Estimates of perceived spatial quality across the listening area

Philip Jackson[1], Martin Dewhirst[2], Robert Conetta[3] and Slawomir Zielinski[2]

[1]*Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK*

[2]*Institute of Sound Recording, University of Surrey, Guildford, UK*

[3]*Acoustics Research Centre, London South Bank University, London, UK*

Correspondence should be addressed to Philip Jackson (p.jackson@surrey.ac.uk)

ABSTRACT

This paper describes a computational model for the prediction of perceived spatial quality for reproduced sound at arbitrary locations in the listening area. The model is specifically designed to evaluate distortions in the spatial domain such as changes in source location, width and envelopment. Maps of perceived spatial quality across the listening area are presented from our initial results.

1. INTRODUCTION

Motivation

The strength of perceptual models has been demonstrated by their use in predicting loudness [1, 2] and also in algorithms such as MPEG audio layer 3 [3, 4], where the data in the audio streams are reduced while minimizing the effect on the perceived attributes of the reproduced sound. These have shown the importance of the listener's perception in the design of audio reproduction systems. Perceptual models that predict the sound quality impairments of speech and audio coding systems based on the timbral and temporal aspects of reproduced sound are well established. These include PEAQ [5] and PESQ [6] models, which are designed primarily to evaluate the audibility of codec distortions in terms of basic audio quality and do not explicitly consider spatial distortions. The 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 and combine these to calculate an overall value of *spatial quality*.

Applications

The range of spatial quality available from different rendering systems has greatly expanded in recent years, giving rise to a greater need for assessing the spatial quality of different processes and systems. Examples are virtual reality, telepresence, home entertainment, automotive audio, games and communications products. As listening tests are both lengthy and highly resource intensive, an alternative means of assessing spatial quality is highly desirable. A model for predicting spatial audio quality has many applications in audio engineering, including automatic system alignment and evaluation of alternative rendering formats and codecs. In addition, such a model could be used to extend existing standard quality models such as PEAQ and also to help extend our understanding of human spatial sound perception. Thus, our model is aimed not only at codec quality evaluation, but also at a wider range of spatial distortions that can arise in audio processing and reproduction systems: downmixing algorithms, spatial audio codecs, virtual surround algorithms, loudspeaker misplacement, level misalignment and phase errors, channel rearrangement and removal, spectral filtering, interchannel crosstalk, and combinations of these.

Background

In addition to evaluating system performance at the 'sweet spot' or 'hot spot' at the centre of the listening area, there is an increasing interest in quantifying perceived spatial attributes at multiple listening positions. For example, the listening tests conducted by Marentakis et al. [7] investigated the minimum audible angle of reproduction systems using Vector Base Amplitude Panning (VBAP) [8] and first- and second-order Ambisonics [9] across a wide listening area. Macpherson [10] and Rose et al. [11] used binaural models to assess the spatial performance of audio reproduction systems when the listener was displaced laterally, yet only one type of re-

production system was tested in each case. Both these papers study how the localisation of sources varies at off-centre listening positions and do not consider the effect on other perceived spatial attributes, such as envelopment, that also contribute to the overall perceived spatial quality. Meanwhile, Härma et al. [12] predict sound quality over a wide listening area by modeling a timbral factor, the sound colouration, at multiple listening positions. However, they do not consider the spatial aspects of sound quality. The model described in this paper can predict the perceived spatial quality at multiple listening positions, and takes into account timbral and spatial factors, incorporating measures associated with the localisation of sources in the foreground of the sound scene as well as background attributes.

Roadmap

The model provides an estimate of the overall spatial quality, expressed as a mean opinion score (MOS), i.e. a global attribute describing any and all changes in the spatial attributes of arbitrary audio reproduction system when compared to a reference reproduction system. The current implementation was designed to evaluate typical audio processes, comparing spatially degraded multichannel audio material against reference five-channel (ITU-R BS.775-1) audio material [13]. An overview of the method is given in Section 2, which includes the calculation of the signals at the listener's ears (*binaural* signals), the auditory processing and binaural cues, the test signals and their associated metrics. Section 3 describes how the metrics were combined to estimate MOS using a multivariate regression that was calibrated with listening test data Section 4 contains the results of simulation experiments predicting spatial quality across the listening area, which are discussed before we conclude.

