



Audio Engineering Society

Convention Paper

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

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

QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener

Francis Rumsey¹, Slawomir Zielinski¹, Philip Jackson¹, Martin Dewhurst¹,
Robert Conetta¹, Sunish George¹, Søren Bech², David Mearns

¹ University of Surrey, Guildford, Surrey GU2 7XH, United Kingdom

² Bang & Olufsen a/s, Peter Bangs Vej 15, 7600 Struer, Denmark

³ DJM Consultancy, Sussex, UK

¹ f.rumsey@surrey.ac.uk

ABSTRACT

Most current perceptual models for audio quality have so far tended to concentrate on the audibility of distortions and noises that mainly affect the timbre of reproduced sound. The QESTRAL model, however, is specifically designed to take account of distortions in the spatial domain such as changes in source location, width and envelopment. It is not aimed only at codec quality evaluation but at a wider range of spatial distortions that can arise in audio processing and reproduction systems. The model has been calibrated against a large database of listening tests designed to evaluate typical audio processes, comparing spatially degraded multichannel audio material against a reference. Using a range of relevant metrics and a sophisticated multivariate regression model, results are obtained that closely match those obtained in listening tests.

1. INTRODUCTION

Most current perceptual models for audio quality have so far tended to concentrate on the audibility of distortions and noises that mainly affect the timbre of reproduced sound. Models such as PEAQ (ITU-R BS1387) [1] are designed primarily to evaluate the audibility of codec distortions in terms of basic audio quality, or mean opinion score, for example, and have not explicitly taken spatial distortions into account. The QESTRAL model described in this paper, however, is specifically designed to evaluate the effect of distortions in the spatial domain such as

changes in source location, width and envelopment. This model aims, among other things, to predict an overall spatial quality score for an audio reproduction, which closely matches the one that would have been obtained in a listening test. In its first embodiment, this model is designed to compare a five channel (ITU-R BS.775) reference signal and altered (degraded) versions of the same. However it has been designed in such a way as to enable its use with any arbitrary spatial format, either with or without a reference signal.

The QESTRAL model is not aimed only at codec quality evaluation but at the evaluation of a wider range of spatial distortions that can arise in audio processing and reproduction systems. This includes such things as downmixing algorithms, spatial audio codecs, loudspeaker misplacement, level misalignment and system phase errors. A current embodiment of the model, introduced in this paper, aims to predict overall spatial quality or a 'spatial mean opinion score', that is a global attribute describing any and all changes in the spatial attributes of the reproduced audio signal. The model has been calibrated against a large database of listening tests designed to evaluate typical audio processes, comparing spatially degraded multichannel audio material against a reference. Using a range of relevant metrics and a sophisticated multivariate regression model, results are obtained that closely match those obtained in listening tests.

A spatial audio quality meter has many applications in audio engineering. These include possible uses in automatic system alignment, evaluation of alternative rendering formats, consumer system optimisation and codec evaluation. As the range of spatial qualities available from fixed and mobile rendering platforms becomes increasingly wide, and now that scalable spatial audio coding is a reality, a means of predicting perceived spatial quality that does not involve lengthy listening tests is highly desirable.

2. BACKGROUND

It is desirable to be able to evaluate the perceived spatial quality of audio processing, coding-decoding (codec) and reproduction systems without needing to involve human listeners. This is because listening tests involving human listeners are time consuming and expensive to run. It is important to be able to gather data about perceived spatial audio quality in order to assist in product development, system setup, quality control or alignment, for example.

Spatial quality evaluation is becoming increasingly important as manufacturers and service providers attempt to deliver enhanced user experiences of spatial immersion and directionality in audio-visual applications. Examples are virtual reality, telepresence, home entertainment, automotive audio, games and communications products. Mobile and telecommunications companies are increasingly interested in the spatial aspect of product sound

quality. Here simple stereophony over two loudspeakers, or headphones connected to a mobile player, is increasingly typical. Binaural spatial audio is likely to become a common feature in mobile devices. Home entertainment involving multichannel surround sound is one of the largest growth areas in consumer electronics, bringing enhanced spatial sound quality into a large number of homes. Home computer systems are increasingly equipped with surround sound replay and recent multimedia players incorporate multichannel surround sound streaming capabilities, for example. Scalable audio coding systems involving multiple data rate delivery mechanisms (e.g. digital broadcasting, internet, mobile communications) enable spatial audio content to be authored once but replayed in many different forms. The range of spatial qualities that may be delivered to the listener will therefore be wide and degradations in spatial quality may be encountered, particularly under the most band-limited delivery conditions or with basic rendering devices.

