Audio Engineering Society
Convention Paper 6686

Presented at the 120th Convention
2006 May 20–23 Paris, France

Initial developments of an objective method for the prediction of basic audio quality for surround audio recordings

Sunish George[1], Slawomir Zielinski[2], and Francis Rumsey[3]

Institute of Sound Recording, University of Surrey, Guildford, Surrey, GU2 7XH, United Kingdom

[1] sunish.george@surrey.ac.uk, [2] s.zielinski@surrey.ac.uk, [3] f.rumsey@surrey.ac.uk

ABSTRACT

This paper describes the development of the objective method for the prediction of the Basic Audio Quality (BAQ) of bandlimited or down-mixed surround audio recordings. A number of physical parameters, including interaural cross-correlation coefficient and spectral descriptors, were extracted from the recordings and used in a linear regression model to predict BAQ scores obtained from listening tests. The results showed a high correlation between the predicted scores and those obtained from the listening test, with the average error of prediction being smaller than 10%. Although the method was originally developed for 5-channel surround recordings, after some modifications it can be upgraded to any number of audio channels.

1. INTRODUCTION

It is well known that listening tests for the evaluation of audio quality are expensive and often time-consuming. Objective methods for the evaluation of audio quality are not an alternative solution to the listening tests but can facilitate sound quality optimisation in a product development process or can be employed in telecommunication systems for quality of service monitoring. Several attempts have been done in the past years to predict the basic audio quality (BAQ) from extracted objective features from the recordings. Some of them are reported in [1] to [8]. ITU's attempt to codify a standard for the objective prediction of BAQ has resulted in a standard known as PEAQ [9], [10].

A number of advancements has been proposed to improve the performance of PEAQ [11], [12], [13], [14], [15] and [16]. Unfortunately, most of them predict the BAQ of mono or 2-channel stereo signals and consequently are not suitable for multichannel audio. The recent years have seen the widespread usage of surround audio in home environment and hence the prediction of BAQ for multichannel audio has great importance. The prediction of multichannel audio quality based on objective features is a challenging task, primarily due to the fact that our

knowledge about the relationships between a perceived spatial audio character and physical characteristics of the signals is limited. The purpose of this paper is to report about the initial research that has been done to predict the BAQ of multichannel audio with selected audio quality degradation types such as band-limitation and down-mixing.

The design of a predictor involves two important phases. The first phase is calibration, which is the fundamental process to achieve the consistency of prediction using a set of variables and a desired output. Before generalising the model obtained from the calibration, it needs to be tested for its consistency. It can be done in two ways. In the first method, known as validation, an independent experiment is conducted to get the necessary data for testing the consistency of the calibrated model. In the second method, a set of experimental data is divided into two. The first set of data is used for calibration and the second is used for testing the consistency of the calibrated model. This way of testing the consistency of the model is known as cross-validation. The division of the database can be done in several ways. Some of these methods of dividing the data set can be seen in [36] One way of dividing the database is to randomly select a small percentage (typically 10%, 20% or 30%) of the data to create the validation database. In this paper, the database for the cross-validation is created by randomly selecting 20% of the data obtained from the listening test. Randomly selecting data for cross-validation is more robust and simple than other methods. In addition, the validation uses a dataset that is obtained from an independent experiment to test the consistency of the model whereas in cross-validation, a set of data selected from the same experiment is used for this purpose.

In this paper three methods for the prediction of BAQ are proposed. The first method seeks to predict the BAQ by extracting physical features such as bandwidth and presence/absence of dialogue in the centre channel of the multichannel audio. The second method extracts a number of features and applies a regression model to predict the BAQ. In the third method, the BAQ is predicted by applying an indirect approach, that is, the attributes of BAQ are predicted separately and use those in a regression equation to predict the BAQ. The three attributes of BAQ are predicted separately and a regression equation is used for the prediction, from the independently predicted attributes of BAQ.

This paper is divided into eight sections, including this introduction. Section 2 describes the listening test score database and a brief description about the experiment used to obtain the database. Sections 3, 5 and 6 describe implementation details and the results of the aforementioned methods for the prediction of BAQ. Section 4 describes the physical features extracted from the multichannel recordings for the prediction of listening test scores. Section 7 discusses about the results obtained from the three methods of prediction and Section 8 closes the paper with conclusions and future work. The Tables are presented in the APPENDIX at the end of the paper.