2. METHOD

In this paper, a model is used to predict measures of spatial quality for a given device under test (DUT) compared to a reference five-channel reproduction system (ITU-R BS.775-1). Probe signals, designed to stress the spatial performance of reproduced audio, are processed with and without the DUT yielding two sets of probe signals: the original reference set and an impaired set corresponding to the DUT. Typical examples of a DUT would be a 64-kbps MPEG codec, a 2.0 (stereo) downmix, or a -6 dB level misalignment on the front left, centre and right (LCR) channels. For each listening position, binaural signals are calculated for the two sets of probe signals by modeling the reproduction system and acoustic listening environment as linear time-invariant systems. Hence, the model extracts five metrics from the binaural signals that are designed to quantify the spatial attributes of the reproduced audio, including accurate rendering of localisable sound sources and listener envelopment: *iacc_9band*, *front_angle_diff*, *mean_spectral_rolloff*, *max_rms_diff*, and *mean_entropy*. Differences between the reference and impaired metric values are combined in linear regression to give an estimate of perceived spatial quality at each listening position.

2.1. Model overview

There are already established perceptual models that predict the sound quality impairments of speech and audio coding systems based on the timbral and temporal aspects of reproduced sound (e.g., PEAQ). The present model was designed in the QESTRAL project to evaluate the effect of distortions in the spatial domain, such as changes in location and envelopment, and calculate an overall value of spatial quality. A detailed description of the construction of the model was presented at the AES Convention in San Francisco, 2008 [14, 15, 16, 17]. For the present purposes, an important feature of the model is its use of binaural signals, allowing spatial quality to be predicted at multiple listening positions to create maps of spatial quality across the listening area. The following sections give a summary of the main elements that are critical for the simulation experiments.

2.2. Calculation of binaural signals

The reproduction of sound in the simulated reproduction environment can be modeled as a linear invariant system, where the sound pressure at any point is the superposition of pressures due to each sound source. Loudspeakers were treated as having perfectly uniform directivity within an anechoic acoustic environment. Impulse responses for any azimuth were obtained by cubic-spline interpolation of the magnitude and unwrapped phase of the frequency response from Gardner and Martin's Head Related Impulse Response (HRIR) database [18]. The HRIRs in the database were recorded at a radius of 1.4 m, so HRIRs at different distances were modeled by adjusting the magnitude and initial delay. Hence, the binaural signals were obtained as the sum of contributions from each loudspeaker at the left and right ears respectively.

2.3. Auditory processing and binaural cues

Auditory processing of the binaural signals was used to extract several perceptually-relevant measures of the soundfield. First, the binaural signals were divided into 24 frequency bands using a gammatone filter bank [19, 20]. The left and right signal envelopes for the *m*th frequency band, $b_{\mathrm{L},m}(t)$ and $b_{\mathrm{R},m}(t)$, were generated by rectifying and smoothing the band-limited signals with a 1.1 kHz low-pass filter, to mimic hair cell behaviour. So, the *interaural cross-correlation function* (IACCF) is

$$r_b(m,t,\tau) = \frac{\sum_{n=1}^{N} b_{\mathrm{L},m}(t+n) b_{\mathrm{R},m}(t+n+\tau)}{\sqrt{\sum_{n=1}^{N} b_{\mathrm{L},m}^2(t+n) \sum_{n=1}^{N} b_{\mathrm{R},m}^2(t+n)}}, \quad (1)$$

over lag τ at time t, based on N samples ($|\tau| < 1$ ms). The *interaural cross-correlation* (IACC) and *interaural time difference* (ITD), $\Delta_{\mathrm{T},m}$, for each band m were taken at the peak IACCF value. The *interaural intensity difference* (IID), $\Delta_{\mathrm{I},m}$, represented the ratio of average sound intensity over N samples, in dB.

For each band m, two look up tables to relate azimuth θ and interaural difference Δ_m were populated using ITD and IID values respectively, obtained with the Gardner and Martin HRIR database as training data. By appropriate normalisation of the tables' columns and rows, and using Bayes' theorem with uniform prior [21], an output was obtained to approximate the posterior probability of the azimuth for a given interaural difference:

$$P(\theta|\Delta_m) = \frac{P(\Delta_m|\theta)}{\sum_{\theta'=-90^\circ}^{+90^\circ} P(\Delta_m|\theta')}. \quad (2)$$

Thus, for each time frame, $\Delta_{\mathrm{T},m}$ and $\Delta_{\mathrm{I},m}$ values from each critical band provide a total of 48 probability histograms across θ. The histograms are weighted by Duplex theory and by loudness within each band, then summed over all 24 frequency bands to yield one summary ITD histogram and one IID histogram, $c_{\mathrm{T}}(t,\theta)$ and $c_{\mathrm{I}}(t,\theta)$. Finally, these are combined by multiplication, $c_\Delta(t,\theta) = c_{\mathrm{T}}(t,\theta) c_{\mathrm{I}}(t,\theta)$, averaged over all time frames, and the peak taken as the localisation angle $\hat{\theta}$. This algorithm for predicting perceived source localisation was validated using a formal listening test [22], achieving a correlation R^2=0.98. [1]

[1] The stimuli for this validation consisted of pink noise, female and male speech and solo musical instruments either played from a single loudspeaker or constant-power panned between a pair of loudspeakers.