Systems that record, process or reproduce audio can give rise to spatial changes including the following: changes in individual sound source-related attributes such as perceived location, width, distance and stability; changes in diffuse or environment related attributes such as envelopment, spaciousness and environment width or depth. In order to be able to analyse the reasons for overall spatial quality changes in audio signals it may also be desirable to be able to predict these individual sub-attributes of spatial quality. There is also a need for a global or holistic grading of spatial quality that weights the importance of these different factors appropriately for the context or task in question.

Under conditions of extreme restriction in delivery bandwidth, major changes in spatial resolution or dimensionality may be experienced (e.g. when downmixing from many loudspeaker channels to one or two). Recent experiments involving multivariate analysis of audio quality show that in home entertainment applications spatial quality accounts for a significant proportion of the overall quality. In one study reported by Rumsey *et al* this proportion was found to be approximately 30% [2].

Because listening tests are expensive and time consuming, there is a need for a quality model that is capable of predicting perceived spatial quality on the basis of measured features of audio signals. Such a model needs to be based on a detailed analysis of

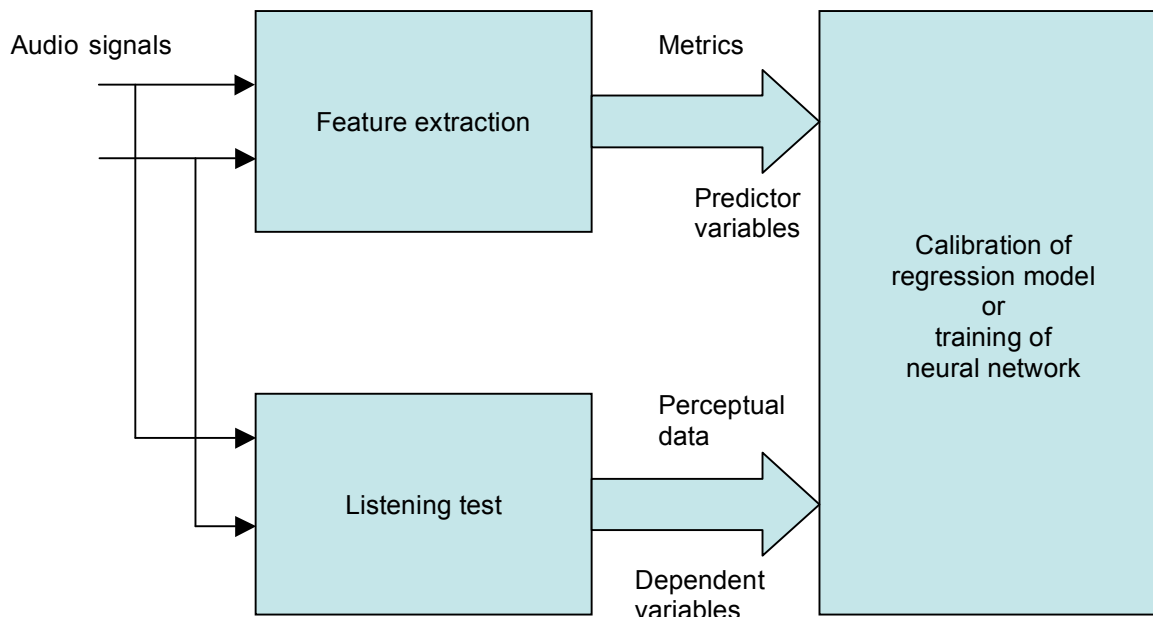


Figure 1 Generic principle of quality prediction model calibration

human listeners' responses to spatially altered audio material, so that the results generated by the model match closely those that would be given by human listeners when listening to typical programme material. The model may optionally take into account the acoustical characteristics of the reproducing space and its effects on perceived spatial fidelity, either using acoustical measurements made in real spaces or using acoustical simulations.

