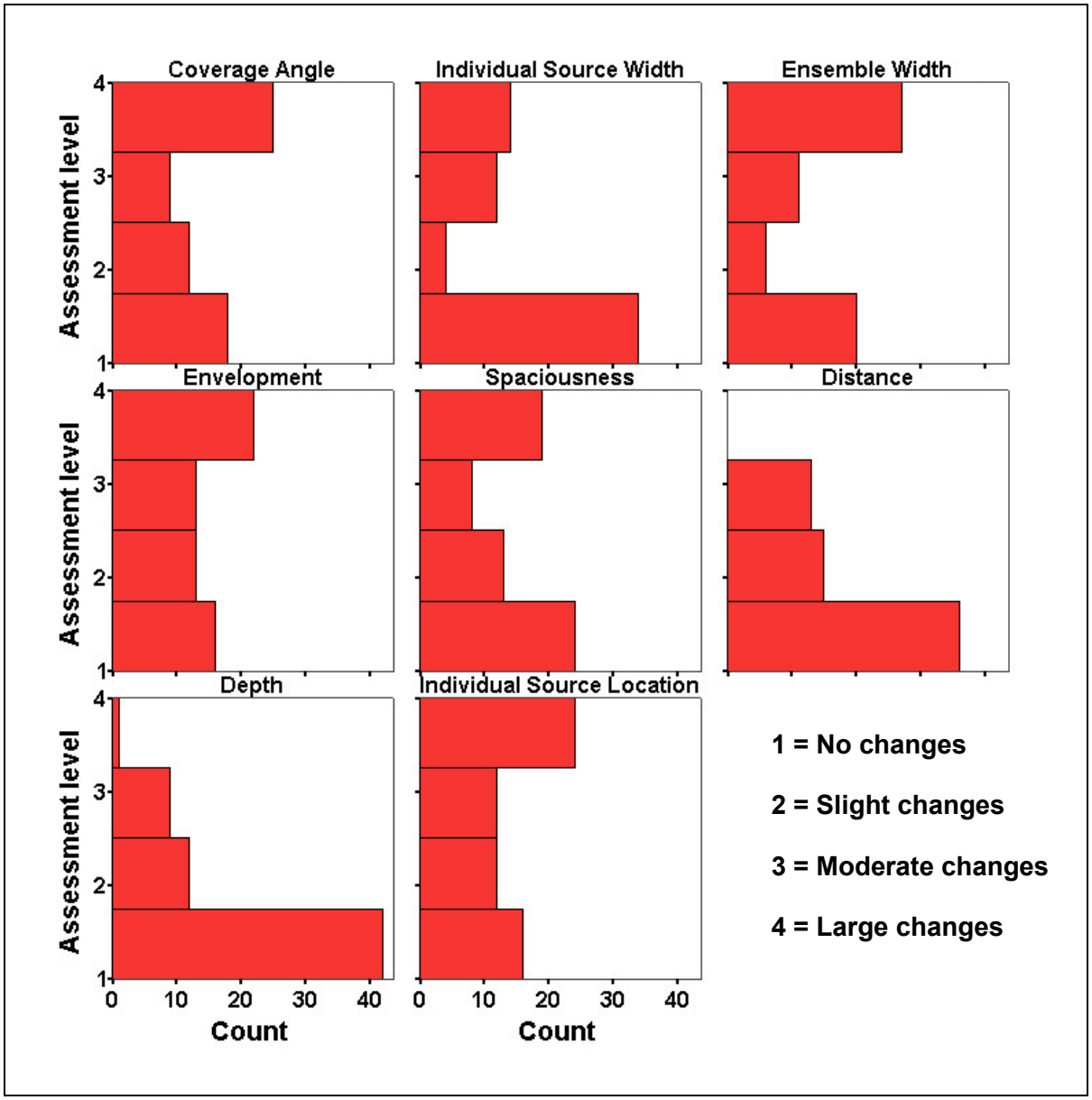


Fig F1. Histograms illustrating the numbers of large, moderate, slight and imperceptible impairments to each of 8 lower level spatial attributes reported in tests using the programme items and SAPs of listening tests 1 and 2

Appendix G - List of spatial audio processes evaluated in listening tests 1 and 2

G.1 All spatial audio processes

No.	Spatial audio process	Description	Group
1	Downmixing from 5CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707l*Rs.	1
2	Downmixing from 5CH 2	3.0: L = L + 0.7071*Ls, R = R + 0.7071*Rs, C = C.	
3	Downmixing from 5CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707l*Rs.	
4	Downmixing from 5CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	
5	Multichannel audio coding 1	160kbs	2
6	Multichannel audio coding 2	64kbs	
7	Multichannel audio coding 3	64kbs	
8	Multichannel audio coding 4	2 stage cascade (80kbs)	
9	Multichannel audio coding 5	4 stage cascade (64kbs)	
10	Altered loudspeaker locations 1	L and R re-positioned at -10° and 10°	3
11	Altered loudspeaker locations 2	C is skewed; re-positioned at 20°	
12	Altered loudspeaker locations 3	Ls and Rs re-positioned at -90° and 90°	
13	Altered loudspeaker locations 4	Ls and Rs re-positioned at -170° and 160°	
14	Altered loudspeaker locations 5	L and C moved 1m to right and not facing listening position	
15	Altered loudspeaker locations 6	Ls moved 1m to right and not facing listening position	
16	Channel rearrangement 1	L and R swapped	4
17	Channel rearrangement 2	L and R swapped for Ls and Rs	
18	Channel rearrangement 3	CH order rotated	
19	Channel rearrangement 4	CH order randomised	
20	Inter-channel level mis-alignment 1	L, C and R -6dB quieter than Ls and Rs	5
21	Inter-channel level mis-alignment 2	Surrounds -6dB	
22	Inter-channel out-of-phase 1	C 180° out-of-phase	6
23	Inter-channel out-of-phase 2	LCR 180° out-of-phase	
24	Channel removal 1	R removed	7
25	Channel removal 2	Ls removed	
26	Channel removal 3	C removed	
27	Spectral filtering 1	500Hz HPF on all channels	8
28	Spectral filtering 2	3.5kHz LPF on all channels	
29	Inter-channel crosstalk 1	1.0 downmix in all CH	9
30	Inter-channel crosstalk 2	Partly correlated (0.5 bleed in adjacent channels)	
31	Virtual surround algorithms 1	Line array virtual surround	10
32	Virtual surround algorithms 2	2 CH virtual surround	
33	Combination 1	CH routing error 4 + Missing channel 1, 2 and 3	11
34	Combination 2	Downmix 2 + Missing channel 1	
35	Combination 3	Downmix 3 + CH routing error 4	
36	Combination 4	Downmix 3 + Loudspeaker miss-placement 1	
37	Combination 5	Downmix 4 + Filtering 1	
38	Combination 6	Loudspeaker miss-placement 4 + Loudspeaker miss-placement 1	
39	Combination 7	Codec A + Downmix 3	
40	Combination 8	Codec A + Loudspeaker miss-placement 3	
41	Combination 9	Codec C + Downmix 4	
42	Combination 10	Codec C + CH routing error 4	
43	Combination 11	Virtual surround algorithms 2 + Missing channel 1	
44	Combination 12	Virtual surround algorithms 2 + Loudspeaker miss-placement 1	
45	Combination 13	Codec C + LS misplacement 6	
46	Anchor recording A	High Anchor - Unprocessed reference	12
47	Anchor recording B	Mid Anchor - Audio codec (80kbs)	
48	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	

Table G1 Complete list of spatial audio processes used in listening tests 1 and 2.

G.2 Spatial audio processes used in listening test 1

No.	Spatial audio process	Description	Group
1	Downmixing from 5CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	1
2	Downmixing from 5CH 2	3.0: L = L + 0.7071*Ls, R = R + 0.7071*Rs, C = C.	
3	Downmixing from 5CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	
4	Downmixing from 5CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	
5	Multichannel audio coding 1	160kbs	2
6	Multichannel audio coding 2	64kbs	
7	Multichannel audio coding 3	64kbs	
8	Multichannel audio coding 4	2 stage cascade (80kbs)	
9	Multichannel audio coding 5	4 stage cascade (64kbs)	
10	Altered loudspeaker locations 1	L and R re-positioned at -10° and 10°	3
11	Altered loudspeaker locations 2	C is skewed; re-positioned at 20°	
12	Altered loudspeaker locations 3	Ls and Rs re-positioned at -90° and 90°	
13	Altered loudspeaker locations 4	Ls and Rs re-positioned at -170° and 160°	
14	Channel rearrangement 1	L and R swapped	4
15	Channel rearrangement 2	L and R swapped for Ls and Rs	
16	Channel rearrangement 3	CH order rotated	
17	Channel rearrangement 4	CH order randomised	
18	Inter-channel level mis-alignment 1	L, C and R -6dB quieter than Ls and Rs	5
19	Inter-channel out-of-phase 1	C 180° out-of-phase	6
20	Channel removal 1	R removed	7
21	Channel removal 2	Ls removed	
22	Channel removal 3	C removed	
23	Spectral filtering 1	500Hz HPF on all channels	8
24	Spectral 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	
41	Anchor recording A	High Anchor - Unprocessed reference	12
42	Anchor recording B	Mid Anchor - Audio codec (80kbs)	
43	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	

Table G2 List of spatial audio processes used in listening test 1.

G.3 Spatial audio processes used in listening test 2

No.	Spatial audio process	Description	Group
1	Down-mixing from 5 CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	1
2	Down-mixing from 5 CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	
3	Down-mixing from 5 CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	
4	Multichannel audio coding 1	160kbs	2
5	Multichannel audio coding 2	64kbs	
6	Multichannel audio coding 3	64kbs	
7	Altered loudspeaker locations 5	L and C moved 1m to left and not facing listening position	3
8	Altered loudspeaker locations 6	Ls moved 1m to left and not facing listening position	
9	Channel rearrangements 1	L and R swapped	4
10	Inter-channel level mis-alignment 1	LCR -6dB	5
11	Inter-channel level mis-alignment 2	Surrounds -6dB	
12	Inter-channel out-of-phase 1	C 180° out-of-phase	6
13	Inter-channel out-of-phase 2	LCR 180° out-of-phase	
14	Channel removal 3	C removed	7
15	Spectral filtering 1	500Hz HPF on all channels	8
16	Spectral filtering 2	3.5kHz LPF on all channels (BS.1534)	8
17	Inter-channel crosstalk 1	1.0 Downmix in all CH	9
18	Combination 5	Down-mixing from 5 CH 4 + Spectral filtering 1	11
19	Combination 7	Multichannel audio coding 1 + Down-mixing from 5 CH 3	
20	Combination 13	Multichannel audio coding 3 + Altered loudspeaker locations 5	
21	Anchor recording A	High Anchor - Unprocessed reference	12
22	Anchor recording B	Mid Anchor - Audio codec (80kbs)	
23	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	
24	Down-mixing from 5 CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	1
25	Down-mixing from 5 CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	
26	Down-mixing from 5 CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	
27	Multichannel audio coding 1	160kbs	2
28	Multichannel audio coding 2	64kbs	
29	Multichannel audio coding 3	64kbs	
30	Channel rearrangements 1	L and R swapped	4
31	Inter-channel level mis-alignment 1	LCR -6dB	5
32	Inter-channel level mis-alignment 2	Surrounds -6dB	
33	Inter-channel out-of-phase 1	C 180° out-of-phase	6
34	Inter-channel out-of-phase 2	LCR 180° out-of-phase	
35	Channel removal 3	C removed	7
36	Spectral filtering 1	500Hz HPF on all channels	8
37	Spectral filtering 2	3.5kHz LPF on all channels (BS.1534)	
38	Inter-channel crosstalk 1	1.0 downmix in all CH	9
39	Combination 5	1.0 Downmix + Spectral filter 1	11
40	Combination 7	Codec A + 2.0 Downmix	
41	Anchor recording A	High Anchor - Unprocessed reference	12
42	Anchor recording B	Mid Anchor - Audio codec (80kbs)	
43	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	

Table G3 List of spatial audio processes used in listening test 2.