2.4. Test signals and calculation of metrics

Listeners perceive a sequence of notes from a musical instrument or the sequence of phones from speech as an auditory stream, where a typical programme material contains multiple streams which can be attended to individually or as a whole [23]. We use the concept of the foreground and background within the auditory scene to develop suitable tests of the DUT, by injecting probe signals and extracting measures of the response from the sound received at the listener [24]. Clearly localisable sources in the foreground reveal the pincipal spatial distortions as changes in their perceived location. However, reverberation and distributed sound sources contribute significantly to the overall spatial impression as the background [25], e.g., late lateral reflections in the sense of listener envelopment [26].

To assess the main effects, two sets of test signals were used to probe the foreground and background distortions respectively: 'spun noise', i.e., a series of pink noise bursts rendered at 10° intervals, and decorrelated pink noise played simultaneously through all channels. The spun noise test signals were generated using pair-wise constant power panning over the reference five-channel reproduction system [13]. The decorrelated pink noise is designed to approximate a diffuse acoustic field. Two metrics were calculated from the binaural signals obtained with the spun noise: *front_angle_diff* and *max_rms_diff*. Three metrics were extracted with the decorrelated pink noise: *iacc_9band*, *mean_spectral_rolloff* and *mean_entropy*. The metrics are described briefly below in order of their importance in the spatial quality prediction; for a more detailed description of metrics used with the model, see [16].

IACC metric: iacc_9band

The degree of correlation between the signals at the two ears has been shown to be related to the perceived envelopment and source width [27, 28]. Higher correlations between the IACC values and the spatial quality values from the listening tests were obtained when only the nine critical bands with centre frequencies from 570 Hz to 2160 Hz were considered, which approximates the three octave bands used to predict acoustical quality in concert halls by Hidaka et al. [29] The *iacc_9band* metric was given an exponential warping, applied to the decor-

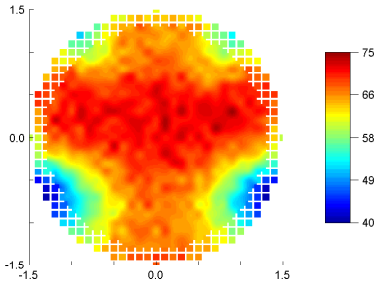

Fig. 1: The *iacc_9band* metric calculated across the listening area for the reference with background test signals, using 2D cubic splines to interpolate the 10 cm grid where possible.

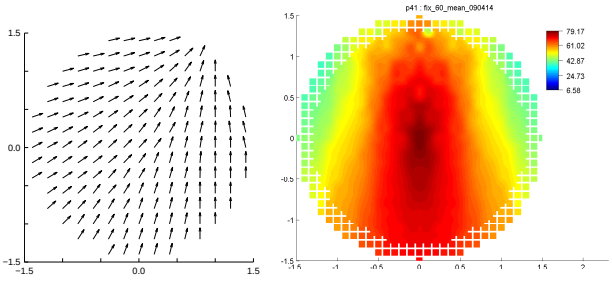

Fig. 2: Foreground source localisation plots across the listening area: (left) localisation azimuths calculated for the reference from noise panned to $+40^\circ$, sub-sampled to a 20-cm grid; (right) *front_angle_diff* metric calculated for the reference system, with cubic-spline interpolation.

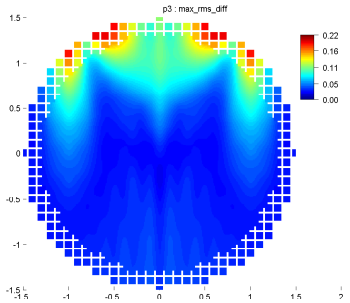

Fig. 3: The *max_rms_diff metric* calculated across the listening area for 2.0 (stereo) downmix, with cubic-spline interpolation.

related pink noise and is defined

$$iacc\_9band = \exp\left(-\frac{3.13}{9}\sum_{m=1}^{9}\left(\max_t \mathrm{IACC}(m,t)\right)\right). \quad (3)$$

The result is illustrated for the reference 5.0 reproduction system in Fig. 1.

