Audio Engineering Society

Convention Paper

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

QESTRAL (Part 4): Test signals, combining metrics and the prediction of overall spatial quality

Martin Dewhirst[1], Robert Conetta[1], Francis Rumsey[1], Philip Jackson[1], Slawomir Zielinski[1], Sunish George[1], Søren Bech[2], and David Meares[3]

[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, on behalf of BBC Research*

Correspondence should be addressed to Martin Dewhirst (Martin.Dewhirst@surrey.ac.uk)

ABSTRACT

The QESTRAL project has developed an artificial listener that compares the perceived quality of a spatial audio reproduction to a reference reproduction. Test signals designed to identify distortions in both the foreground and background audio streams are created for both the reference and the impaired reproduction systems. Metrics are calculated from these test signals and are then combined using a regression model to give a measure of the overall perceived spatial quality of the impaired reproduction compared to the reference reproduction. The results of the model are shown to match closely the results obtained in listening tests. Consequently, the model can be used as an alternative to listening tests when evaluating the perceived spatial quality of a given reproduction system, thus saving time and expense.

1. INTRODUCTION

The QESTRAL project is developing an artificial listener to compare the perceived spatial quality of an audio reproduction to a reference reproduction. The QESTRAL model is designed to evaluate changes in the spatial quality of reproduced audio, rather than changes in the timbral quality. This contrasts with previous models of audio quality, such as PEAQ (ITU-R BS1387) [16], which have not explicitly considered the spatial distortions of different reproduction systems and processes.

The previous QESTRAL papers have already given an overview of the entire model [13], a description of the listening tests that were undertaken in order to provide data to calibrate the model [5] and a more detailed description of the different components in

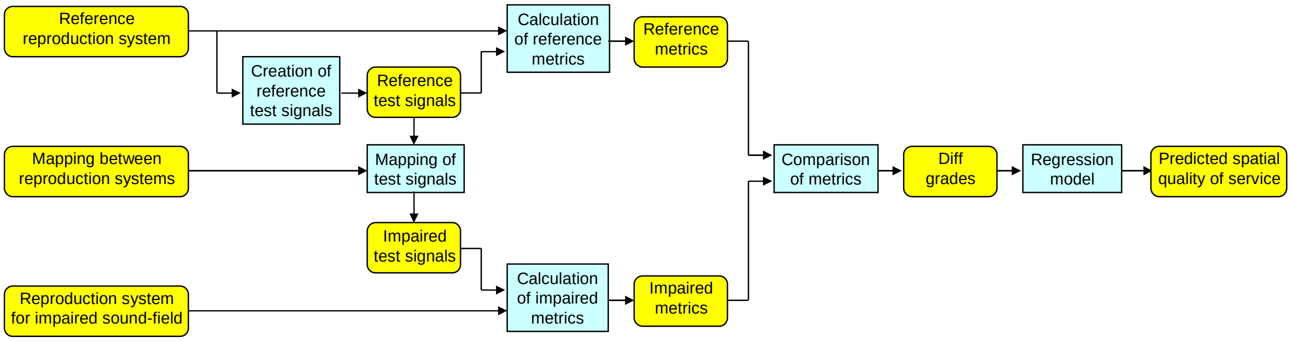
Reference reproduction system
Creation of reference test signals
Reference test signals
Calculation of reference metrics
Reference metrics
Mapping between reproduction systems
Mapping of test signals
Impaired test signals
Reproduction system for impaired sound-field
Calculation of impaired metrics
Impaired metrics
Comparison of metrics
Diff grades
Regression model
Predicted spatial quality of service

Fig. 1: *Flow diagram showing the processes in the QESTRAL model, including the use of the test signals. The yellow rounded boxes show data, while the blue rectangular boxes show processes.*

the model, including the measures calculated from binaural and microphone signals [8]. This paper describes the test signals used by the model, how these test signals are incorporated into the model framework, and finally how the metrics calculated from the test signals are combined into regression models to predict the spatial quality.