G.4 Division of spatial audio processes for each session of listening test 1

No.	Spatial audio process	Description	Group
3	Downmixing from 5CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	1
4	Downmixing from 5CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	
11	Altered loudspeaker locations 2	C is skewed; re-positioned at 20°	3
12	Altered loudspeaker locations 3	Ls and Rs re-positioned at -90° and 90°	
16	Channel rearrangement 3	CH order rotated	4
17	Channel rearrangement 4	CH order randomised	
20	Channel removal 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 G4. SAPs selected for listening test 1 session 1.

No.	Spatial audio process	Description	Group
2	Downmixing from 5CH 2	3.0: L = L + 0.7071*Ls, R = R + 0.7071*Rs, C = C.	1
10	Altered loudspeaker locations 1	L and R re-positioned at -10° and 10°	3
13	Altered loudspeaker locations 4	Ls and Rs re-positioned at -170° and 160°	
15	Channel rearrangement 2	L and R swapped for Ls and Rs	4
18	Inter-channel level misalignment 1	L, C and R -6dB quieter than Ls and Rs	5
21	Channel removal 2	Ls removed	7
22	Channel removal 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 G5. SAPs selected for listening test 1 session 2.

No.	Spatial audio process	Description	Group
1	Downmixing from 5CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	1
8	Multichannel audio coding 4	2 stage cascade (80kbs)	2
9	Multichannel audio coding 5	4 stage cascade (64kbs)	
14	Channel rearrangement 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 G6. SAPs selected for listening test 1 session 3.

No.	Spatial audio process	Description	Group
5	Multichannel audio coding 1	160kbs	2
6	Multichannel audio coding 2	64kbs	
7	Multichannel audio coding 3	64kbs	
19	Inter-channel out-of-phase 1	C 180° out-of-phase	6
23	Spectral filtering 1	500Hz HPF on all channels	8
24	Spectral 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 G7. SAPs selected for listening test 1 session 4.

G.5 Division of spatial audio processes for each session of listening test 2

No.	Spatial audio process	Description	Group
1	Down-mixing from 5 CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	1
3	Down-mixing from 5 CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	
6	Multichannel audio coding 3	64kbs	2
7	Altered loudspeaker locations 5	L and C moved 1m to left and not facing listening position	3
8	Altered loudspeaker locations 6	Ls moved 1m to left and not facing listening position	
17	Inter-channel crosstalk 1	1.0 downmix in all CH	9
25	Down-mixing from 5 CH 3	2.0: L=L+0.7071*C+0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	1
26	Down-mixing from 5 CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	
27	Multichannel audio coding 1	160kbs	2
41	Anchor recording A	High Anchor - Unprocessed reference	12

Table G8. SAPs selected for listening test 2 session 1.

No.	Spatial audio process	Description	Group
2	Down-mixing from 5 CH 3	2.0: L=L+0.7071*C+0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	1
5	Multichannel audio coding 2	64kbs	2
10	Inter-channel level misalignment 1	LCR -6dB	5
16	Spectral filtering 2	3.5kHz LPF on all channels (BS.1534)	8
20	Combination 13	Codec C + LS misplacement 6	11
24	Down-mixing from 5 CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	1
28	Multichannel audio coding 2	64kbs	2
30	Channel rearrangements 1	L and R swapped	4
38	Inter-channel crosstalk 1	1.0 downmix in all CH	9
43	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	12

Table G9. SAPs selected for listening test 2 session 2.

No.	Spatial audio process	Description	Group
4	Multichannel audio coding 1	160kbs	2
9	Channel rearrangements 1	L and R swapped	4
12	Inter-channel level out-of-phase 1	C 180° out-of-phase	6
14	Channel removal 3	C removed	7
15	Spectral filtering 1	500Hz HPF on all channels	8
31	Inter-channel level misalignment 1	LCR -6dB	5
34	Inter-channel level out-of-phase 2	LCR 180° out-of-phase	6
37	Spectral filtering 2	3.5kHz LPF on all channels (BS.1534)	8
39	Combination 5	Down-mixing from 5 CH 4 + Spectral filtering 1	11
40	Combination 7	Multichannel audio coding 1 + Down-mixing from 5CH 3	11

Table G10. SAPs selected for listening test 2 session 3.

No.	Spatial audio process	Description	Group
11	Inter-channel level misalignment 2	Surrounds -6dB	5
13	Inter-channel level out-of-phase 2	LCR 180° out-of-phase	6
18	Combination 5	Down-mixing from 5 CH 4 + Spectral filtering 1	11
19	Combination 7	Multichannel audio coding 1 + Down-mixing from 5CH 3	11
29	Multichannel audio coding 3	64kbs	2
32	Inter-channel level misalignment 2	Surrounds -6dB	5
33	Inter-channel level out-of-phase 1	C 180° out-of-phase	6
35	Channel removal 3	C removed	7
36	Spectral filter 1	500Hz HPF on all channels	8
42	Anchor recording B	Mid Anchor - Audio codec (80kbs)	12

Table G11. SAPs selected for listening test 2 session 4.

Appendix H - Flowchart illustrating a listeners path through sessions 1 and 2 for listening test 1

This path was repeated for sessions 3 and 4.

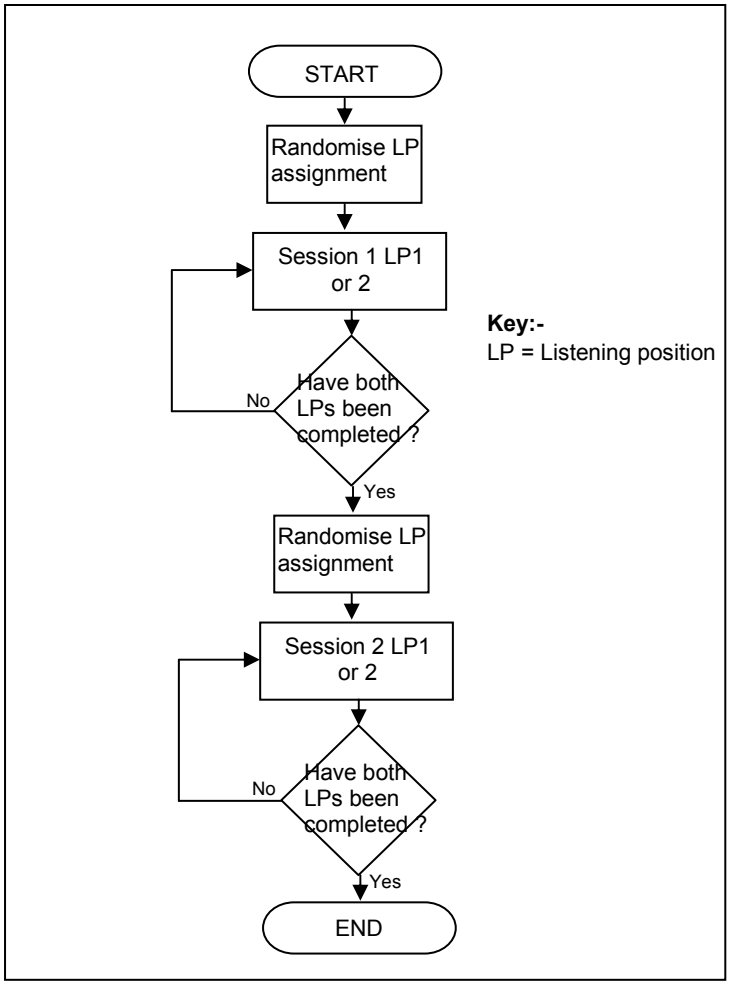


Fig H1. Flowchart illustrating a listener's path through sessions 1 and 2 of listening test 1.

Appendix I - Assessment of listener performance in listening tests 1 and 2

Each listener's responses were assessed, so that the most reliable could be selected for further analysis and investigation. Two methods of assessment were used:

I.1 Discrimination ability

The discrimination ability of each listener was determined by conducting a one-sampled t-test on their scores for 'Anchor recording A' (high anchor – unprocessed reference). A one-sampled t-test tests whether a mean is statistically significant ($p < 0.05$) different from a specified value. If a listener was capable of identifying this stimulus and scoring it as instructed, they were deemed as having suitable discrimination ability.

I.2 Consistency

The consistency of a listener's responses was determined by investigating the magnitude of their error in repeat judgements. Root mean square error was calculated between repeated assessments of stimuli. To pass this test a listener's RMS error must not be greater than 15% (based on a 100 point test scale). Although smaller values of RMS error such as 10% have been considered as acceptable in similar experiments [Rumsey, 1998] a higher threshold was chosen due to the difficulty of the task. (NB. The anchor recordings are assessed many more times than the other stimuli so to balance the assessment they are removed).

Figures I.1 – I.12 illustrate the results of these assessments for listening tests 1 and 2. Tables I.1 and I.2 summarise these results.

I.3 Listening test 1

Listener	Mean score	p
1	100	N/A
2	100	N/A
3	100	N/A
4	100	N/A
5	100	N/A
6	99.58	0.339
7	99.76	0.339
8	100	N/A
9	100	N/A
10	99.83	0.339
11	100	N/A
12	99.67	0.220
13	100	N/A
14	99.92	0.339

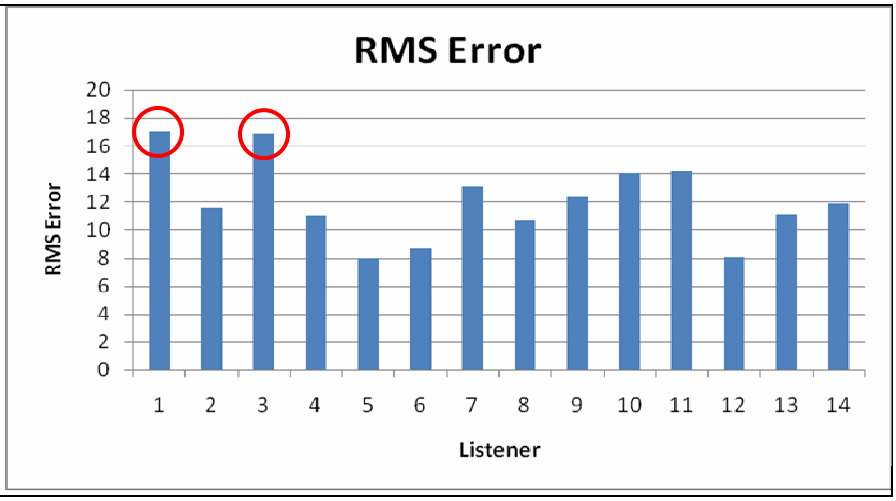


Fig I.1 Listening test 1, Session 1, listening position 1 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
1	99.83	0.339
2	100	N/A
3	98.17	0.293
4	99.83	0.339
5	97.67	0.048
6	98.75	0.191
7	95.33	0.044
8	100	N/A
9	99.58	0.339
10	99.50	0.111
11	100	N/A
12	98.42	0.196
13	100	N/A
14	100	N/A