Localisation metric: front_angle_diff

As the front scene dominates the placement of foreground sources in typical 5-channel programme material, the *front_angle_diff* metric concentrates on localisation errors to spun noise sounds in the intended range $\phi \in \pm 30^\circ$:

$$front\_angle\_diff = \exp\left(-\frac{0.016}{N}\sum_{|\phi|\le 30}\left|\hat{\theta}_{\mathrm{Ref}}(\phi)-\hat{\theta}_{\mathrm{DUT}}(\phi)\right|\right) \quad (4)$$

which takes the mean difference between localisation angles for the reference and processed test signals, $\hat{\theta}_{\mathrm{Ref}}(\phi)$ and $\hat{\theta}_{\mathrm{DUT}}(\phi)$ respectively, and $N = 7$ is the number of test angles. Fig. 2 (left) shows $\hat{\theta}_{\mathrm{Ref}}(+40^\circ)$ as an example, and Fig. 2 (right) gives the *front_angle_diff* map for the reference system.

Spectrum metric: mean_spectral_rolloff

Changes in the high frequency content of the audio signals due to the impairment processes were found to affect the perceived spatial quality, in particular the envelopment and source distance. The spectral rolloff [30], is defined at each time frame by the frequency up to which covers 95 % of the magnitude spectrum. The spectrum was obtained by Fourier transform of the sum of the left- and right-ear signals using decorrelated noise, and the *mean_spectral_rolloff* is the mean over time. The DUTs that affect this metric, such as low-pass filters or certain codecs, tend to influence all positions and give lower spatial quality right across the listening area.

Level metric: max_rms_diff

Using spun noise, the average sound pressure for the left and right signals is defined at each azimuth $\phi \in \pm 180^\circ$:

$$\mathrm{RMS}(\phi) = \sqrt{\frac{\sum_{t=1}^{T}\left(a_{\mathrm{L},\phi}(t)+a_{\mathrm{R},\phi}(t)\right)^2}{4T}}, \quad (5)$$

where $a_{\mathrm{L},\phi}(t)$ and $a_{\mathrm{R},\phi}(t)$ are the left- and right-ear signals respectively at time t and T is the number of samples. The *max_rms_diff* metric is calculated as

$$max\_rms\_diff = \max_{\phi}\left|\mathrm{RMS}_{\mathrm{Ref}}(\phi)-\mathrm{RMS}_{\mathrm{DUT}}(\phi)\right|, \quad (6)$$

where $\mathrm{RMS}_{\mathrm{Ref}}(\phi)$ is the reference value of $\mathrm{RMS}(\phi)$ and $\mathrm{RMS}_{\mathrm{DUT}}(\phi)$ is the value with the DUT. This metric was

designed to show distortions in the sound pressure level for different scene components, reflected by the reproduction system's ability to reproduce the correct levels of sources around the listener. The plot in Fig. 3 shows large differences close to the loudspeakers active with the 2-channel downmix.

Entropy metric: mean_entropy

The sense of envelopment was found to include factors such as the number of sound sources and the amount of reverberation. The signal entropy was used as measure of the auditory information, which correlates with some of these factors as well as loudness [31], calculated for the left-ear signal $a_{\mathrm{L}}(t)$ as

$$\mathrm{entropy}_{\mathrm{L}} = -\sum_{t=1}^{T} P(a_{\mathrm{L}}(t)) \ln P(a_{\mathrm{L}}(t)), \qquad (7)$$

where $P(\cdot)$ is the probability of a sample value, estimated from the histogram of the sample distribution [32]. The *mean_entropy* metric is defined as the mean of left and right entropies. As signal entropy is associated with amplitude, plots of this metric show high values close to the active loudspeakers that reduce gradually with increasing distance.

3. CALIBRATION

The results from two subjective listening tests [15] were used to calibrate the model. Both listening tests assessed changes to the spatial quality created by processes degrading the reference audio reproduction.

Three different five-channel reference recordings were used in the tests, each being typical of five-channel programme material with various spatial characteristics, representing different genres: TV sport (a tennis match with commentators panned midway between the left and centre channels and the centre and right channels and with applause in all channels), classical music (with a wide continuous front stage and the surround channels containing ambient sound and reverberation) and pop music (with a wide continuous front stage, the main vocal in the centre channel, and additional instruments in the left and right surround channels). Informal listening showed that each reference recording was highly enveloping and contained distinctive source locations. Forty-three different processes were applied to the reference recordings to create stimuli with different spatial quality for the two listening tests. These DUTs covered a wide variety of degradations at all stages of processing and rendering, included downmixes, low bit-rate audio codecs and loudspeaker misplacements.