3. GENERIC PRINCIPLES

The QESTRAL model adopts a similar generic principle to other perceptual quality prediction models in that it extracts a number of physical features, by means of measurement, from one or more audio signals, as shown in Figure 1. From these are derived a number of perceptually motivated low-level metrics, some of which are also used to derive higher-level metrics relating to spatial features or distortions of the reproduced sound scene. In a separate process, perceptual ratings of the spatial quality of the reproduced audio signals are obtained and used to calibrate a statistical model or neural network, in such a way that an appropriate weighting and combination of the metrics is determined, which enables the output of the predictor to match the results of the listening tests as closely as possible.

4. INTRUSIVE AND UNINTRUSIVE MEASUREMENT OF QUALITY

When measuring perceived quality the question of appropriate reference conditions must be addressed. In most extant audio quality prediction models the quality scale is calibrated against an unimpaired reference signal, with the assumption that signals having the same perceived quality as the reference signal will be graded at the top of the scale. Essentially these are impairment models of audio quality and there is the implicit or explicit assumption that any changes to the perceived characteristics of the reproduced signal, compared with the reference, are to be considered as impairments, making the quality poorer. The assumption is that the reference is 'correct' and anything else is to a greater or lesser degree 'incorrect'. It is not possible to rate any alternative versions of the reproduction higher up the scale or 'better' than the reference. Such quality scales usually include a strong implied hedonic component, and are often labelled with hedonic terms such as 'good' or 'bad'. This is partly because sound quality is a high level construct that is hard to define in absolute terms and much easier to define in relative terms. The bottom end of these scales tends to float,

depending on the range of qualities implicit in the stimuli presented and the nature of any anchor stimuli present, as shown by Zielinski *et al* [3]. However gradings on such scales are capable of showing at least the rank order of stimuli quality, and a guide to the magnitude of the relationships between them.

The listening tests used to calibrate such models require listeners to compare the sound of impaired stimuli with a reference stimulus. When building perceptual quality prediction models that aim to emulate the results of listening tests of this type, an intrusive approach is usually adopted whereby measurements of both reference and impaired versions of the signal are made and a number of perceptually motivated metrics used to derive a comparative quality grade or difference grade. This is the primary approach adopted in the QESTRAL model when evaluating holistic or global spatial quality in terms of an ‘opinion score’. However the use of alternative model calibrations allowing bidirectional or floating quality scales is not precluded.

Single-ended or unintrusive evaluation is sometimes possible when evaluating individual attributes of spatial quality. Such attributes can be defined in simpler terms, are more likely to be of a perceptually

unidimensional nature, and metrics can be calibrated against known anchor points. Examples might include measurements of perceived location, width or envelopment. One example of this, arising from the QESTRAL project, is an envelopment measurement algorithm that can operate in unintrusive fashion, enabling the prediction of perceived envelopment for any five-channel programme material. This is described in a separate paper by George *et al* [4].

5. SPECIFIC PRINCIPLES OF THE QESTRAL MODEL

An outline of the key principles of the QESTRAL model is given here. Further details are given in a separate paper by Jackson *et al* [5].

5.1. System concept

The QESTRAL model was developed to be independent of the reproduction format of spatial audio content. In other words it is intended to work with an arbitrary layout of loudspeakers or headphones, although in a prototype demonstrator version it adopts a reference format according to the ITU BS. 775 3-2 stereo format. In order to achieve this it relies primarily on measurements of the reproduced sound field made at one or more listening