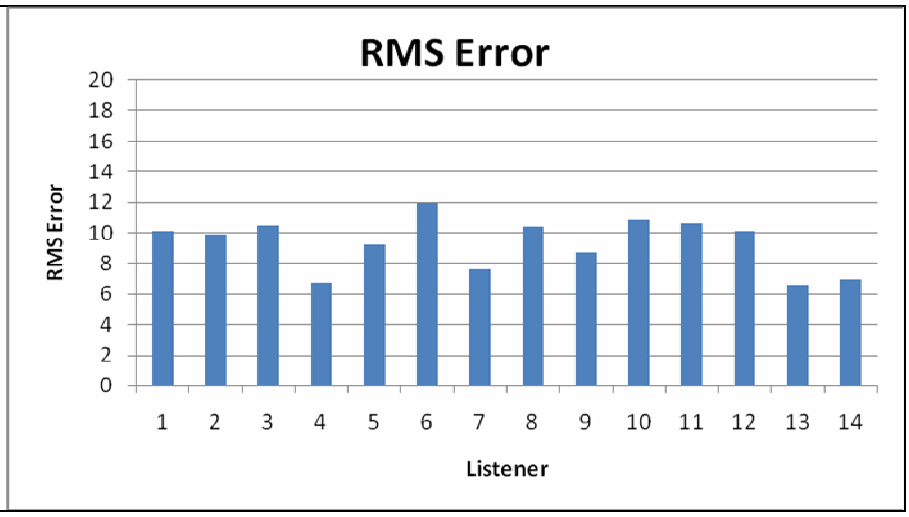


Fig I.2 Listening test 1, Session 2, listening position 1 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
1	100	N/A
2	100	N/A
3	100	N/A
4	100	N/A
5	97.25	N/A
6	100	N/A
7	99.58	0.175
8	99.42	0.206
9	100	N/A
10	99.83	0.339
11	98.83	0.339
12	100	N/A
13	99.67	0.339
14	100	N/A

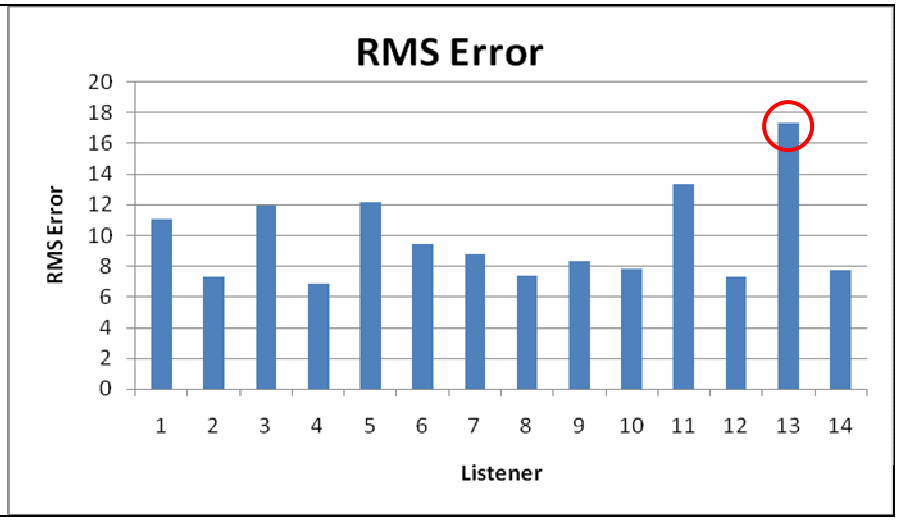


Fig I.3 Listening test 1, Session 3, listening position 1 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
1	99.67	0.339
2	100	N/A
3	99.17	0.339
4	100	N/A
5	99.67	0.339
6	100	N/A
7	100	N/A
8	100	N/A
9	100	N/A
10	100	N/A
11	100	N/A
12	99.58	0.339
13	100	N/A
14	100	N/A

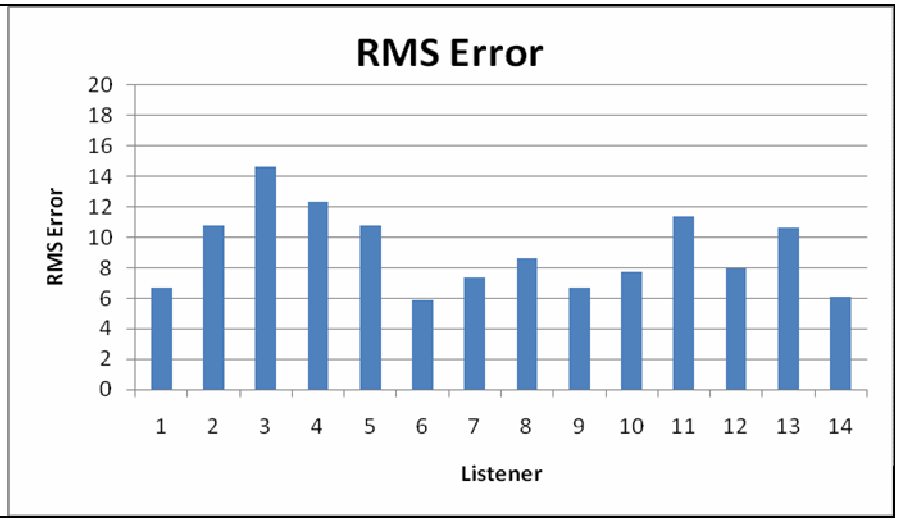


Fig I.4 Listening test 1, Session 4, listening position 1 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
1	99.33	0.054
2	100	N/A
3	100	N/A
4	100	N/A
5	99.08	0.067
6	100	N/A
7	97.42	0.006
8	100	N/A
9	100	N/A
10	99.58	0.339
11	99.50	0.339
12	98.25	0.127
13	100	N/A
14	100	N/A

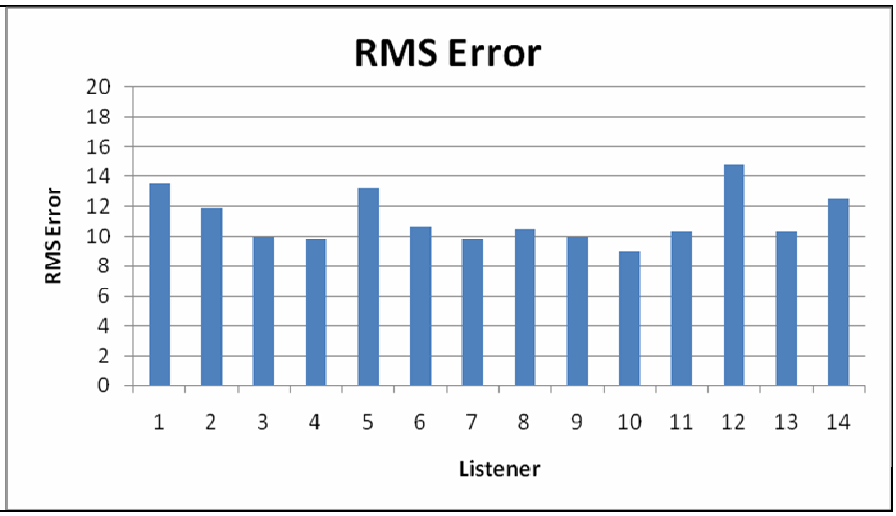


Fig I.5 Listening test 1, Session 1, listening position 2 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
1	98.67	0.108
2	100	N/A
3	99.5	0.339
4	99.83	0.339
5	96.83	0.020
6	100	N/A
7	98.92	0.065
8	100	N/A
9	98.42	0.215
10	98.75	0.339
11	99.58	0.339
12	98.5	0.166
13	100	N/A
14	100	N/A

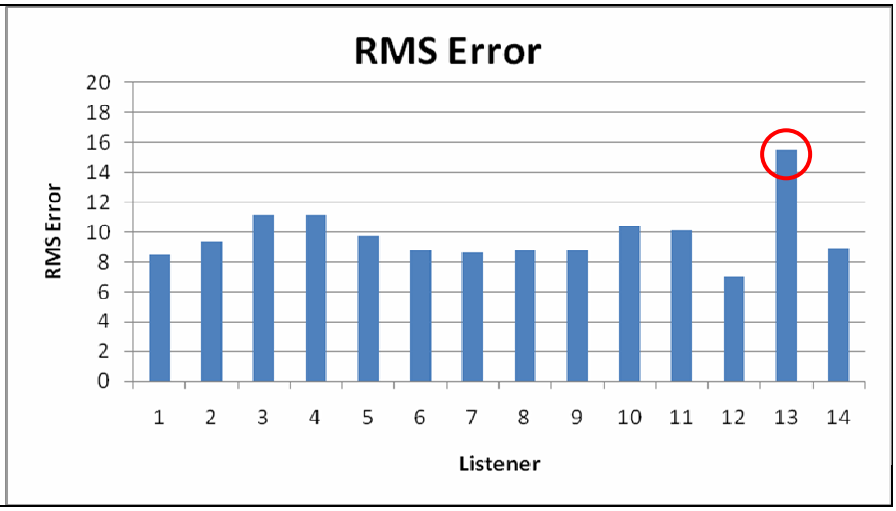


Fig I.6 Listening test 1, Session 2, listening position 2 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
1	99.83	0.339
2	100	N/A
3	100	N/A
4	100	N/A
5	98.25	0.084
6	99.58	0.339
7	99.17	0.201
8	99.17	0.175
9	99.75	0.339
10	99.83	0.339
11	99.67	0.339
12	99.83	0.339
13	99.67	0.339
14	100	N/A

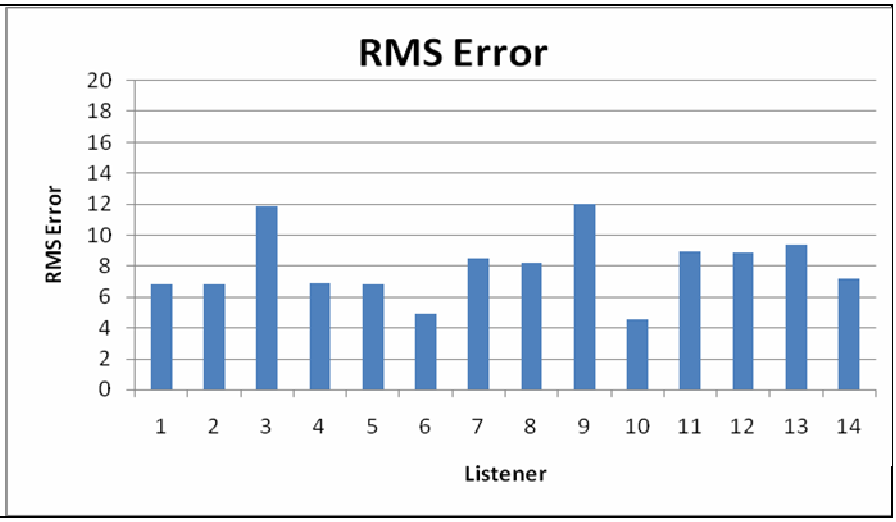


Fig I.7 Listening test 1, Session 3, listening position 2 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
1	99.5	0.339
2	100	N/A
3	98.08	0.152
4	100	N/A
5	99.67	0.339
6	100	N/A
7	96.83	0.185
8	99	0.339
9	100	N/A
10	99.75	0.191
11	100	N/A
12	96.33	0.293
13	100	N/A
14	100	N/A

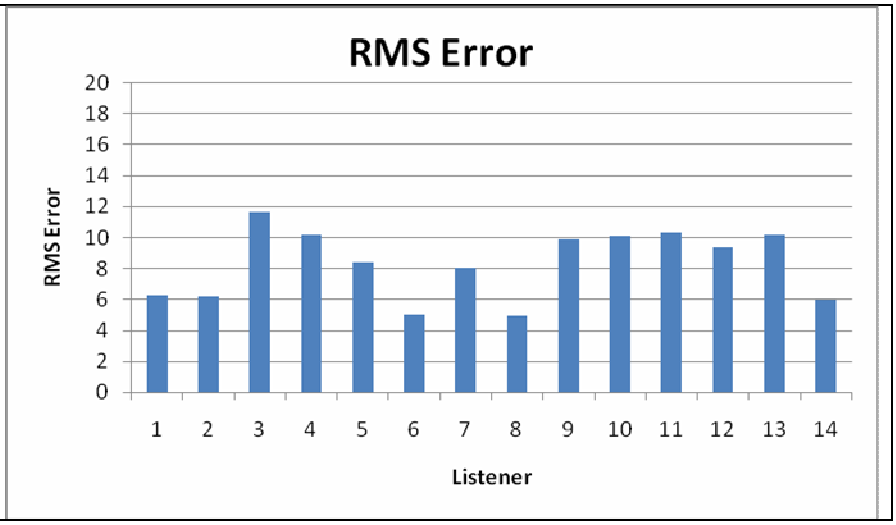


Fig I.8 Listening test 1, Session 4, listening position 2 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

The outcome of this analysis resulted in a number of listeners being removed from the listening test 1 data set (see Table I.1).