Both listening tests used a listening room and five-channel loudspeaker array conforming to ITU standards [33, 13] and drew the test subjects from experienced listeners at the Institute of Sound Recording at the University of Surrey. The method, which is described in detail in [15], used a 100-point scale and listeners were instructed to give the top score for recordings with the same spatial quality as the reference recording and to judge any changes in spatial quality as impairments.

Two different listener positions were used within each of the listening tests: the centre of the listening area (the *sweet spot*) and one metre to the right. A major difference between the two listening tests was that the reference was presented at the off-centre listening position in the first listening test (viz. the reference recording heard at the off-centre position) whereas in the second experiment additional loudspeakers were used to present the reference with a configuration centred on the listener. Thus, the results for the off-centre position in the first listening test had a different scale to all the other results. A quadratic function was found to map the results from the off-centre listening position in the first listening test to the same scale as the rest of the results (R^2=0.94).

Once all data from both listening tests were on a single scale, they were used to calibrate a least-squares regression for predicting spatial quality from the calculated metric values. As the top of the scale in the listening tests was fixed for the reference reproduction, the regression was constrained to give a MOS of 100, implemented using QR factorization of the metric diff grades [34]. Table 1 shows the coefficients of the resulting regression model. A leave-one-out cross-validation gave R^2=0.78 and root-mean-squared error of prediction (RMSEP) of 12.0 %, shown in Fig. 4. A further cross-validation was performed by dividing the set of all listening test results into two subsets of equal size, for calibration and validation, which yielded the same R^2 and RMSEP values.

4. RESULTS

For our simulation experiments, the listening area inside the five-channel loudspeaker configuration was sampled using a 10-cm grid. The model was used to estimate

Metric name	SE(B)	B
iacc_9band	-0.47	-271.21
front_angle_diff	-0.42	-45.95
mean_spectral_rolloff	-0.25	-0.003
max_rms_diff	0.22	371.08
mean_entropy	-0.16	-16.76
Constant	–	100.00

Table 1: The coefficients of the regression model fitted to the results from both listening tests. The second and third columns contain the standardized and raw coefficients respectively.

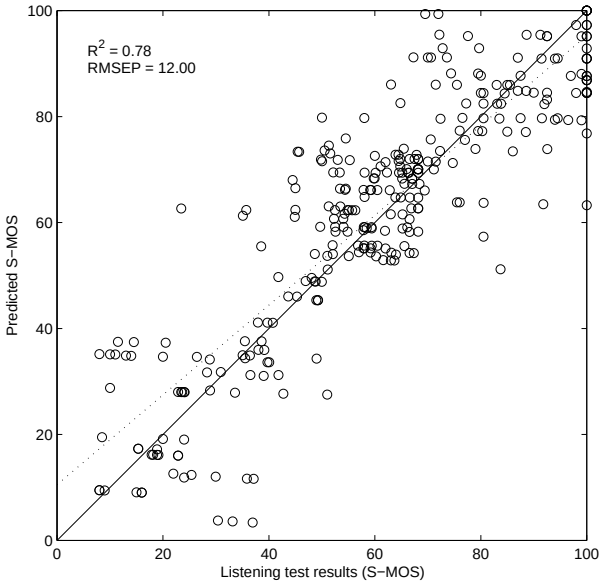

Fig. 4: Results of the leave-one-out cross-validation of the regression model. The solid line shows the ideal relationship and the line of best fit is shown dotted.

the spatial quality at each point on the grid for different DUTs, and the results were plotted as maps of spatial quality, using 2D cubic splines to interpolate the 10cm grid where possible. The benefits of modeling spatial quality across the listening area include ensuring that the model is generalisable and also mitigation of the risk of overfitting to the listening test data.

Fig. 5 shows plots of predicted spatial quality across the listening area for eight processes. The 95% confidence interval for predicted MOS of 100 was calculated (i.e. the spatial quality of the reference soundfield at the centre listening position). The lower limit of the confidence interval is shown as a black contour in the relevant plots in Fig. 5. For each one, the area inside the contour provides a measure of the extent of the DUT's good spatial reproduction.