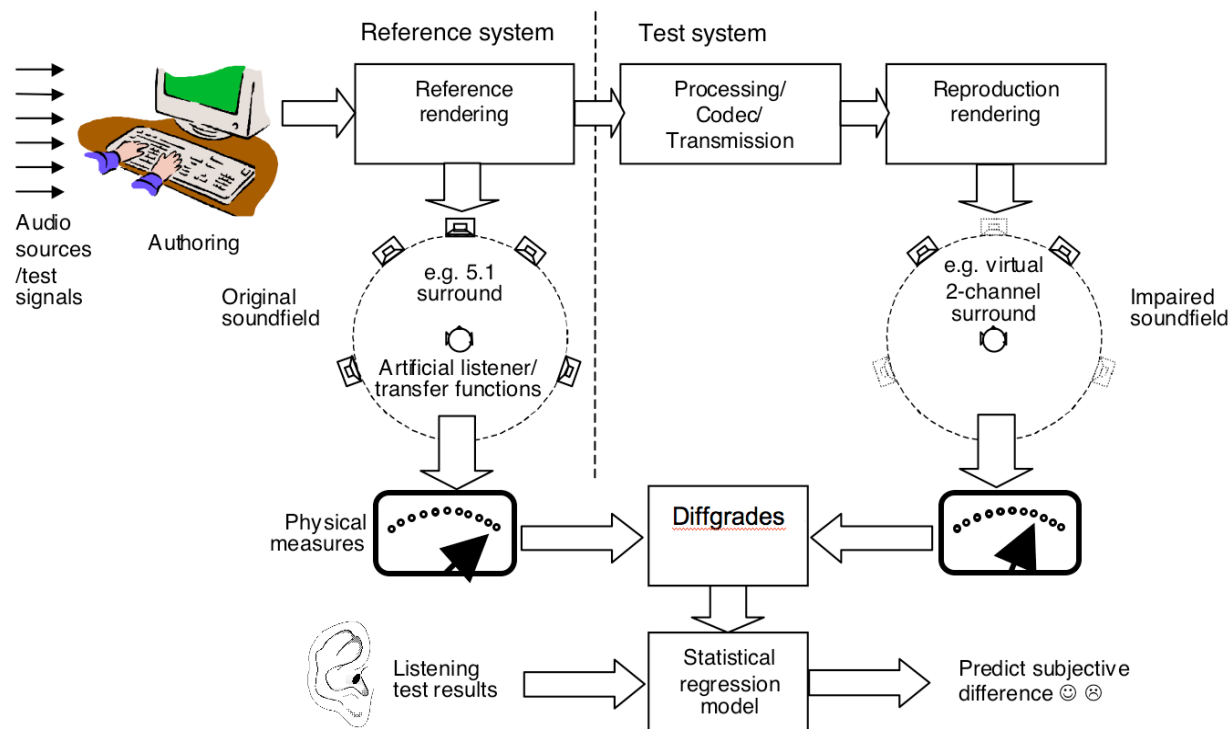


Figure 2 Conceptual diagram of the QESTRAL system

positions, using binaural and other microphone-derived signals hereafter referred to as probes. An important feature of the model is that it can incorporate an acoustical simulation of the reproduced sound field and can measure the spatial quality at any listening position within it. This enables the effects of reflections within the sound field to be incorporated. A conceptual diagram of the approach is shown in Figure 2, and a flow diagram of the processes involved is shown in Figure 3.

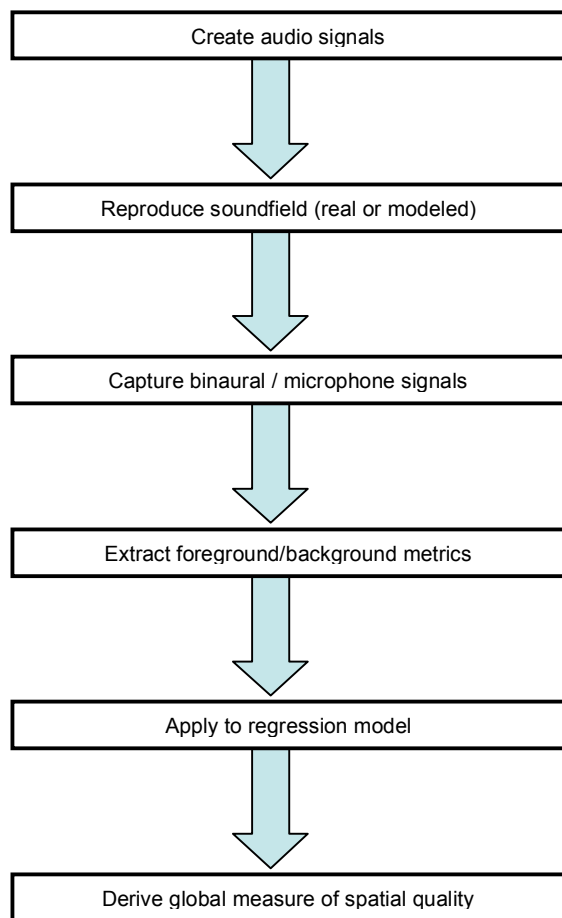


Figure 3 Sequence of processes involved in measuring perceived spatial quality

The device under test (DUT) can be any spatial audio processing device and can include alternative rendering methods. In other words, the model is designed to evaluate the change in spatial quality that results from electrical or acoustical alteration of the spatial characteristics of reproduced sound scenes, or