2. MODEL OVERVIEW

The data input to the model can be divided into three parts. The first of these consists of a description of the reproduction system and acoustic environment used in the creation of the reference sound-field. This includes the location and orientation of the different loudspeakers used in the reproduction system and the location and orientation of the listener within the sound-field. The second group of input data consists of a description of the reproduction system and acoustic environment used in the creation of the impaired sound-field. The third group of input data is a process that maps the signals for the reference reproduction system to the signals for the system used for the impaired sound-field. The second and third groups of input data together comprise a description of the device under test (DUT).

An overview of the model was given in the QESTRAL (Part 1) paper [13]. Fig. 1 shows the structure of the QESTRAL model, including how the generation of the test signals is incorporated into the model. The model has four main stages. The first stage is the creation of the test signals, described in Section 3. The second stage comprises the calculation of metrics. This stage has two parts: first binaural signals and microphone signals at the listener location are calculated, which are then used to calculate the different metrics. This is described in detail in the QESTRAL (Part 3) paper [8]. The output of this stage is a pair of values for each metric; in each pair, one value corresponds to the reference sound-field and the other value corresponds to the impaired sound-field. In the third stage of the model the impaired sound-field value is subtracted from the reference sound-field value for each metric, resulting in a *diff grade* for each metric. The fourth stage of the model consists of applying a regression model on the calculated diff grades to give a single value of the predicted quality of service. This final stage is discussed in Section 4.

3. TEST SIGNALS

Listeners perceive a sequence of notes from a musical instrument or the sequence of phones from speech as auditory streams [3]. When more than one musical instrument or speaker is present then the listener assigns each sequence to a separate auditory stream. In addition to these foreground perceptual streams, there are also sounds which are not easily assigned by the listener to any single distinct auditory stream, for example, the reflected sound from a reverberant room. These can be grouped together as the background auditory stream [7].

The model generates two groups of test signals appropriate to the reference reproduction system. The first group consists of signals designed to identify any distortions in the foreground audio stream, while the second group consists of signals designed to identify any distortions in the background audio stream.

3.1. Foreground stream

The most obvious perceived spatial distortions in the foreground stream will be changes in the perceived location of sources. However, other types of perceived spatial distortion related to foreground objects may also be present, such as changes in individual source width, ensemble width, source stability and source focus [12].

The test signals designed to allow the model to evaluate the distortions in the foreground stream consist of thirty-six one second pink noise bursts, positioned at 10° intervals in the horizontal plane. These test signals are termed the “spun noise” test signals. The test signals only have to be specified for the reference reproduction system, as the test signals for the reproduction system for the impaired sound-field are created using the process included in the DUT. Therefore, the method of positioning the thirty-six noise bursts depends on the reference reproduction system. Where the reference reproduction system consists only of a loudspeaker arrangement and does not include a specific method of creating the loudspeaker signals for the reproduction system (for example, the five channel loudspeaker setup specified in ITU-R BS.775-1 Recommendation [11]), then each of the noise bursts is positioned using pair-wise constant power panning. Where the reference reproduction system does include a specific method of creating the loudspeaker signals, such as higher order ambisonics or wave field synthesis [6], then each noise burst is positioned according to that method.

The regression models described in this paper were created using a reference reproduction system consisting of the ITU standard 5 channel loudspeaker setup with no specific method of creating the loudspeaker signals. Hence, the foreground stream test signals for the models used to generate the results in this paper were created by pair-wise constant power panning the a one second noise burst to the thirty-six different equally spaced angles.

3.2. Background stream

The test signals designed to allow the model to evaluate the perceived spatial distortions in the background stream consist of a 10 second burst of decorrelated pink noise played through all the channels in the reference reproduction system. Each channel in the reference reproduction system has a signal consisting of a 10 second burst of pink noise, where the pink noise burst in each channel is decorrelated in relation to the pink noise bursts in the other channels. The pink noise signals in all the loudspeaker channels are played simultaneously. This contrasts with the test signal for the foreground stream described above, where the pink noise bursts were played sequentially.