Listening position	Session	Listeners whose data was removed
1	1	1, 3
	2	No listeners removed
	3	13
	4	No listeners removed
2	1	No listeners removed
	2	13
	3	No listeners removed
	4	No listeners removed

Table I.1. Listeners removed from the subjective database of listening test 1.

I.4 Listening test 2

Listener	Mean score	p
1	100	N/A
2	100	N/A
3	100	N/A
4	95.17	0.339
5	100	N/A
6	95.50	0.125
7	89.42	0.004
8	100	N/A
9	89.75	0.000
10	99.25	N/A
11	100	N/A
12	100	N/A
13	96.67	0.009
14	95.25	0.013
15	100	N/A
16	90.50	0.000
17	100	N/A

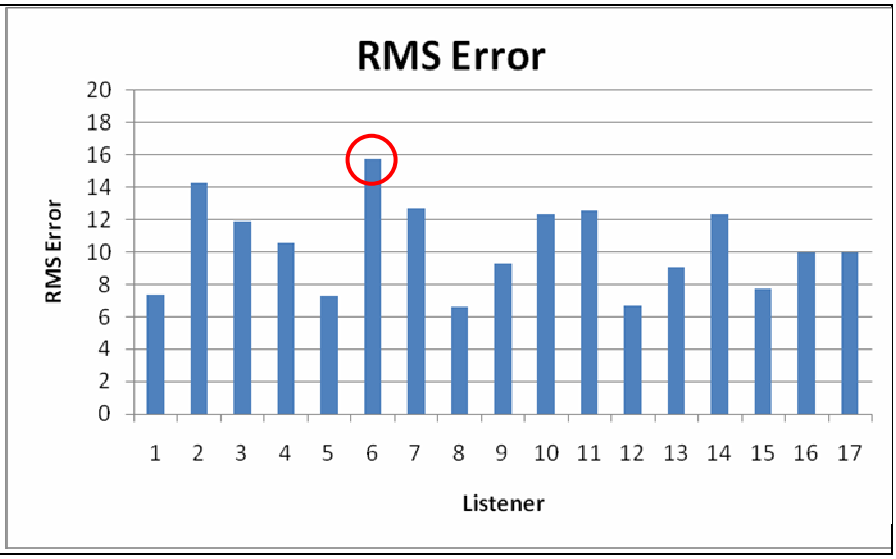


Fig I.9 Listening test 2, Session 1 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
1	100	N/A
2	100	N/A
3	100	N/A
4	100	N/A
5	99.5	0.166
6	99.58	0.339
7	88.58	0.000
8	100	N/A
9	91.92	0.000
10	99.08	0.085
11	98.58	0.309
12	100	N/A
13	99.08	0.042
14	100	N/A
15	99.42	0.111
16	94.92	0.000
17	100	N/A

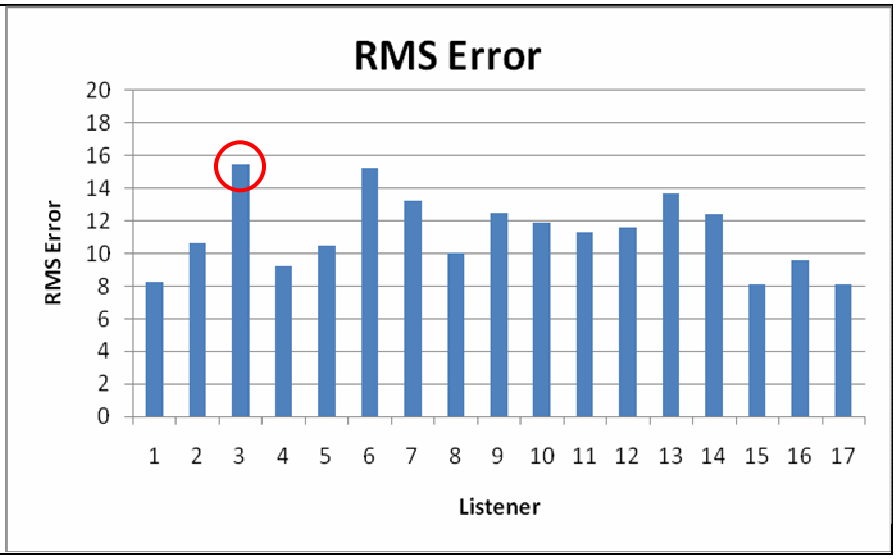


Fig I.10 Listening test 2, Session 2 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
1	99.67	0.104
2	100	N/A
3	99.67	0.339
4	99.25	0.339
5	100	N/A
6	100	N/A
7	91.67	0.000
8	100	N/A
9	86.42	0.000
10	94.83	0.015
11	100	N/A
12	100	N/A
13	98.58	0.006
14	98.25	0.206
15	99.75	0.339
16	94.08	0.000
17	100	N/A

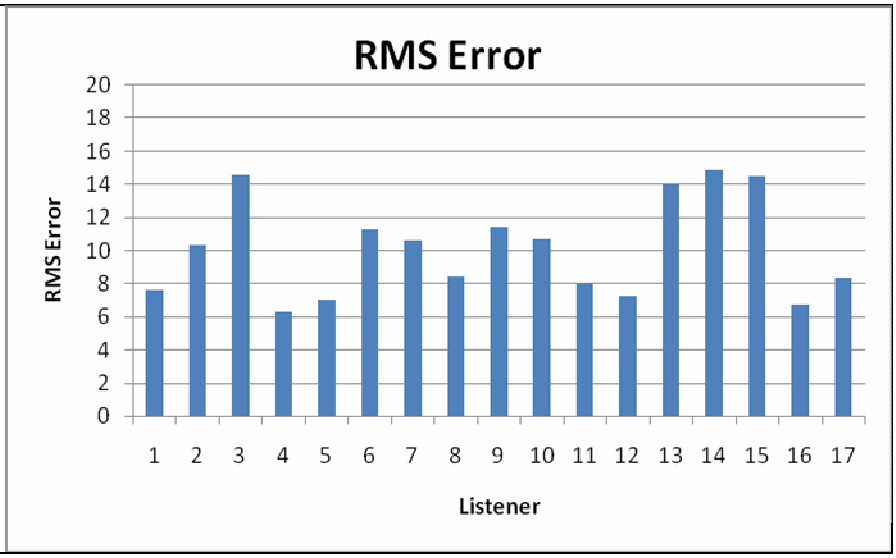


Fig I.11 Listening test 2, Session 3 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

Listener	Mean score	p
1	100	N/A
2	100	N/A
3	100	N/A
4	100	N/A
5	100	N/A
6	100	N/A
7	88.25	0.000
8	100	N/A
9	90.25	0.000
10	99.58	0.339
11	95.58	0.310
12	99.08	0.168
13	99.25	0.021
14	97.75	0.166
15	100	N/A
16	94.5	0.000
17	100	N/A

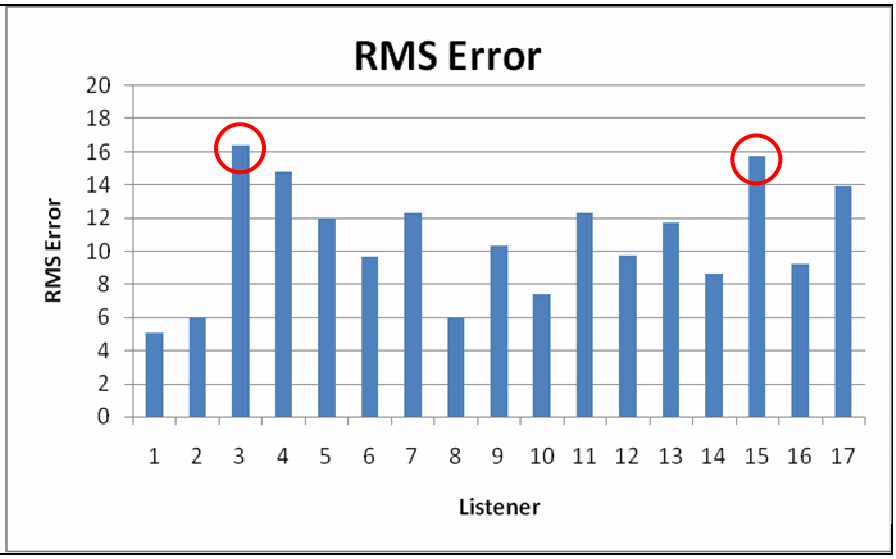


Fig I.12 Listening test 2, Session 4 listener assessment. Left panel: Discrimination – Listener vs. Spatial quality score (for hidden reference), Right panel: Consistency – Listener vs. RMS Error (%).

The outcome of this analysis resulted in a number of listeners being removed from the listening test 2 data set (see Table I.2).

Session	Listeners whose data was removed
1	6, 7, 9, 16
2	3, 7, 9
3	7, 9
4	3, 7, 9, 15

Table I.2. Listeners removed from the subjective database of listening test 2.

I.5 Average intra-listener error (RMSE)(%)

The average intra-listener error (RMSE)(%) for listening tests 1 and 2 = 10%

Appendix J – The generalisablity of the QESTRAL model before correction

Although the model shows a low level of multicolinearity (represented by the low VIF values) and the cross-validation results indicate that the QESTRAL model would have a similar performance when used to predict other databases, a true test of the model's generalisability would be to use it to predict a different data set. However in the absence of a validation database, Field [2005] suggests a number of statistical tests that can be used to check this. He explains that to consider a regression model as generalisable it must satisfy a number of statistical conditions.

To check these conditions required that the calibrated QESTRAL model be recalculated using PCR regression which resulted in a slightly different weighting of the objective metrics to the model calibrated using PLS regression (presented in the main body of the thesis). A one-way ANOVA indicated that the models were not statistically significantly different ($p < 0.05$).

J.1 Homoscedasticity and linearity

A test for homoscedasticity determines if the residuals at each level of the predictor variables have the same variance. In other words, that the range of the error between the predicted scores and the dependent variable (subjective) scores is constant. This is a measure of the models ability to predict the subjective scores across the scale. If the range of the error is not constant (heteroscedastic) it indicates that the prediction of different dependent variables varies. Linearity is investigated to determine whether the relationship being modelled is linear (i.e. if there is a linear relationship between the objective metrics and subjective scores).

Homoscedasticity and linearity can be assessed by plotting the regression standardised residuals against regression standardised predicted values (Fig J.1) (NB. This was calculated and plotted using SPSS). Field explains that the conditions of homoscedasticity and linearity are met if the samples are randomly distributed throughout the plot which would indicate that the samples are spread evenly along the regression line. As can be seen in figure J.1 the samples were randomly distributed, indicating that the model met these assumptions.

J.2 Normally distributed errors (residuals)

It is assumed that the residuals in the model are randomly or normally distributed. This means that the difference between the predicted and measured samples should most frequently be zero or close to zero. A different distribution (e.g. multi-modal) indicates that there is variance in the data that the model does not predict. This assumption can be checked using a normal probability plot calculated in SPSS (see figure J.2). The samples in figure J.2 formed a straight line. Field explains that this means that the residuals are normally distributed, indicating homoscedasticity and that there are no obvious outliers .