some combination of the two. Examples of primarily electrical alteration include such devices as audio codecs and downmixers, whereas examples of acoustical alteration includes changes of loudspeaker position, directivity and room reflection characteristics. Combination scenarios could include changes of spatial rendering format such as the comparison of a five-channel loudspeaker surround reference reproduction with a virtual surround equivalent, rendered transaurally over two loudspeakers, perhaps in a different room.

5.2. Probe signals and metrics

In order to measure global or holistic changes in perceived spatial quality a number of perceptually motivated metrics are employed, based on signals derived from the probes. These are supplied to a statistical regression model that has been calibrated using a large database of listening test scores that describe the subjective ratings of many types of spatially impaired reproduction. The current version of the model makes use of specially designed probe signals (also called test signals) that stress the performance of the spatial audio process concerned. These probe signals aim to emulate the generic features of the spatial components of typical programme material. The algorithms and metrics employed in the model are designed to work specifically with these probe signals, but the regression model based on these metrics is calibrated using the results of real listening tests. Therefore the predicted results are very similar to those given in listening tests that used programme material as a stimulus. The advantage of using specially designed probe signals is that they have known signal characteristics (e.g. known source locations), and a reference version can be easily compared against an impaired version. The difficulty of using such specially tailored probe signals and metrics is ensuring that one has appropriate and sufficient signals and related metrics to account for all of the spatial attributes and quality variations encountered in real programme material and DUTs. The use of programme material as a probe signal is not precluded, although this requires more sophisticated automatic scene analysis processing in the case of some metrics.

Conceptually the evaluation of spatial quality can be divided into foreground and background scene components, along the lines of the scene-based paradigm for spatial quality evaluation proposed by

Rumsey [6]. The foreground components consist of the localisable objects (sources) in the scene whereas the background components consist of the diffuse or environment-related aspects of the scene. These are not always easy to distinguish exactly and a number of spatial attributes result from contributions of both these components. Foreground components, when processed by a DUT that affects spatial quality, typically suffer changes in location-related attributes, whereas background components typically suffer changes in attributes such as envelopment and spaciousness.

In order to attempt a scene-based evaluation of spatial quality, the QESTRAL model aims to utilise probe signals, probes and metrics that respond to both foreground and background components of the reproduced sound scene. A source localisation model is incorporated into the system that is capable of measuring changes in the foreground scene, and this is partnered by a set of metrics that aims to measure changes in background envelopment and spaciousness (related to subjective diffuseness). In one current prototype implementation two probe signals are employed, the first being a point source panned to discrete locations around the listening position, and the second being an uncorrelated noise signal fed to all channels simultaneously. These are used in conjunction with foreground and background metrics respectively.

5.3. Calibration listening tests

In order to ensure that the QESTRAL model would be capable of predicting accurately a very wide range of spatial audio quality changes, it was calibrated using a large database of listening tests. It was vital to ensure that this database represented the perceived quality changes arising from numerous commonly encountered audio processes, and a rigorous selection method was adopted to ensure that the results represented a wide range of programme material genres. Care was also taken to ensure that the perceptual attribute space in terms of relevant spatial features and the magnitude of their changes was adequately spanned by the listening test stimuli. Selected hidden spatial anchor stimuli were included in order to evaluate the repeatability of gradings and the relationship between test stimuli and the scale in different iterations of the calibration procedure. This work is described in greater detail in a separate paper by Conetta *et al* [7].

6. EXAMPLES OF RESULTS

Initial results from the QESTRAL model are promising and suggest that it is possible to predict a holistic or global spatial quality measure that represents a form of spatial mean opinion score or ‘S-MOS’. This essentially describes both the audibility/magnitude of change in the spatial domain, when comparing the reference and impaired stimulus, and the degree of annoyance or displeasure associated with any change. In this way the scale employed exhibits similar conceptual characteristics to the basic audio quality scale, or MOS scale, employed in other audio quality tests, which also conflates an evaluation of the perceived magnitude of the difference between the impaired and reference stimuli with a judgment about the subjective acceptability of the same [8].