This test signal is designed to approximate some of the signal characteristics of a diffuse acoustic field, such as can arise as a result of late reflections from a reverberant acoustic environment. The presence of late reflections is one of the most typical situations in which a background auditory stream can arise and Bradley and Soulodre [2] have shown that the relative level of the late arriving lateral sound energy was highly correlated to the sense of listener envelopment. Consequently, listener envelopment is an important spatial attribute associated with the background auditory stream. As in the case of the foreground auditory stream, other types of perceived spatial distortion may also be present in the background streams, for instance the scene depth and the scene width [12]. While the use of decorrelated pink noise as the sole test signal for the background stream may not be suitable for evaluating all of these spatial distortions, high correlations between the QESTRAL model and listening test data were obtained using only this test signal for the background stream.

4. CALIBRATION OF MODEL AND PREDICTION OF LISTENING TEST RESULTS

The results of the listening tests described in the QESTRAL (Part 2) paper [5] were used to calibrate regression models using the metrics described in the QESTRAL (Part 3) paper [8]. The models were calibrated using partial least squares (PLS) regression [17] using the Unscrambler v9.6 software. The

Metric name	SE(B)	B
IACC0	0.40	94.86
EntropyL	-0.34	-39.42
Mean_Ang_Diff	-0.40	-0.26
Constant	3.51	88.51

Table 1: *The coefficients of the regression model fitted to the listening test results for the centre listening position. The second and third columns contain the standardised and raw coefficients respectively.*

Metric name	SE(B)	B
IACC0	0.25	76.36
IACC0*IACC90	0.24	52.15
Max_Ang_Diff	-0.23	-0.10
Mean_Ang_Diff	-0.28	-0.20
Constant	3.45	94.13

Table 2: *The coefficients of the regression model fitted to the listening test results for the right (off-centre) listening position. The second and third columns contain the standardised and raw coefficients respectively.*

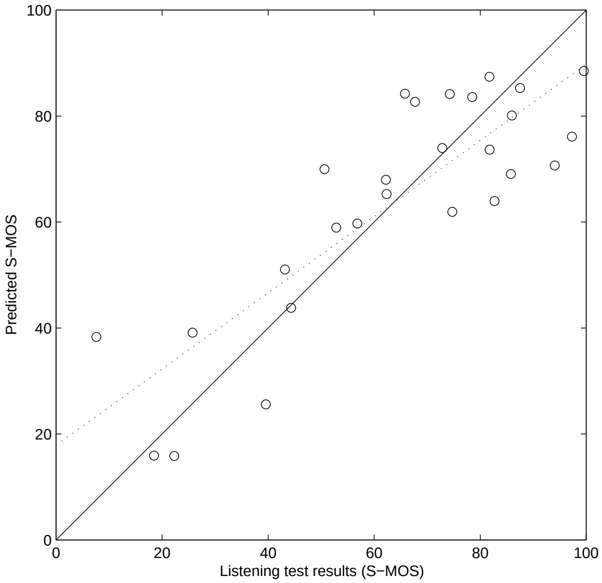

Fig. 2: *Results of the cross validation of the regression model calibrated for the centre listening position. The solid line shows the ideal relationship and the dotted line shows the line of best fit.*

process of arriving at the final regression models was an iterative process, beginning with all the metrics described in the QESTRAL (Part 3) paper and altering the selection of metrics used in the model based on the analysis provided by the Unscrambler software. The resulting regression models were then cross-validated using the listening test results.

Two different listener positions were used in these listening tests: one at the sweet spot at the centre of the listening area, and the second one metre to the right of the sweet spot. These will be referred to as the *centre* and *right* listening positions respectively. Note that for a given set of signals fed to the loudspeakers, the binaural signals at the two listener positions are extremely likely to differ from each other. This is also true of the simulated microphone signals used in the calculation of some of the metrics [8]. As the binaural signals at the two listener positions differ from each other for the same reference sound-field it follows that the perception of the listener is different at each listener position. In the results from the listening tests, the top of the scale corresponded to the reference sound-field (i.e. a stimulus graded at the top of the scale has the same spatial perception as the reference sound-field). Therefore, the listening test results for the centre listening position have a different scale than the listening test results for the right listening position. Consequently, separate regression models were created for each of the two listening positions.