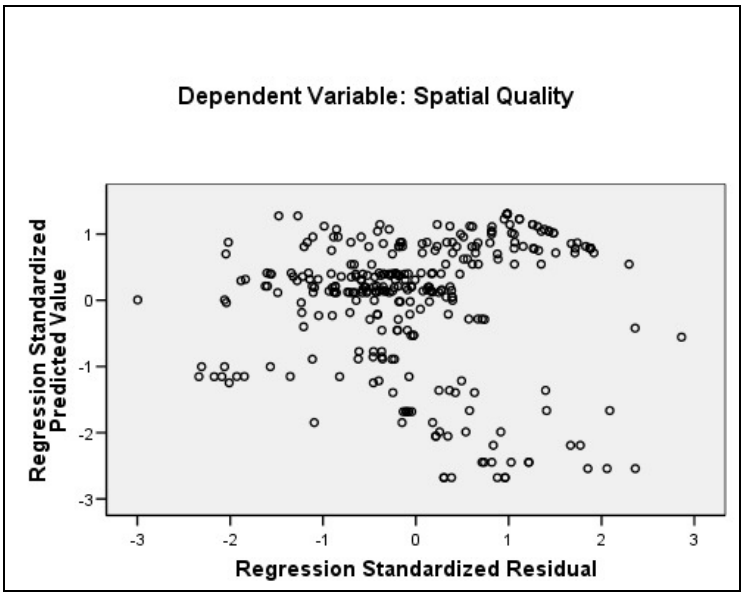


Fig J.1 Regression Standardised Residuals vs. Regression Standardised Residuals (predicted).

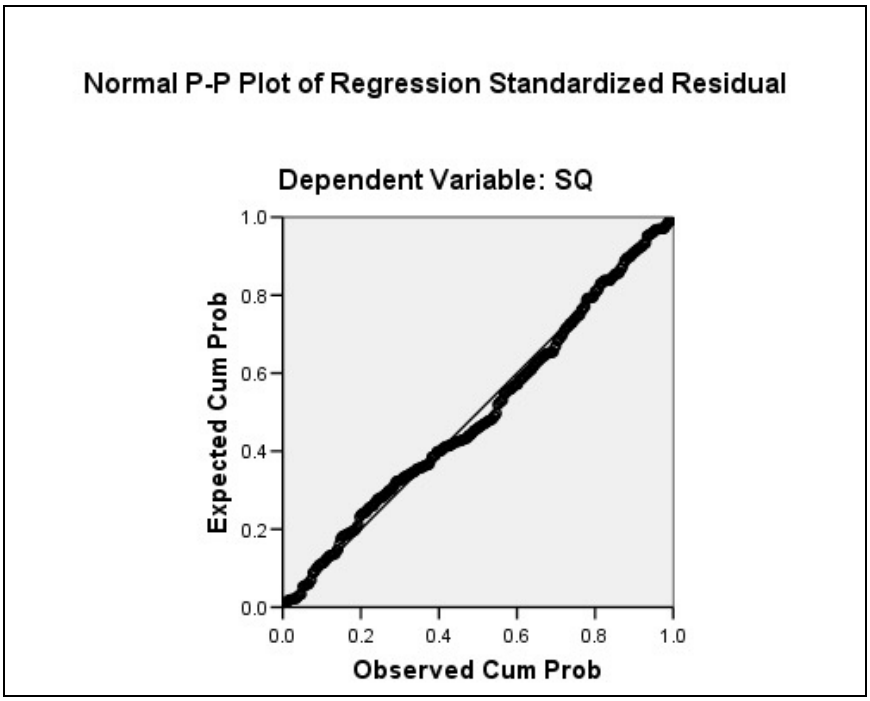


Fig J.2 Observed probability vs. Expected probability.

J.3 Conclusion

The results indicated that the conditions that Field suggests for testing the generalisability of the model were met by the calibration of the QESTRAL model. This indicates that the model is generalisable.

Appendix K – QESTRAL model results

No.	SAP	Description	Programme Item	Listening Position	Spatial Quality	
					Perceived	Predicted
1	Downmixing from 5CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 I*Rs.	1	1	90	77
2	Downmixing from 5CH 2	3.0: L = L + 0.7071*Ls, R = R + 0.7071*Rs, C = C.	1	1	73	88
3	Downmixing from 5CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 I*Rs.	1	1	65	80
4	Downmixing from 5CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	1	1	13	34
5	Multichannel audio coding 1	160kbs	1	1	78	86
6	Multichannel audio coding 2	64kbs	1	1	67	83
7	Multichannel audio coding 4	2 stage cascade (80kbs)	1	1	45	61
8	Multichannel audio coding 5	4 stage cascade (64kbs)	1	1	50	60
9	Altered loudspeaker locations 1	L and R re-positioned at -10° and 10°	1	1	72	75
10	Altered loudspeaker locations 2	C is skewed; re-positioned at 20°	1	1	95	87
11	Altered loudspeaker locations 3	Ls and Rs re-positioned at -90° and 90°	1	1	100	97
12	Altered loudspeaker locations 4	Ls and Rs re-positioned at -170° and 160°	1	1	74	78
13	Channel rearrangement 1	L and R swapped	1	1	85	72
14	Channel rearrangement 3	CH order rotated	1	1	62	66
15	Channel rearrangement 4	CH order randomised	1	1	81	46
16	Inter-channel level mis-alignment 1	L, C and R -6dB quieter than Ls and Rs	1	1	77	71
17	Inter-channel out-of-phase 1	C 180° out-of-phase	1	1	80	93
18	Channel removal 1	R removed	1	1	77	80
19	Channel removal 2	Ls removed	1	1	89	77
20	Channel removal 3	C removed	1	1	81	79
21	Spectral filtering 1	500Hz HPF on all channels	1	1	72	98
22	Spectral filtering 2	3.5kHz LPF on all channels	1	1	52	56
23	Inter-channel crosstalk 1	1.0 downmix in all CH	1	1	35	31
24	Inter-channel crosstalk 2	Partly correlated (0.5 bleed in adjacent channels)	1	1	81	70
25	Combination 2	Downmix 2 + Missing channel 1	1	1	35	49
26	Combination 3	Downmix 3 + CH routing error 4	1	1	21	34
27	Combination 4	Downmix 3 + Loudspeaker miss-placement 1	1	1	42	52
28	Combination 5	Downmix 4 + Filtering 1	1	1	8	34
29	Combination 7	Codec A + Downmix 3	1	1	63	77
30	Combination 8	Codec A + Loudspeaker miss-placement 3	1	1	72	90
31	Combination 9	Codec C + Downmix 4	1	1	22	22
32	Combination 10	Codec C + CH routing error 4	1	1	39	28
33	Combination 11	Virtual surround algorithms 2 + Missing channel 1	1	1	9	22
34	Anchor recording A	High Anchor - Unprocessed reference	1	1	100	100
35	Anchor recording B	Mid Anchor - Audio codec (80kbs)	1	1	50	63
36	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	1	1	8	15

37	Downmixing from 5CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	1	2	65	65
38	Downmixing from 5CH 2	3.0: L = L + 0.7071*Ls, R = R + 0.7071*Rs, C = C.	1	2	52	56
39	Downmixing from 5CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	1	2	52	57
40	Downmixing from 5CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	1	2	33	25
41	Multichannel audio coding 1	160kbs	1	2	60	62
42	Multichannel audio coding 2	64kbs	1	2	58	59
43	Multichannel audio coding 3	64kbs	1	2	50	42
44	Multichannel audio coding 4	2 stage cascade (80kbs)	1	2	51	59
45	Multichannel audio coding 5	4 stage cascade (64kbs)	1	2	51	48
46	Altered loudspeaker locations 1	L and R re-positioned at -10° and 10°	1	2	60	64
47	Altered loudspeaker locations 2	C is skewed; re-positioned at 20°	1	2	64	69
48	Altered loudspeaker locations 3	Ls and Rs re-positioned at -90° and 90°	1	2	58	61
49	Altered loudspeaker locations 4	Ls and Rs re-positioned at -170° and 160°	1	2	53	53
50	Channel rearrangement 1	L and R swapped	1	2	63	60
51	Inter-channel out-of-phase 1	C 180° out-of-phase	1	2	64	64
52	Channel removal 1	R removed	1	2	60	58
53	Channel removal 2	Ls removed	1	2	66	52
54	Channel removal 3	C removed	1	2	63	57
55	Spectral filtering 1	500Hz HPF on all channels	1	2	60	64
56	Spectral filtering 2	3.5kHz LPF on all channels	1	2	49	40
57	Inter-channel crosstalk 1	1.0 downmix in all CH	1	2	24	54
58	Inter-channel crosstalk 2	Partly correlated (0.5 bleed in adjacent channels)	1	2	63	48
59	Virtual surround algorithms 2	2 CH virtual surround	1	2	49	46
60	Combination 1	CH routing error 4 + Missing channel 1, 2 and 3	1	2	22	20
61	Combination 2	Downmix 2 + Missing channel 1	1	2	34	35
62	Combination 3	Downmix 3 + CH routing error 4	1	2	39	32
63	Combination 4	Downmix 3 + Loudspeaker miss-placement 1	1	2	37	39
64	Combination 5	Downmix 4 + Filtering 1	1	2	24	25
65	Combination 7	Codec A + Downmix 3	1	2	50	56
66	Combination 8	Codec A + Loudspeaker miss-placement 3	1	2	55	61
67	Combination 9	Codec C + Downmix 4	1	2	30	16
68	Combination 10	Codec C + CH routing error 4	1	2	41	30
69	Combination 11	Virtual surround algorithms 2 + Missing channel 1	1	2	25	19
70	Combination 12	Virtual surround algorithms 2 + Loudspeaker miss-placement 1	1	2	37	36
71	Anchor recording A	High Anchor - Unprocessed reference	1	2	66	65
72	Anchor recording B	Mid Anchor - Audio codec (80kbs)	1	2	53	55
73	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	1	2	22	18
74	Downmixing from 5CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	2	1	100	77
75	Downmixing from 5CH 2	3.0: L = L + 0.7071*Ls, R = R + 0.7071*Rs, C = C.	2	1	100	88
76	Downmixing from 5CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	2	1	81	80
77	Downmixing from 5CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	2	1	14	34

78	Multichannel audio coding 1	160kbs	2	1	93	86
79	Multichannel audio coding 2	64kbs	2	1	74	83
80	Multichannel audio coding 4	2 stage cascade (80kbs)	2	1	55	61
81	Multichannel audio coding 5	4 stage cascade (64kbs)	2	1	72	60
82	Altered loudspeaker locations 1	L and R re-positioned at -10° and 10°	2	1	51	75
83	Altered loudspeaker locations 2	C is skewed; re-positioned at 20°	2	1	88	87
84	Altered loudspeaker locations 3	Ls and Rs re-positioned at -90° and 90°	2	1	100	97
85	Altered loudspeaker locations 4	Ls and Rs re-positioned at -170° and 160°	2	1	99	78
86	Channel rearrangement 1	L and R swapped	2	1	100	72
87	Channel rearrangement 3	CH order rotated	2	1	50	66
88	Inter-channel level mis-alignment 1	L, C and R -6dB quieter than Ls and Rs	2	1	84	71
89	Channel removal 1	R removed	2	1	55	80
90	Channel removal 2	Ls removed	2	1	100	77
91	Channel removal 3	C removed	2	1	92	79
92	Spectral filtering 2	3.5kHz LPF on all channels	2	1	45	56
93	Inter-channel crosstalk 1	1.0 downmix in all CH	2	1	49	31
94	Inter-channel crosstalk 2	Partly correlated (0.5 bleed in adjacent channels)	2	1	80	70
95	Virtual surround algorithms 2	2 CH virtual surround	2	1	84	42
96	Combination 1	CH routing error 4 + Missing channel 1, 2 and 3	2	1	10	28
97	Combination 3	Downmix 3 + CH routing error 4	2	1	12	34
98	Combination 5	Downmix 4 + Filtering 1	2	1	10	34
99	Combination 7	Codec A + Downmix 3	2	1	76	77
100	Combination 8	Codec A + Loudspeaker miss-placement 3	2	1	91	90
101	Combination 9	Codec C + Downmix 4	2	1	30	22
102	Combination 10	Codec C + CH routing error 4	2	1	37	28
103	Combination 11	Virtual surround algorithms 2 + Missing channel 1	2	1	24	22
104	Anchor recording A	High Anchor - Unprocessed reference	2	1	100	100
105	Anchor recording B	Mid Anchor - Audio codec (80kbs)	2	1	55	63
106	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	2	1	8	15
107	Downmixing from 5CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 I*Rs.	2	2	66	65
108	Downmixing from 5CH 2	3.0: L = L + 0.7071*Ls, R = R + 0.7071*Rs, C = C.	2	2	65	56
109	Downmixing from 5CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 I*Rs.	2	2	58	57
110	Multichannel audio coding 1	160kbs	2	2	66	62
111	Multichannel audio coding 2	64kbs	2	2	61	59
112	Multichannel audio coding 3	64kbs	2	2	49	42
113	Multichannel audio coding 4	2 stage cascade (80kbs)	2	2	54	59
114	Multichannel audio coding 5	4 stage cascade (64kbs)	2	2	60	48
115	Altered loudspeaker locations 2	C is skewed; re-positioned at 20°	2	2	64	69
116	Altered loudspeaker locations 3	Ls and Rs re-positioned at -90° and 90°	2	2	66	61
117	Altered loudspeaker locations 4	Ls and Rs re-positioned at -170° and 160°	2	2	65	53
118	Channel rearrangement 1	L and R swapped	2	2	66	60