Degradation	Description	Listening Position
1	Ls and Rs are positioned at -90° and 90°	1
2	L and R are positioned at -10° and 10°	1
3	Ls turned-off	1
4	1.0 downmix in all channels	1
5	1.0 downmix to C	1
6	Unimpaired	1
7	2.0 downmix to L and R	1
8	1.0 downmix to Ls	1
9	Ls and Rs are positioned at -90° and 90°	2
10	L and R are positioned at -10° and 10°	2
11	Ls turned-off	2
12	1.0 downmix in all channels	2
13	1.0 downmix to C	2
14	Unimpaired	2
15	2.0 downmix to L and R	2
16	1.0 downmix to Ls	2

Table 1 Quality degradations used in prediction example

Detailed examples of model calibration and results from selected predictions are provided in a separate paper by Dewhirst *et al* [9]. A simple early example is given here, showing the prediction of the spatial quality of a limited set of stimuli that had been impaired by different forms of downmixing, missing channels and loudspeaker position changes, as shown in Table 1.

The 3-2 stereo format is used as the reference condition. These stimuli were evaluated subjectively in two separate listening positions, shown in Figure 4, and predictions were also made in these positions. Two simple probe signals were employed in the QESTRAL model, namely a panned pink noise burst and a decorrelated pink noise signal in all channels. The graph in Figure 5 shows the predicted versus actual scores for spatial quality, exhibiting a high correlation between them of 0.9, together with an RMS prediction error of 14.9 on a 100-point scale.