Table 1 shows the coefficients of the regression model using the metrics described in [8] calibrated using the listening test results for the centre listening position. A leave-one-out cross-validation [15] was performed on the the model, the results of which are shown in Fig. 2, which resulted in a correlation of 0.82 and a root-mean-square error of prediction (RMSEP) of 14%. The horizontal axis shows the Spatial Mean Opinion Score (S-MOS) results from the listening tests, while the vertical axis shows the predicted S-MOS results. Similarly, the coefficients for the regression model calibrated using the listening test results for the right listening position are given in Table 2, and Fig. 3 shows the results of the

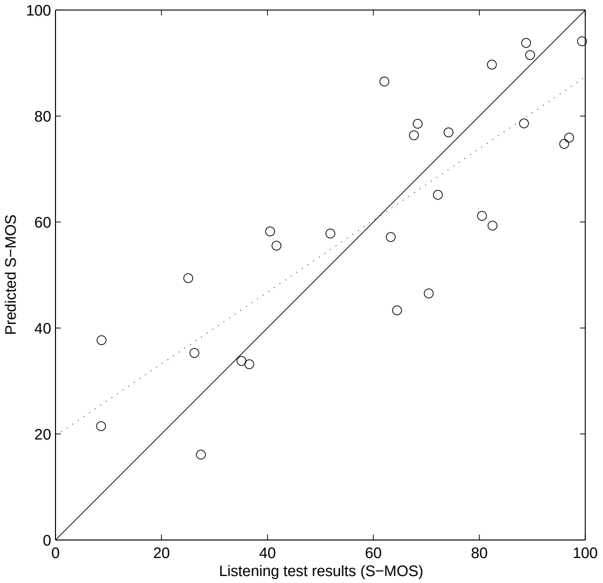

Fig. 3: *Results of the cross validation of the regression model calibrated for the right listening position.*

corresponding cross-validation. The cross-validation for the second model resulted in a correlation of 0.79 and a RMSEP value of 17%.

4.1. Discussion of metrics used in the regression models

While the two regression models both contain the IACC0 and Mean_Ang_Diff metrics, the remaining metrics differ between the two models. The IACC0 metric is the interaural cross-correlation (IACC) for the listener facing forwards, and this measure has been associated with both envelopment [1] and width [9, 10]. Hence the inclusion of the IACC0 metric in the models is expected, as the spatial quality of a DUT depends partly on its ability to recreate both of these two perceived spatial attributes.

The IACC90 metric consists of the IACC for the listener facing 90° to the right. While the IACC90 metric was not selected for the final regression model for either of the two listening positions, the metric IACC0*IACC90 (the product of the two IACC metrics) is used for the model for the right listening position. One possible explanation is that the ability of a DUT to reproduce sources or reflections from all directions affects the perceived spatial quality, which relates to the diffuseness of the reproduced soundfield.

The Mean_Ang_Diff and Max_Ang_Diff metrics are both measures of the change in the foreground source locations caused by the DUT. This affects the perceived spatial quality, which is shown by the Mean_Ang_Diff metric being in both regression models and also the Max_Ang_Diff metric being in the regression model for the right listening position.

The TotEnergy and EntropyL metrics were calculated for the test signals because these were found to be useful for predicting perceived envelopment in an earlier study [4]. However, the listening tests in this earlier study differed substantially from the listening tests described in the QESTRAL (Part 2) paper. The earlier study into envelopment included stimuli generated from different source material, quite apart from differences in the positioning of the original sources and any spatial processing. For example, one stimulus consisted of a single anechoic voice from the centre loudspeaker, while another stimulus consisted of eight different anechoic voices panned to different locations. One of the other stimuli consisted of a five-channel music recording from a commercially available DVD. The test subjects used the same scale for their responses to all the stimuli.

In contrast, the listening tests described in the QESTRAL (Part 2) paper were principally concerned with the effects on the perceived spatial quality of different impairments to the original five-channel signals. As none of the impairments considered in this listening test caused large changes to the source material, it was not expected that the metrics introduced to account for these changes (TotEnergy and particularly EntropyL) would appear in the regression models. However, the EntropyL metric was selected by the Unscrambler for the regression model for the centre listener position, and this is discussed in the next section.