119	Channel rearrangement 2	L and R swapped for Ls and Rs	2	2	44	62
120	Channel rearrangement 4	CH order randomised	2	2	38	46
121	Inter-channel level mis-alignment 1	L, C and R -6dB quieter than Ls and Rs	2	2	59	52
122	Inter-channel out-of-phase 1	C 180° out-of-phase	2	2	62	64
123	Channel removal 1	R removed	2	2	59	58
124	Channel removal 2	Ls removed	2	2	66	52
125	Channel removal 3	C removed	2	2	65	57
126	Spectral filtering 1	500Hz HPF on all channels	2	2	58	64
127	Spectral filtering 2	3.5kHz LPF on all channels	2	2	49	40
128	Inter-channel crosstalk 2	Partly correlated (0.5 bleed in adjacent channels)	2	2	62	48
129	Virtual surround algorithms 2	2 CH virtual surround	2	2	64	46
130	Combination 1	CH routing error 4 + Missing channel 1, 2 and 3	2	2	20	20
131	Combination 3	Downmix 3 + CH routing error 4	2	2	28	32
132	Combination 4	Downmix 3 + Loudspeaker miss-placement 1	2	2	39	39
133	Combination 5	Downmix 4 + Filtering 1	2	2	24	25
134	Combination 6	Loudspeaker miss-placement 4 + Loudspeaker miss-placement 1	2	2	45	47
135	Combination 7	Codec A + Downmix 3	2	2	60	56
136	Combination 8	Codec A + Loudspeaker miss-placement 3	2	2	65	61
137	Combination 9	Codec C + Downmix 4	2	2	36	16
138	Combination 10	Codec C + CH routing error 4	2	2	30	30
139	Combination 11	Virtual surround algorithms 2 + Missing channel 1	2	2	36	19
140	Anchor recording A	High Anchor - Unprocessed reference	2	2	66	65
141	Anchor recording B	Mid Anchor - Audio codec (80kbs)	2	2	56	55
142	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	2	2	24	18
143	Downmixing from 5CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 I*Rs.	3	1	98	77
144	Downmixing from 5CH 2	3.0: L = L + 0.7071*Ls, R = R + 0.7071*Rs, C = C.	3	1	80	88
145	Downmixing from 5CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 I*Rs.	3	1	84	80
146	Downmixing from 5CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	3	1	20	34
147	Multichannel audio coding 1	160kbs	3	1	93	86
148	Multichannel audio coding 2	64kbs	3	1	70	83
149	Multichannel audio coding 4	2 stage cascade (80kbs)	3	1	53	61
150	Multichannel audio coding 5	4 stage cascade (64kbs)	3	1	58	60
151	Altered loudspeaker locations 1	L and R re-positioned at -10° and 10°	3	1	86	75
152	Altered loudspeaker locations 2	C is skewed; re-positioned at 20°	3	1	92	87
153	Altered loudspeaker locations 3	Ls and Rs re-positioned at -90° and 90°	3	1	98	97
154	Altered loudspeaker locations 4	Ls and Rs re-positioned at -170° and 160°	3	1	80	78
155	Channel rearrangement 1	L and R swapped	3	1	80	72
156	Channel rearrangement 2	L and R swapped for Ls and Rs	3	1	93	79
157	Channel rearrangement 3	CH order rotated	3	1	75	66
158	Inter-channel level mis-alignment 1	L, C and R -6dB quieter than Ls and Rs	3	1	81	71
159	Inter-channel out-of-phase 1	C 180° out-of-phase	3	1	100	93

160	Channel removal 1	R removed	3	1	71	80
161	Channel removal 2	Ls removed	3	1	87	77
162	Channel removal 3	C removed	3	1	76	79
163	Spectral filtering 1	500Hz HPF on all channels	3	1	70	98
164	Spectral filtering 2	3.5kHz LPF on all channels	3	1	46	56
165	Inter-channel crosstalk 1	1.0 downmix in all CH	3	1	36	31
166	Inter-channel crosstalk 2	Partly correlated (0.5 bleed in adjacent channels)	3	1	85	70
167	Combination 1	CH routing error 4 + Missing channel 1, 2 and 3	3	1	29	28
168	Combination 2	Downmix 2 + Missing channel 1	3	1	45	49
169	Combination 3	Downmix 3 + CH routing error 4	3	1	15	34
170	Combination 4	Downmix 3 + Loudspeaker miss-placement 1	3	1	48	52
171	Combination 5	Downmix 4 + Filtering 1	3	1	11	34
172	Combination 7	Codec A + Downmix 3	3	1	83	77
173	Combination 8	Codec A + Loudspeaker miss-placement 3	3	1	89	90
174	Combination 9	Codec C + Downmix 4	3	1	25	22
175	Combination 10	Codec C + CH routing error 4	3	1	28	28
176	Combination 11	Virtual surround algorithms 2 + Missing channel 1	3	1	20	22
177	Anchor recording A	High Anchor - Unprocessed reference	3	1	100	100
178	Anchor recording B	Mid Anchor - Audio codec (80kbs)	3	1	53	63
179	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	3	1	9	15
180	Downmixing from 5CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 I*Rs.	3	2	66	65
181	Downmixing from 5CH 2	3.0: L = L + 0.7071*Ls, R = R + 0.7071*Rs, C = C.	3	2	54	56
182	Downmixing from 5CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 I*Rs.	3	2	57	57
183	Multichannel audio coding 1	160kbs	3	2	64	62
184	Multichannel audio coding 2	64kbs	3	2	59	59
185	Multichannel audio coding 3	64kbs	3	2	47	42
186	Multichannel audio coding 5	4 stage cascade (64kbs)	3	2	55	48
187	Altered loudspeaker locations 1	L and R re-positioned at -10° and 10°	3	2	63	64
188	Altered loudspeaker locations 2	C is skewed; re-positioned at 20°	3	2	66	69
189	Altered loudspeaker locations 3	Ls and Rs re-positioned at -90° and 90°	3	2	62	61
190	Altered loudspeaker locations 4	Ls and Rs re-positioned at -170° and 160°	3	2	55	53
191	Channel rearrangement 1	L and R swapped	3	2	64	60
192	Channel rearrangement 3	CH order rotated	3	2	58	51
193	Inter-channel level mis-alignment 1	L, C and R -6dB quieter than Ls and Rs	3	2	58	52
194	Inter-channel out-of-phase 1	C 180° out-of-phase	3	2	66	64
195	Channel removal 1	R removed	3	2	65	58
196	Channel removal 2	Ls removed	3	2	66	52
197	Channel removal 3	C removed	3	2	59	57
198	Spectral filtering 2	3.5kHz LPF on all channels	3	2	49	40
199	Inter-channel crosstalk 2	Partly correlated (0.5 bleed in adjacent channels)	3	2	63	48
200	Virtual surround algorithms 2	2 CH virtual surround	3	2	52	46

201	Combination 1	CH routing error 4 + Missing channel 1, 2 and 3	3	2	20	20
202	Combination 2	Downmix 2 + Missing channel 1	3	2	38	35
203	Combination 3	Downmix 3 + CH routing error 4	3	2	39	32
204	Combination 4	Downmix 3 + Loudspeaker miss-placement 1	3	2	40	39
205	Combination 5	Downmix 4 + Filtering 1	3	2	25	25
206	Combination 6	Loudspeaker miss-placement 4 + Loudspeaker miss-placement 1	3	2	43	47
207	Combination 7	Codec A + Downmix 3	3	2	54	56
208	Combination 8	Codec A + Loudspeaker miss-placement 3	3	2	63	61
209	Combination 9	Codec C + Downmix 4	3	2	32	16
210	Combination 11	Virtual surround algorithms 2 + Missing channel 1	3	2	35	19
211	Combination 12	Virtual surround algorithms 2 + Loudspeaker miss-placement 1	3	2	38	36
212	Anchor recording A	High Anchor - Unprocessed reference	3	2	66	65
213	Anchor recording B	Mid Anchor - Audio codec (80kbs)	3	2	57	55
214	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	3	2	24	18
215	Down-mixing from 5 CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	4	1	100	77
216	Down-mixing from 5 CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	4	1	92	80
217	Multichannel audio coding 1	160kbs	4	1	100	86
218	Altered loudspeaker locations 5	L and C moved 1m to left and not facing listening position	4	1	79	77
219	Altered loudspeaker locations 6	Ls moved 1m to left and not facing listening position	4	1	100	84
220	Channel rearrangements 1	L and R swapped	4	1	76	72
221	Inter-channel level mis-alignment 1	LCR -6dB	4	1	99	71
222	Inter-channel level mis-alignment 2	Surrounds -6dB	4	1	100	74
223	Inter-channel out-of-phase errors 1	C 180° out-of-phase	4	1	88	93
224	Inter-channel out-of-phase errors 2	LCR 180° out-of-phase	4	1	91	91
225	Channel removal 3	C removed	4	1	76	79
226	Combination 5	Down-mixing from 5 CH 4 + Spectral filtering 1	4	1	36	34
227	Combination 7	Multichannel audio coding 1 + Down-mixing from 5 CH 3	4	1	86	77
228	Combination 13	Multichannel audio coding 3 + Altered loudspeaker locations 5	4	1	51	46
229	Anchor recording A	High Anchor - Unprocessed reference	4	1	100	100
230	Anchor recording B	Mid Anchor - Audio codec (80kbs)	4	1	70	63
231	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	4	1	16	15
232	Down-mixing from 5 CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	4	2	65	65
233	Down-mixing from 5 CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	4	2	66	57
234	Down-mixing from 5 CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	4	2	51	25
235	Multichannel audio coding 1	160kbs	4	2	67	62
236	Channel rearrangements 1	L and R swapped	4	2	58	60
237	Inter-channel level mis-alignment 1	LCR -6dB	4	2	62	52
238	Inter-channel level mis-alignment 2	Surrounds -6dB	4	2	67	55
239	Inter-channel out-of-phase errors 1	C 180° out-of-phase	4	2	65	64
240	Inter-channel out-of-phase errors 2	LCR 180° out-of-phase	4	2	65	65