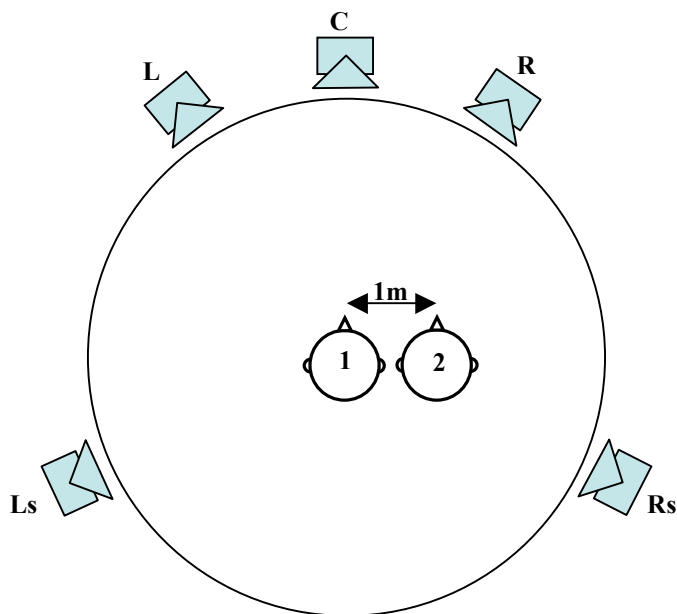


Figure 4 Listening positions and loudspeaker layout for prediction example

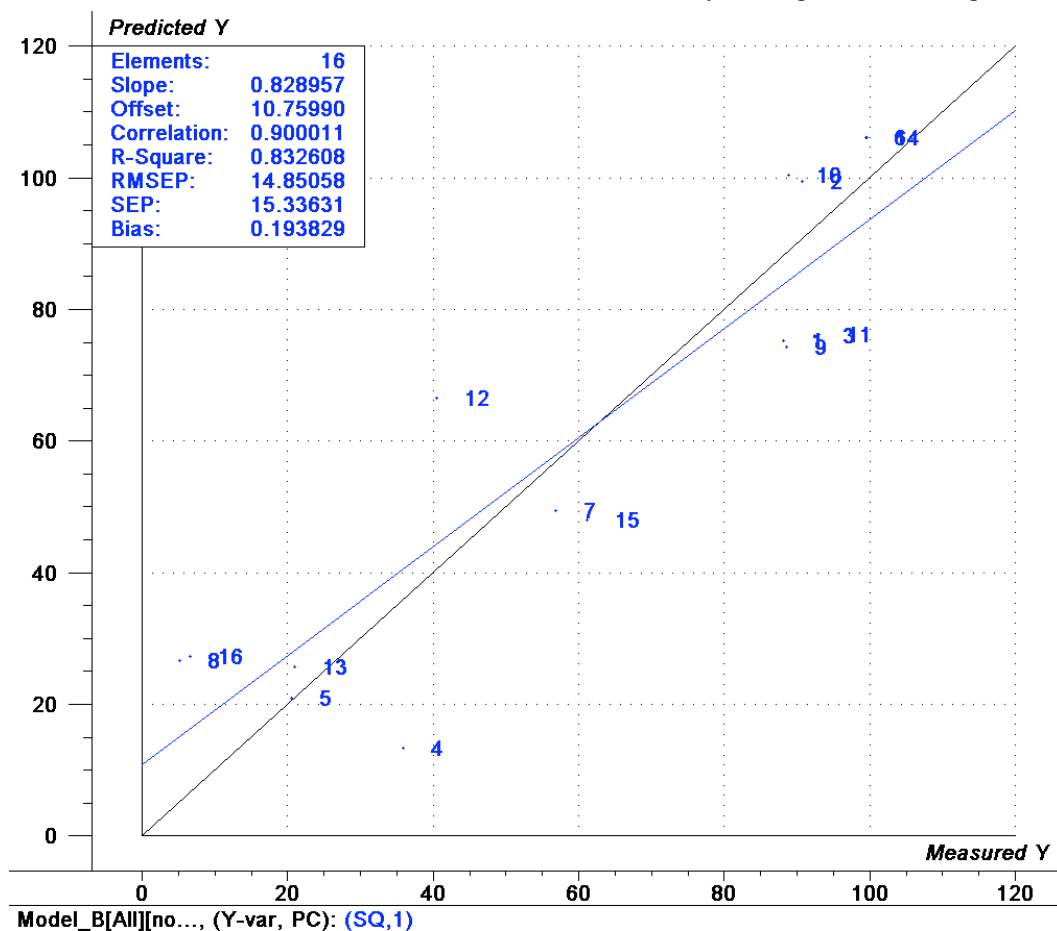


Figure 5 Results of prediction showing output of model (predicted Y) against results of listening test (measured Y)

7. FURTHER WORK

Although the QESTRAL model can be shown to perform reasonably well when predicting relatively simple spatial quality changes, the current challenge is to improve its performance on a range of more sophisticated and subtle spatial quality changes. This includes the development of more appropriate or sensitive metrics and probe signals, and the calibration of the model with a wider range of programme material and DUTs. It may also be necessary to introduce context and content dependency to the model so as to enable predictions that more accurately represent the perceived quality for different listening scenarios, tasks and content types.

8. REFERENCES

- [1] ITU-R BS.1387-1, Method for Objective Measurement of Perceived Audio Quality, International Telecommunication Union, Geneva, Switzerland (1999).
- [2] Rumsey, F., Zielinski, S., Kassier, R. & Bech, S. (2005) On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *J. Acoust. Soc. Amer.*, 118 (2), 968–977, August
- [3] Zielinski, S., Rumsey, F. and Bech, S. (2008) On some biases encountered in modern listening tests – a review. *J. Audio Eng. Soc.* 56 (6), June
- [4] George, S. et al (2008) An unintrusive objective model for predicting the sensation of envelopment arising from surround sound recordings. Presented at 125th AES Convention, San Francisco, Oct 2–5. Audio Engineering Society
- [5] Jackson, P. et al (2008) QESTRAL (Part 3): system and metrics for spatial quality prediction. Presented at 125th AES Convention, San Francisco, Oct 2–5. Audio Engineering Society
- [6] Rumsey F. (2002) Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm *J. Audio Eng. Soc.*, 50 (9), September
- [7] Conetta, R. et al (2008) QESTRAL (Part 2): Calibrating the QESTRAL spatial quality model using listening test data. Presented at 125th AES Convention, San Francisco, Oct 2–5. Audio Engineering Society
- [8] ITU-R BS.1116-1 (1993) Recommendation: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. International Telecommunication Union, Geneva
- [9] Dewhurst, M. et al (2008) QESTRAL (Part 4): Test signals, combining metrics and the prediction of overall spatial quality. Presented at 125th AES Convention, San Francisco, Oct 2–5. Audio Engineering Society