4.2. The centre listening position and the EntropyL metric

The inclusion of the EntropyL metric in the regression model for the centre listening position is initially surprising. The EntropyL metric is the Shannon entropy [14] calculated for the left ear signal from the binaural signals at the listening position. The entropy quantifies the information contained in this signal. As stated previously, this was introduced in an earlier study of the perceived envelopment arising from being surrounded by different voices [4]. From inspecting the listening test results in this study it was found that the number of different voices had a positive correlation with the envelopment scores elicited from the test subjects when the other aspects of the stimuli were kept constant. As the amount of information in the binaural signals increases as the number of voices increases, entropy was introduced as a metric. For the stimuli used in this study there was not a significant difference between the values of entropy calculated from the left or right ear signals, and so an arbitrary decision was made to use the left ear signal in the calculation of the entropy metric.

For the regression models shown in this paper, the EntropyL metric was calculated using the decorrelated pink noise test signal. The information content of the noise signals is much larger than that of the stimuli based on voices in the previous study into envelopment. The values of EntropyL are calculated from the left ear signal for the test signal once it has been subjected to the different degradations. Note that none of these processes will result in a change in the information content of the signals of the same order of magnitude as the changes which were seen from changing the number of voices in the previous envelopment study. However, smaller changes to the information content of the left ear signal do arise due to filtering effects of the ear and head.

Fig. 4 shows the values of the EntropyL metric plotted against the S-MOS results from the listening test for the centre listening position. Inspection of the values of the EntropyL metric calculated for the different DUTs showed that most of these values fell within a narrow range (78% of the DUTs had values of EntropyL in the range -1.16 to -1.4). The most prominent EntropyL value outside this range was -2.1, corresponding to process 31. This degradation consisted of a down-mix to 2.0 followed by the two resulting channels being randomly allocated to two of the five loudspeakers in the five-channel setup (namely the front right and right surround loudspeakers). Thus, for process 31, all the active loudspeakers were to the right of the listener's head. The left ear signal is used for the calculation of the EntropyL metric, which is subject to the filtering due to the pinna and the shadowing of the head. This results in the frequency response of the signal at the left ear being much less flat, so there is an increase in the periodicity of the left ear signal and consequently this signal has less information content and so has a lower value of entropy.

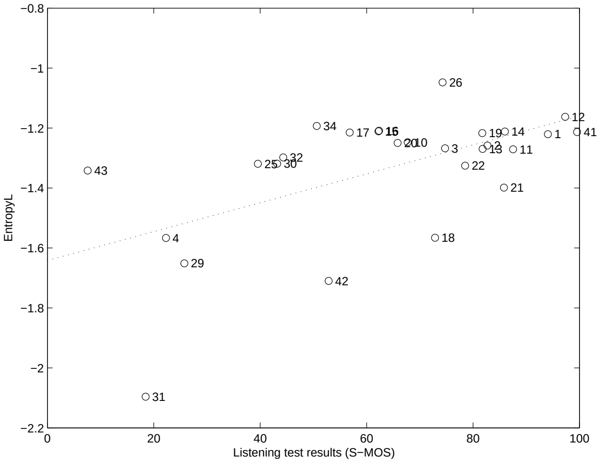

Fig. 4: *The values of the EntropyL metric plotted against the S-MOS results from the listening test for the centre listening position. The numbers on the graph show the different DUTs used in the listening tests (see [5]).*

It can be seen from Fig. 4 that the listening test scores for the centre listening position are mainly in the upper half of the scale (74% of the listening test scores were above 50). It can also be seen from the figure that the range of calculated values for the EntropyL metric is much larger for the points in the lower half of the listening test scale than it is for the points in the upper half of the scale. Consequently, the few DUTs with low listening test scores have a greater influence on the fitting of the regression model. As the EntropyL metric is able to differentiate between these DUTs, this metric improves the fit of the regression model, which is why it appears in the regression model calculated by the Unscrambler.