241	Channel removal 3	C removed	4	2	57	57
242	Spectral filtering 1	500Hz HPF on all channels	4	2	52	64
243	Combination 5	1.0 Downmix + Spectral filter 1	4	2	24	25
244	Combination 7	Codec A + 2.0 Downmix	4	2	58	56
245	Anchor recording A	High Anchor - Unprocessed reference	4	2	68	65
246	Anchor recording B	Mid Anchor - Audio codec (80kbs)	4	2	55	55
247	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	4	2	19	18
248	Down-mixing from 5 CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	5	1	93	77
249	Down-mixing from 5 CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	5	1	77	80
250	Multichannel audio coding 1	160kbs	5	1	99	86
251	Multichannel audio coding 2	64kbs	5	1	71	83
252	Multichannel audio coding 3	64kbs	5	1	51	59
253	Altered loudspeaker locations 5	L and C moved 1m to left and not facing listening position	5	1	93	77
254	Altered loudspeaker locations 6	Ls moved 1m to left and not facing listening position	5	1	95	84
255	Channel rearrangements 1	L and R swapped	5	1	89	72
256	Inter-channel level mis-alignment 1	LCR -6dB	5	1	94	71
257	Inter-channel level mis-alignment 2	Surrounds -6dB	5	1	100	74
258	Inter-channel out-of-phase errors 1	C 180° out-of-phase	5	1	100	93
259	Inter-channel out-of-phase errors 2	LCR 180° out-of-phase	5	1	94	91
260	Channel removal 3	C removed	5	1	100	79
261	Spectral filtering 2	3.5kHz LPF on all channels (BS.1534)	5	1	60	56
262	Combination 5	Down-mixing from 5 CH 4 + Spectral filtering 1	5	1	26	34
263	Combination 7	Multichannel audio coding 1 + Down-mixing from 5 CH 3	5	1	72	77
264	Anchor recording A	High Anchor - Unprocessed reference	5	1	100	100
265	Anchor recording B	Mid Anchor - Audio codec (80kbs)	5	1	65	63
266	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	5	1	15	15
267	Down-mixing from 5 CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	5	2	59	57
268	Multichannel audio coding 2	64kbs	5	2	54	59
269	Multichannel audio coding 3	64kbs	5	2	49	42
270	Inter-channel level mis-alignment 2	Surrounds -6dB	5	2	68	55
271	Inter-channel out-of-phase errors 1	C 180° out-of-phase	5	2	66	64
272	Channel removal 3	C removed	5	2	67	57
273	Inter-channel crosstalk 1	1.0 downmix in all CH	5	2	36	54
274	Combination 5	1.0 Downmix + Spectral filter 1	5	2	24	25
275	Combination 7	Codec A + 2.0 Downmix	5	2	53	56
276	Anchor recording 1	High Anchor - Unprocessed reference	5	2	71	65
277	Anchor recording 2	Mid Anchor - Audio codec (80kbs)	5	2	45	55
278	Anchor recording 3	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	5	2	18	18
279	Down-mixing from 5 CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	6	1	100	77

280	Down-mixing from 5 CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	6	1	83	80
281	Multichannel audio coding 1	160kbs	6	1	100	86
282	Multichannel audio coding 2	64kbs	6	1	87	83
283	Altered loudspeaker locations 5	L and C moved 1m to left and not facing listening position	6	1	65	77
284	Altered loudspeaker locations 6	Ls moved 1m to left and not facing listening position	6	1	100	84
285	Channel rearrangements 1	L and R swapped	6	1	80	72
286	Inter-channel level mis-alignment 1	LCR -6dB	6	1	97	71
287	Inter-channel level mis-alignment 2	Surrounds -6dB	6	1	100	74
288	Inter-channel out-of-phase errors 1	C 180° out-of-phase	6	1	97	93
289	Inter-channel out-of-phase errors 2	LCR 180° out-of-phase	6	1	100	91
290	Combination 7	Multichannel audio coding 1 + Down-mixing from 5 CH 3	6	1	85	77
291	Anchor recording A	High Anchor - Unprocessed reference	6	1	100	100
292	Anchor recording B	Mid Anchor - Audio codec (80kbs)	6	1	72	63
293	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	6	1	16	15
294	Down-mixing from 5 CH 1	3/1: L = L, R = R, C = C, S = 0.7071*Ls + 0.707 l*Rs.	6	2	68	65
295	Down-mixing from 5 CH 3	2.0: L = L + 0.7071*C + 0.7071*Ls, R = R + 0.7071*C + 0.707 l*Rs.	6	2	61	57
296	Down-mixing from 5 CH 4	1.0: C = 0.7071*L + 0.7071*R + C + 0.5*Ls + 0.5*Rs.	6	2	43	25
297	Multichannel audio coding 1	160kbs	6	2	60	62
298	Multichannel audio coding 2	64kbs	6	2	69	59
299	Channel rearrangements 1	L and R swapped	6	2	67	60
300	Inter-channel level mis-alignment 1	LCR -6dB	6	2	58	52
301	Inter-channel level mis-alignment 2	Surrounds -6dB	6	2	66	55
302	Inter-channel out-of-phase errors 1	C 180° out-of-phase	6	2	66	64
303	Inter-channel out-of-phase errors 2	LCR 180° out-of-phase	6	2	61	65
304	Spectral filtering 1	500Hz HPF on all channels	6	2	53	64
305	Combination 7	Codec A + 2.0 Downmix	6	2	66	56
306	Anchor recording A	High Anchor - Unprocessed reference	6	2	66	65
307	Anchor recording B	Mid Anchor - Audio codec (80kbs)	6	2	55	55
308	Anchor recording C	Low Anchor - Mono downmix reproduced asymmetrically by the rear left loudspeaker only	6	2	18	18

Table K1 QESTRAL model results - comparing subjective and predicted scores.

References

Abdi, H. (2007) “Partial Least Square Regression PLS-Regression” in Salkind, N. (Ed.) *Encyclopedia of Measurement Statistics*, Thousand Oaks, California, USA.

Bang & Olufsen (2010) “Car Audio” http://www.bang-olufsen.com/car-audio [Accessed 20/09/10].

Bang & Olufsen (2011) “Beolab 3 specifications” http://www.bang-olufsen.com/specifications?productid=38 [Accessed 03/03/11].

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

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

Bech S (1999) “Methods for subjective evaluation of spatial characteristics of sound” presented at the Audio Engineering Society 16th International Conference, April 10 – 12, Rovaniemi, Finland, Preprint 16-044.

Beerends J. G., Stemerdink J. A (1992) “A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation”, J. Audio Eng. Soc., Vol. 40 (12), pp. 963 – 978.

Beranek L (1996) “Concert and opera halls: how they sound” Acoustical Society of America, USA.

Berg, J., Rumsey, F. (1999a) “Identification of Perceived Spatial Attributes of Recordings by Repertory Grid Technique and Other Methods” Presented at the Audio Engineering Society 106th Convention, May 8 – 11, Munich, Germany, Preprint 4924

Berg, J., Rumsey, F. (1999b) “Spatial Attribute Identification and Scaling by Repertory Grid Technique and Other Methods” presented at the Audio Engineering Society 16th International Conference, April 10 – 12, Rovaniemi, Finland, Preprint 16-005.

Berg, J., Rumsey, F. (2000a) “In Search of the Spatial Dimensions of Reproduced Sound: Verbal Protocol Analysis and Cluster Analysis of Scaled Verbal Descriptors” presented at the Audio Engineering Society 108th Convention, Feb 19 – 22, Paris, France, Preprint 5139.

Berg, J., Rumsey, F. (2000b) "Correlation between Emotive, Descriptive and Naturalness Attributes in Subjective Data Relating to Spatial Sound Reproduction" presented at the Audio Engineering Society 109th Convention, Sep 22 – 25, Los Angeles, USA, Preprint 5206.

Berg, J., Rumsey, F. (2001) "Verification and correlation of attributes used for describing the spatial quality of reproduced sound" presented at the Audio Engineering Society 19th International Conference, June 21 – 24, Schloss Elmau, Germany, Preprint 1932.

Berg, J., Rumsey, F. (2003) "Systematic evaluation of perceived spatial quality" presented at the Audio Engineering Society 24th International Conference, June 26 – 28, Banff, Canada, Preprint 43.

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.

Blauert, J. & Jekosch, U. (1997) "Sound-Quality Evaluation – A Multi-Layered Problem" Acustica, Vol.83, pp. 747-753.

Blumlein, A. (1958) "British Patent Specification 394,325 (Improvements in and relating to Soundtransmission, Sound-recording and Sound-reproducing Systems)" J. Audio Eng. Soc., Vol. 6(2), pp. 91–98.

Bose (2010) "Automotive systems" http://www.bose.co.uk/GB/en/automotive-systems/automotive-systems/index.jsp [Accessed 20/09/10].

Blauert, J. (2001) "Spatial hearing: the psychoacoustics of human sound localization" The MIT press. USA.

Brandenburg, K. (1987) "Evaluation of quality for audio encoding at low bit rates" presented at the Audio Engineering Society 82nd Convention, London, UK, preprint 2433.

Bregman, A.S. (1990) "Auditory scene analysis: the perceptual organisation of sound" MIT Press, Cambridge, Massachusetts, USA.

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

Choi I, Shinn-Cunningham B.G, Chon S. B, And Sung K (2007) "Prediction of perceived quality in multi-channel audio compression coding systems" presented at Audio Engineering Society 30th International Conference, Mar 15 – 17, Saariselkä, Finland.

Choi, I., Shinn-Cunningham, B.G., Chon, S. B, & Sung, K. (2008) "Objective Measurement of Perceived Auditory Quality in Multichannel Audio Compression Coding Systems" J. Audio Eng. Soc., Vol. 56 (1/2), pp. 3 – 17.

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.

Choisel, S. & Wickelmaier, F. (2006a) "Relating Auditory Attributes of Multichannel Reproduced Sound to Preference and to Physical Parameters" presented at the Audio Engineering Society 120th Convention, May 20 – 23, Paris, France, Preprint 6684.

Choisel, S. & Wickelmaier, F. (2006b) "Extraction of Auditory Features and Elicitation of Attributes for the Assessment of Multichannel Reproduced Sound" J. Audio Eng. Soc., Vol. 54 (9), pp. 815 - 826.

Colomes C., Lever M., Rault J. B., Dehery Y. F (1995) "A Perceptual Model Applied to Audio Bit-Rate Reduction" J. Audio Eng. Soc., Vol. 43 (4), pp. 233 – 240.

Conetta, R., Jackson, P.J.B., Zielinski, S. & Rumsey, F. (2007) "Envelopment: What is it? A definition for multichannel audio" presented at the 1st SpACE-Net Workshop, Jan 25, University of York, UK.

Conetta, R. (2007) "Scaling and predicting spatial attributes of reproduced sound using an artificial listener" MPhil-PhD Transfer Report, Institute of Sound Recording, University of Surrey.

Conetta, R., Rumsey, F., Zielinski, S., Jackson, P.J.B, Dewhirst, M., Bech, S., Meares D. & George, S (2008a). "QESTRAL (Part 2): Calibrating the QESTRAL spatial quality model using listening test data" presented at the Audio Engineering Society 125th Convention, Oct 2 – 5, San Francisco, Preprint 7596.

Conetta, R., Rumsey, F., Zielinski, S., Jackson, P.J.B, Dewhirst, M., Bech, S., Meares D. & George, S (2008b). "Calibration of the QESTRAL model for the prediction of spatial quality" proceedings of the Institute of Acoustics 24th Reproduced Sound Conference, Nov 20-21, Brighton, UK.

Davis, M. (2003) “History of Spatial Coding” J. Audio Eng. Soc, Vol.51 No.6, pp. 554-569.
De Vries, D. (2007) “Wave Field Synthesis: Reality or Illusion at your choice” presented at the Audio Engineering Society 22nd UK Conference, April 11 – 12, Cambridge, UK.

Dewhirst, M., Zielinski, S., Jackson, P.J.B & Rumsey, F. (2005) “Objective Assessment of Spatial Localisation Attributes of Surround-Sound Reproduction Systems” presented at the Audio Engineering Society 118th Convention, May 28-31, Barcelona, Spain, Preprint 6441.

Dewhirst, M. (2008) “Modelling perceived spatial attributes of reproduced sound” PhD Thesis, Institute of Sound Recording, University of Surrey.

Dewhirst, M., Conetta, R., Rumsey, F., Jackson, P.J.B, Zielinski, S., Bech, S., Meares D. & George, S (2008) “QESTRAL (Part 4): Test signals, combining metrics and the prediction of overall spatial quality”. presented at the Audio Engineering Society 125th Convention, Oct 2 – 5, San Francisco, USA, Preprint 7598.