The results for the reference soundfield (top left) and 3/1 downmix (middle left) both show the limitations of the loudspeaker setup: the spatial image collapses into the nearest loudspeaker at listener positions close to the loudspeaker locations. The *iacc_9band* metric in particular shows this behaviour (see Fig. 1). The results for the 3/1 downmix are similar to those for the reference soundfield, with the addition of a thin area of lower predicted spatial quality between the two surround loudspeakers. This is due to the Left and Right Surround channels dominating the binaural signals at these positions. As these two channels are identical for the 3/1 downmix, this results in a low value for the *iacc_9band* metric (the only significant difference between the 3/1 downmix and the reference).

The results for the process attenuating the Left, Right and Centre channels by 6dB (Fig. 5, top centre) show the Left and Right Surround channels dominating the binaural signals for a large proportion of the listening area. Two of the metrics substantially differ for this process compared to the reference soundfield; *mean_entropy*, which is substantially lower throughout the listening area for the processed soundfield, and *iacc_9band*, which is significantly lower near the two surround loudspeakers. The *iacc_9band* metric in particular causes the predicted spatial quality to be much lower in the region of the two surround loudspeakers (shown by the large green and blue areas on the left and right of the plot).

The pattern of results for the 2.0 downmix (Fig. 5, middle centre) is due largely to the *max_rms_diff* and *iacc_9band* metrics. The *max_rms_diff* metric has higher values when the listener position is located close to a loudspeaker, as those spun noise signals that use the near loudspeaker will be louder at the listener's ears. The effect is even greater in the region close to the left and right loudspeakers for the 2.0 downmix, as the left and right loudspeaker signals also incorporate the centre and surround channels from the original recording. The vertical line down the middle of the spatial quality plot for the 2.0 downmix is determined by the *iacc_9band* metric, which is attributable to the interference between two loudspeakers and two receivers (the listener's ears).

All the metrics were lower for the 1.0 downmix (Fig. 5, middle right), leading to its generally poor spatial quality. The value of *front_angle_diff* was lower when the active loudspeaker was not directly in front of the listener, which contributes to the pattern visible in the plot.

The pattern in spatial quality for the 3.5 kHz low-pass filter (top right) is similar to that of the reference (top left), albeit at a lower level. The metric values differ little from those of the reference soundfield, with the exception of the *mean_spectral_rolloff*, which is relatively uniform over the listening area but at a much lower level, as expected.

The differences between the predicted spatial quality for the reference and the 80 kbps codec (bottom left) are mainly due to *iacc_9band*. The codec seems to alter the correlation between the different loudspeaker channels, resulting in lower *iacc_9band*, especially at positions directly in line with the Centre loudspeaker.

For the 1.0 downmix played through the Left Surround loudspeaker, all metric values were worse. The spatial quality was slightly higher nearer to the active loudspeaker (mainly due to *mean_entropy*) but overall predictions were deservedly the lowest (Fig. 5, bottom centre).

5. CONCLUSION

This paper presents maps of estimated spatial quality across the listening area, created using a model for predicting spatial quality refined from previous work [14, 15, 16, 17]. Part of the motivation for the development of the model used to generate these estimates is to find an alternative to lengthy, resource intensive listening tests. Potential applications for the model include automatic system alignment, evaluation of alternative rendering formats and codecs, and extensions to existing quality standards. The model shows a high correlation ($R^2 = 0.78$) and low error (RMSEP=12.0%) when cross-validated using the listening test results.

The simulation results show the combination of foreground, background and timbral factors in determining the spatial quality of a reproduction at any listening position. While the spatial quality is not significantly different between the reference five-channel and downmixed two-channel reproduction at the sweet spot, the area for which this holds true is much smaller in the latter case. However, elementary operational errors (e.g., in routing, channel alignment or missing channels) can cause significant degradation across most of the listening area. Codecs at moderate bit rates appear able to maintain a reasonable quality compared with the reference, yet bandwidth reduction and increased inter-channel correlation showed a substantial effect. These predicted spatial quality maps have the potential not only to provide better understanding of the relative importance of factors in rendering a sound scene, but also to assist in the design and development of audio reproduction.

As well as work to relax some of the modeling assumptions, an important area of future work is to extend the model validation at additional listening positions using formal listening tests. Equally, the relationship between the type of programme material and the perception of spatial quality degradations merits further investigation.