5. FUTURE WORK

The spun noise test signals and the decorrelated pink noise signals are a first attempt at developing test signals which allow the QESTRAL model to evaluate changes in spatial quality. Consequently, these may not be the most suitable test signals to expose the changes for all the perceived spatial attributes affecting the spatial quality. One area of future work for the QESTRAL project therefore involves both refining the existing test signals and, if necessary, designing additional test signals. This needs to be done in parallel with improvements to the existing metrics and also the design of new metrics. This is motivated not only by the need to improve the performance of the existing models, but also so that the model can predict the effect on perceived spatial quality of other common impairments, such as MPEG codecs, which have not so far been included in the regression models.

One of the motivations for having metrics which use only either binaural signals or microphone signals at the listener position was that they would not be confined to the sweet spot or to using a single loudspeaker setup. Currently, two different listener positions have been considered and this has led to two different regression models. Ideally the same model would be used by any position in the listening area, so one area of future work is to combine the different models from the different listener locations into a single model.

6. REFERENCES

[1] J.S. Bradley and G.A. Soulodre. Objective Measures of Listener Envelopment. *J. Acoust. Soc. Am.*, 98(5):2590–2597, November 1995.

[2] J.S. Bradley and G.A. Soulodre. The Influence of Late Arriving Energy in Spatial Impression. *J. Acoust. Soc. Am.*, 97(4):2263–2271, April 1995.

[3] A.S. Bregman. *Auditory scene analysis: the perceptual organistaion of sound*. MIT Press, Cambridge, Massachusetts, 1990.

[4] R. Conetta. SQoS 25: Direct envelopment scaling and prediction: Part 1. Technical report, QESTRAL project internal document, Guildford, UK, April 2007. QESTRAL project website: `http://www.surrey.ac.uk/soundrec/QESTRAL/`.

[5] R. Conetta *et al*. QESTRAL (Part 2): Calibrating the QESTRAL spatial quality model using listening test data. Presented at the *125th AES Convention*, San Francisco, October 2008. Audio Engineering Society.

[6] J. Daniel, R. Nicol, and S. Moreau. Further Investigations of High Order Ambisonics and Wavefield Synthesis of Holophonic Sound Imaging. Amsterdam, The Netherlands, March 2003, 114th Conv. Audio Eng. Soc., Preprint 5788.

[7] D. Griesinger. Objective Measures of Spaciousness and Envelopment. In *Proceedings of the AES 16th International Conference*, Rovaniemi, Finland, April 1999.

[8] P. Jackson *et al*. QESTRAL (Part 3): System and metrics for spatial quality prediction. Presented at the *125th AES Convention*, San Francisco, October 2008. Audio Engineering Society.

[9] R. Mason, T. Brookes, and F. Rumsey. Integration of measurements of interaural cross-correlation coefficient and interaural time difference within a single model of perceived source width, October 2004, 117th Conv. Audio Eng. Soc., Preprint 6317.

[10] T. Okano, L.L. Beranek, and T. Hidaka. Relations among interaural cross-correlation coefficient (IACC_E), lateral fraction (LF_E), and apparent source width (ASW) in concert halls. *J. Acoust. Soc. Am.*, 104(1):255–265, July 1998.

[11] Rec. ITU-R BS.775-1. Multichannel stereophonic sound system with and without acompanying picture, 1994.

[12] F. Rumsey. Spatial Quality Evaluation For Reproduced Sound: Terminology, Meaning and a Scene-Based Paradigm. *J. Audio Eng. Soc.*, 50(8):651–666, September 2002.

[13] F. Rumsey *et al*. QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener. Presented at the *125th AES Convention*, San Francisco, October 2008. Audio Engineering Society.

[14] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:623–656, July and October 1948.

[15] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36(2):111–147, 1974.

[16] T. Thiede, W.C. Treurniet, R. Bitto, C Schmidmer, T. Sporer, J.G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten. PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality. *J. Audio Eng. Soc.*, 48(1/2):3–29, January/February 2000.

[17] H. Wold. Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiaah (Ed.). *Multivariate Analysis.* (pp.391-420), New York: Academic Press, 1966.