Dougherty, D. (2009) “The sizzling sound of music” http://radar.oreilly.com/2009/03/the-sizzling-sound-of-music.html [Accessed 11/08/10].

Draper, N.R. & Smith, H. (1981) “Applied Regression Analysis” 2nd Edition, Wiley, USA.

Esbensen, K. (2002) “Multivariate Data Analysis - in practice” 5th Edition, CAMO Process AS, Norway.

Feiten, B., Wolf, I. & Graffunder, A. (2005) “Audio Adaptation According to Usage Environment and Perceptual Quality Metrics” IEEE transactions on Multimedia, Vol. 7 (3), pp. 446 – 453.

Field, A. (2005) “Discovering Statistics Using SPSS” 2nd Edition, SAGE Publications Ltd, UK.

Gabrielsson, A. & Lindström, B. “Perceived Sound Quality of High-Fidelity Loudspeakers” J. Audio Eng. Soc., Vol. 33 (1/2), pp. 33 – 53.

George, S., Zielinski, S. & Rumsey, F. (2006a) “Feature Extraction for the Prediction of Multichannel Spatial Audio Fidelity”, IEEE Transactions on Audio, Speech, and Language Processing, Vol.14, No.6, pp.1994-2005.

George, S., Zielinski, S. & Rumsey, F. (2006b) “Initial developments of an objective method for the prediction of basic audio quality for surround audio recordings” presented at the Audio Engineering Society 120th International Convention, May 20-23, Paris, France, Preprint 6686.

George, S., Zielinski, S., Rumsey, F. & Bech, S. (2008) "Evaluating the sensation of envelopment arising from 5-channel surround sound recordings" presented at the Audio Engineering Society 124th Convention, May 17-20, Amsterdam, The Netherlands, Preprint 7382.

George, S. (2009) “Objective models for predicting selected multichannel audio quality attributes” PhD Thesis, Institute of Sound Recording, University of Surrey.

Glasberg B.R. & Moore, B.C.J. (2002) “A Model of Loudness Applicable to Time-Varying Sounds” J. Acoust. Soc. Am. Vol 50 (5), pp. 331 – 342.

Griesinger, D. (1997) “Spatial Impression and Envelopment in Small Rooms” presented at the Audio Engineering Society 103rd International Convention, Sep 26 – 29, New York, USA, Preprint 4638.

Guastavino, C. & Katz, B.F.G. (2004) “Perceptual evaluation of multi-dimensional spatial audio reproduction” J. Acoust. Soc. Am. Vol. 116 (2), pp. 1105 – 1115.

Hands, D. (2004) “A basic multimedia quality model” IEEE Transactions on multimedia, Vol. 6 (6), pp. 806 – 816.

Hamasaki, K., Hiyama, K. & Okumura, R. (2005) "The 22.2 multichannel sound system and its application" presented at the Audio Engineering Society 118th Convention, May 28 – 31, Barcelona, Spain, Preprint 6406.

Hiyama, K., Komiyama, S. & Hamasaki, K. (2002) “The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field” presented at the Audio Engineering Society 113th Convention, Oct 5 – 8, Los Angeles, USA, Preprint 5674.

ITU-R BS.775-1 (1992-1994) “Multichannel stereophonic sound system with and without accompanying picture” International Telecommunication Union recommendation.

ITU-T P.800 (1996) “Methods for subjective determination of transmission quality” International Telecommunication Union recommendation.

ITU-R BS.1116-1 (1997) “Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems” International Telecommunication Union recommendation.

ITU-R BS.1284 (1997-2003) “General methods for the subjective assessment of sound quality” International Telecommunication Union recommendation.

ITU-T P.862 (2001) “PESQ an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs” International Telecommunication Union recommendation.

ITU-R BS.1387 (2001) “Method for objective measurements of perceived audio quality” International Telecommunication Union recommendation.

ITU-R BS.1534 (2001) “Method for the subjective assessment of intermediate audio quality” International Telecommunication Union recommendation.

ITU-R BS.1387 (2001) “Method for objective measurements of perceived audio quality” International Telecommunication Union recommendation.

Jackson, P.J.B, Dewhirst, M., Conetta, R., Rumsey, F., Zielinski, 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.

Jackson, P.J.B., Dewhirst, M., Conetta, R. & Zielinski, S. (2010) “Estimates of perceived spatial quality across the listening area” presented at the Audio Engineering Society 31st International Conference, Jun 13 – 15, Pitea, Sweden.

Jiao, Y., Zielinski, S. & Rumsey, F. (2007) “Adaptive Karhunen-Lòeve Transform for Multichannel Audio” presented at the Audio Engineering Society 123rd International Convention, Oct 5 – 8, New York, USA, Preprint 7298.

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.

Macpherson, E. (1991) “A computer model of binaural localisation for stereo imaging measurement” J. Audio Eng. Soc., Vol.39 (9), pp. 604 – 622.

Martin, G., Woszczyk, W., Corey, J. & Quesnel, R. (1999) "Sound Source Localization in a Five-Channel Surround Sound Reproduction System" presented at the Audio Engineering Society 107th Convention, Sep 24 – 27, New York, USA, Preprint 4994.

Marins, P. (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.

Mason, R. (2002) "Elicitation and measurement of auditory spatial attributes in reproduced sound" PhD Thesis, Institute of Sound Recording, University of Surrey.

Mason, R. (2006) "Implementation and application of a binaural hearing model to the objective evaluation of spatial hearing" presented at the Audio Engineering Society 28th International Conference, June 30 – July 2, Pitea, Sweden.

Martin, G. (2006) "Introduction to Sound Recording" http://www.tonmeister.ca/main/textbook/node363.html [Accessed 11/08/10].

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

Morimoto, M. (1997) "The Role of Rear Loudspeakers in Spatial Impression" presented at the Audio Engineering Society 103rd Convention, Sep 26 – 29, New York, USA, Preprint 4554.

Nakayama, T., Miura, T., Kosaka, O., Michio, O & Shiga, T. (1971) "Subjective Assessment of Multichannel Reproduction" J. Audio Eng. Soc., Vol.19 (9), pp. 744 – 751, Preprint 2825.

Oxford University Press (2010) http://oxforddictionaries.com/view/entry/m_en_gb0678350#m_en_gb0678350 [Accessed 11/08/10].

Paillard, B., Mabilleau, B., Morisette, S., Soumagne, J. (1992) "Perceval: Perceptual Evaluation of the Quality of Audio Signals" J. Audio Eng. Soc., Vol. 40 (1/2), pp. 21 – 31 .

Pocock, M. (1982) "A computer model of binaural localisation" presented at the Audio Engineering Society 72nd International Convention, California, USA, Preprint 1951.

Pulkki, V., Karjalainen, M. & Huopaniemi, J. (1999) "Analyzing virtual sound source attributes using a binaural model" J. Audio Eng. Soc., Vol.47 (4), pp. 203 – 217.

Rumsey, F. (1998) “Subjective Assessment of the Spatial Attributes of Reproduced Sound” presented at the Audio Engineering Society 15th International Conference: Audio, Acoustics & Small Space. Oct 31 – Nov 2. Copenhagen, Denmark.

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

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.

Rumsey, F., Zielinski S., Bech, S. & Kassier, R. (2005a) “Relationships Between Experienced Listener Ratings of Multichannel Audio Quality and Naïve Listener Preferences” J. Acoust. Soc. Am., Vol. 117 (6), pp. 3832 – 40.

Rumsey, F., Zielinski S., Bech, S. & Kassier, R. (2005b) “On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality” J. Acoust. Soc. Am., Vol. 118 (2), pp. 968 – 976.

Rumsey, F., Jackson, P.J.B. & Zielinski, S. (2005c) “Quality of service evaluation for spatial audio coding and processing systems” EPSRC Grant EP/D041244/1.

Rumsey, F., Zielinski, S., Jackson, P.J.B, Dewhirst, M., Conetta, R., George, S., Bech, S. & Meares D. (2008) “QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener” presented at the Audio Engineering Society 125th Convention, Oct 2 – 5, San Francisco, Preprint 7595.

Seefeldt, A., Crockett, B. & Smithers, M. (2004) “A New Objective Measure of Perceived Loudness” presented at the Audio Engineering Society 117th International Convention, Oct 28 – 31, San Francisco, USA, Preprint 6236.

Seefeldt, A. & Lyman, S. (2006) “A Comparisons of Various Multichannel Loudness Measurement Techniques” presented at the Audio Engineering Society 121st International Convention, Oct 5 – 8, San Francisco, USA, Preprint 6918.

Soulodre, G., Lavoie, M. & Norcross, S. (2002) “Investigation of Listener Envelopment in Multichannel Surround Systems” presented at the Audio Engineering Society 113th International Convention, Oct 5 – 8, Los Angeles, USA, Preprint 5676.

Soulodre, G.A., Lavoie, M.C. & Norcross, S.G. (2003a) "Temporal Aspects of Listener Envelopment in Multichannel Surround Systems" presented at the Audio Engineering Society 114th International Convention, Mar 22 – 25, Amsterdam, Netherlands, Preprint 5803.

Soulodre, G.A., Lavoie, M.C., Norcross, S.G. (2003b) "Objective Measures of Listener Envelopment in Multichannel Surround Systems". J. Audio Eng. Soc., Vol.51 (9), pp. 826 – 840.

Sporer, T. (1997) "Objective Audio Signal Evaluation Applied Psychoacoustics for Modeling the Perceived Quality of Digital Audio" Presented at the 103rd AES Convention, New York, September 1997

Supper. B. (2005) "An onset-guided spatial analyser for binaural audio" PhD Thesis, Institute of Sound Recording, University of Surrey.

Thiede, T., Kabot, E. (1996) "A New Perceptual Quality Measure for the Bit Rate Reduced Audio" presented at the Audio Engineering Society 100th Convention, May 11 – 14, Berlin, Germany, Preprint 4280.

Toole, F. & Olive, S. (1994) "Hearing is Believing vs. Believing is Hearing: Blind vs. Sighted Listening Tests, and Other Interesting Things" presented at the Audio Engineering Society 97th International Convention, Nov 10 – 13, San Francisco, USA, Preprint 3894.

University of Surrey (2010) "Research students and projects" http://www3.surrey.ac.uk/soundrec/php/dkoya.php [Accessed 20/09/10].

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.

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.

Zacharov, N., Koivuniemi, K. (2001c) "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.

Zielinski S., Rumsey, F. & Bech, S. (2002) "Subjective audio quality trade-offs in consumer multichannel audio-visual delivery systems. Part I: Effects of high frequency limitation" presented at the Audio Engineering Society 112th Convention, May 10 – 13, Munich, Germany, Preprint 5562.

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

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

Zielinski, S., Rumsey, F., Bech, S. & Kassier, R. (2004) "Quality Adviser – A Multichannel Audio Quality Expert System" presented at the Audio Engineering Society 116th Convention, 8-11 May, Berlin, Germany, Preprint 6140.

Zielinski, S., Rumsey, F., Kassier, R., & Bech, S. (2005a) "Development and Initial Validation of a Multichannel Audio Quality Expert System". J. Audio Eng. Soc., Vol.53 (1/2), pp 4-21.

Zielinski, S., Rumsey, F., Bech, S. & Kassier, R. (2005b) "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.

Zielinski, S., Hardisty, P., Hummersone, C. & Rumsey, F. (2007a) "Potential Biases in MUSHRA Listening Tests" presented at the Audio Engineering Society 123rd Convention, Oct 5 – 8, New York, USA, Preprint 7179.

Zielinski, S., Brooks, P. & Rumsey, F. (2007b) "On the Use of Graphic Scales in Modern Listening Tests" presented at the Audio Engineering Society 123rd Convention, Oct 5 – 8, New York, USA, Preprint 7176.

Zielinski, 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.