6. ACKNOWLEDGMENTS

This research was completed as part of the QESTRAL project which was funded by the Engineering and Physical Sciences Research Council (EPSRC) - grant EP/D041244/1. This project was a collaboration between the University of Surrey, Bang and Olufsen and BBC Research. The authors would also like to thank Søren Bech, David Meares, Ben Supper, Russell Mason, Sunish George and the listening test subjects for their contributions to the research presented in this paper.

7. REFERENCES

[1] E. Zwicker and B. Scharf. A model of loudness summation. *Psychol. Rev.*, 72(1):3–26, January 1965.

[2] B.C.J. Moore, B.R. Glasberg, and T Baer. A model for the prediction of thresholds, loudness and partial loudness. *J. Audio Eng. Soc.*, 45(4):224–240, April 1997.

[3] K. Brandenburg and G. Stoll. The ISO/MPEG-audio codec: a generic standard for coding of high quality digital audio. 92nd Conv. Audio Eng. Soc., Preprint 3336, Vienna, March 1992.

[4] B. Grill. Using ISO/MPEG audio layer 3 for high quality music transmission. UK 8th Conference: Digital Audio Interchange (DAI), London, UK, May 1993.

[5] T. Thiede, W.C. Treurniet, R. Bitto, C Schmidmer, T. Sporer, J.G. Beerends, C. Colomes, M. Keyhl,

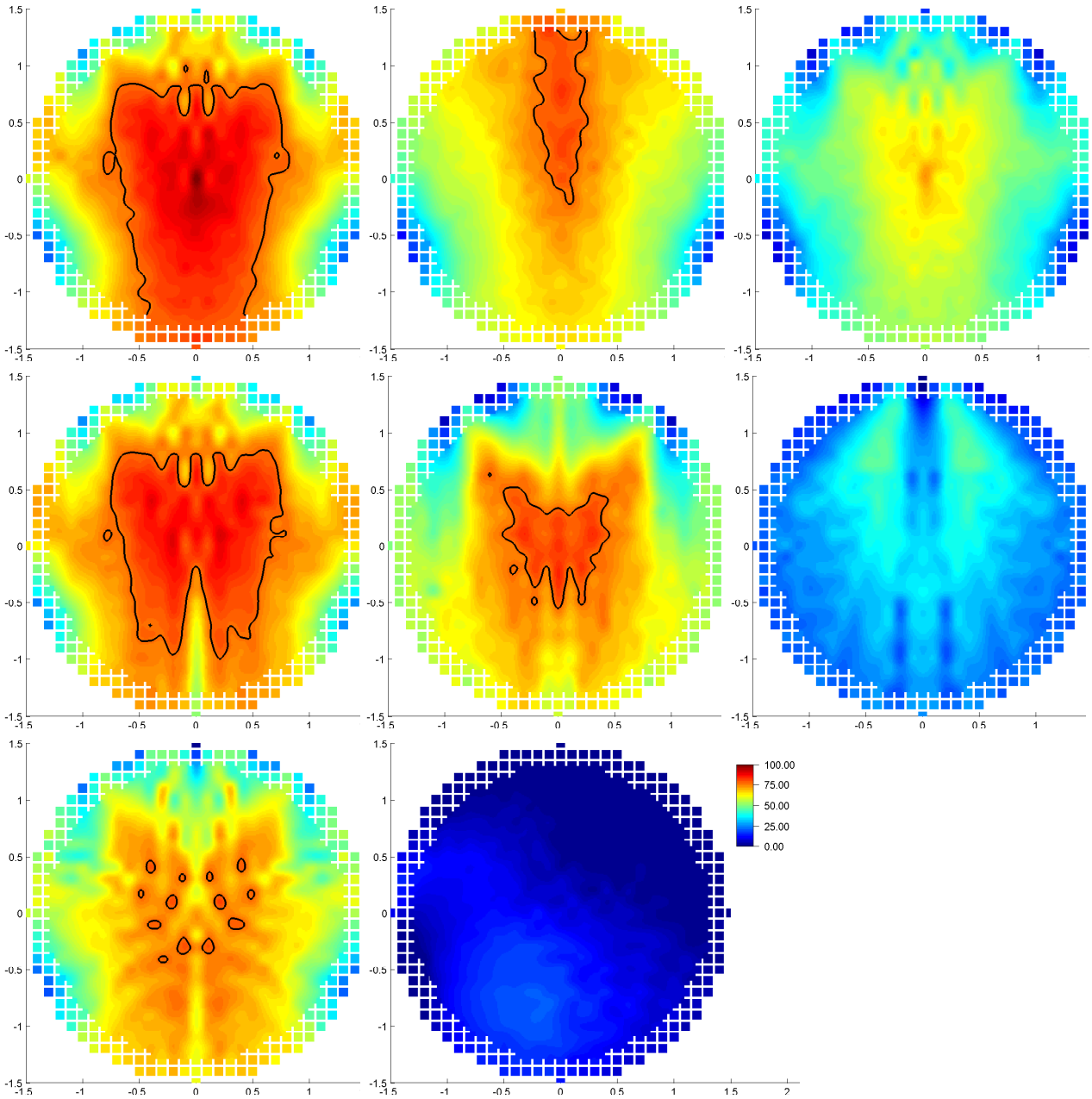

Fig. 5: Predicted spatial quality (MOS) at multiple locations across the listening area: (top, from left) reference soundfield, Left, Right and Centre channels attenuated by 6dB, 3.5kHz low pass filter on all channels; (middle row) 3/1 downmix, 2.0 downmix, 1.0 downmix; (bottom row) 80kbs codec, 1.0 downmix from the Left Surround loudspeaker.

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.

[6] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. PESQ - The New ITU Standard for End-to-End Speech Quality Assessment. 109th Conv. Audio Eng. Soc., Preprint 5260, Los Angeles, California, September 2000.

[7] G. Marentakis, N. Peters, and S. McAdams. Auditory resolution in virtual environments. Effects of algorithm, off-centre listener positioning and speaker configuration. Acoustics 2008 Conference, Paris, June-July 2008.

[8] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc.*, 45(6):456–466, June 1997.

[9] D. Malham. Higher order ambisonic systems for the spatialization of sound. In *Proc. Int. Computer Music Conf.*, Beijing, China, October 1999.

[10] E.A. Macpherson. A computer model of binaural localization for stereo imaging measurement. *J. Audio Eng. Soc.*, 39(9):604–622, September 1991.

[11] J. Rose, P. Nelson, and T. Takeuchi. Sweet spot size of virtual acoustic imaging systems at asymmetric listener locations. *J. Acoust. Soc. Am.*, 112(5):1992–2002, November 2002.

[12] A. Härmä, T. Lokki, and V. Pulkki. Drawing quality maps of the sweet spot and its surroundings in multichannel reproduction and coding. In *Proceedings of the AES 21st International Conference*, St. Petersburg, Russia, 2002.

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

[14] R. 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.

[15] 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.

[16] 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.

[17] M. Dewhirst *et al*. QESTRAL (Part 4): Test signals, combining metrics and the prediction of overall spatial quality. Presented at the *125th AES Convention*, San Francisco, October 2008. Audio Engineering Society.

[18] B. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone. `http://sound.media.mit.edu/KEMAR.html`, May 18, 1994 (last revised July 18, 2000).

[19] M. Slaney. An efficient implementation of the Patterson-Holdsworth auditory filter bank. *Apple Computer Technical Report #35*, 1993.

[20] W. Gaik. Combined svaluation of interaural time and intensity differences: psychoacoustic results and computer modeling. *J. Acoust. Soc. Am.*, 94(1):96–110, July 1993.

[21] A. Papoulis and Unnikrishna Pillai. *Probability, random variables and stochastic processes*. McGraw-Hill, New York, 4th edition, 2002.

[22] M. Dewhirst. *Modelling perceived spatial attributes of reproduced sound*. PhD thesis, Institute of Sound Recording, University of Surrey, 2008.

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

[24] 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.

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

[26] 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.

[27] R. Mason, T. Brookes, and F. Rumsey. Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli. *J. Acoust. Soc. Am.*, 117(3):1337–1350, March 2005.

[28] S. George, S. Zielinski, and F. Rumsey. Feature extraction for the prediction of multichannel spatial audio fidelity. *IEEE Transactions on Audio, Speech and Language Processing*, 13(6):1994–2005, November 2006.

[29] T. Hidaka, L.L. Beranek, and T. Okano. Interaural Cross-Correlation, Lateral Fraction and Low- and High-Frequency Sound Levels as Measures of Acoustical Quality in Concert Halls. *J. Acoust. Soc. Am.*, 98(2):988–1007, August 1995.

[30] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 10(5):293–302, 2002.

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

[32] R. Modemeijer. On estimation of entropy and mutual information of continuous distributions. *Signal Processing*, 16(3):233–246.

[33] Rec. ITU Draft BS. 1116. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, 1997.

[34] G.H. Golub and C.F.V. Loan. *Matrix computations*. The Johns Hopkins University Press, Baltimore, 3rd edition, 1996.