2. LISTENING TEST DATABASE

As mentioned previously, the listening test scores for the calibration and the cross-validation were obtained from the experiments that has been conducted at the Institute of Sound Recording during the project investigating subjective quality trade-offs in consumer multichannel sound and video delivery systems. The experiment was conducted in an ITU-R. BS. 1116 Recommendation [37] compliant listening room at the University of Surrey, UK. The audio setup used for the listening tests is shown in Figure 1. The following paragraphs bring out a summary of the listening test score database. A more detailed coverage of the experimental setup and the listening test results can be found in [21].

There were twelve audio programme materials for the listening test, selected from movies, music recordings, TV programme etc. The recordings were of two types- depending on the audio scene characteristic they carried- ‘F-F’ and ‘F-B’. An audio scene with ‘F-F’ characteristics means that the front and the rear audio channels contained clearly distinguishable audio sources. That is, the recordings with ‘F-F’ audio scene characteristics bring out the listening impression that is similar to that when a listener is surrounded by a group of instruments in an orchestra. The listening impression from ‘F-B’ type recordings is similar to that experienced in a concert hall. It means that the front channels contain clearly distinguishable audio sources and the rear channels contain mainly reverberant sounds and room response. A detailed discussion about the audio scene characteristics can be found in [22] and [23].

The audio programme material was processed in two different ways: band-limiting and down-mixing. The band-limiting was done by following two different

approaches. In the first approach, all the channels were passed through a low pass filter of equal cut-off frequencies. In the second approach the cut-off frequency of the low pass filter differed across the channels. In down-mixing, the number of channels was reduced by re-directing the content of certain channels to others. The detailed description of the processes applied to the programme material is given in Tables 1 and 2 in Appendix. There were a total of 138 audio recordings. The reference recordings were selected from various sources including commercially released DVD disks and used for the listening test at a sampling frequency of 48 kHz and 16 bit resolution.

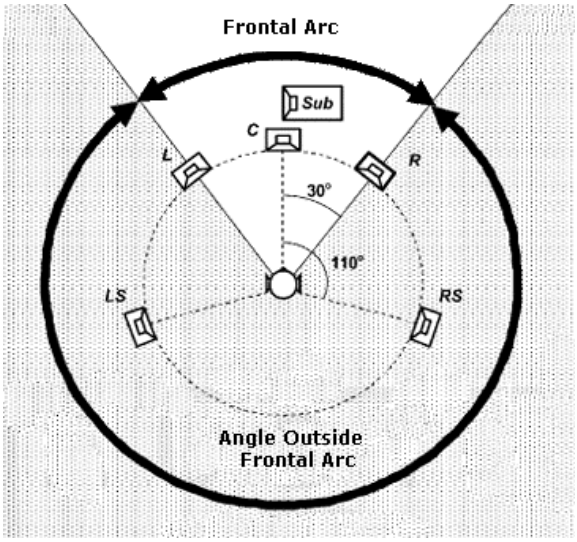

Figure 1: multichannel audio setup: Frontal Arc and the angle outside Frontal Arc.

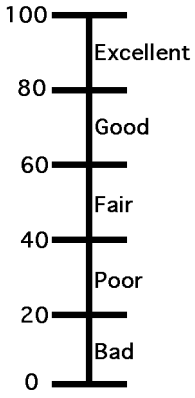

Figure 2: Grading Scale used for the listening tests

Each subject was asked to grade BAQ and the three attributes of BAQ: timbral fidelity, frontal spatial fidelity and surround spatial fidelity. Frontal spatial fidelity can be defined as the global attribute that describes any and all detected differences in the 'spatial impression' inside the frontal arc (see non-shaded area in Figure 1) of the multichannel audio setup, between the reference and the evaluated recording. The definition of the surround spatial fidelity can be given as the global attribute that describes any and all detected differences in the spatial impression outside the frontal arc (see shaded area in Figure 1) of the multichannel audio setup, between the reference and the evaluated recording.

The relative importance of timbral fidelity, frontal spatial fidelity and surround spatial fidelity in basic audio quality are described in [20]. The grading scale used for the test is presented in Figure 2. The listening test database (138 aggregated scores in total) was divided into two subsets for the cross-validation purpose, as described in the previous Section.

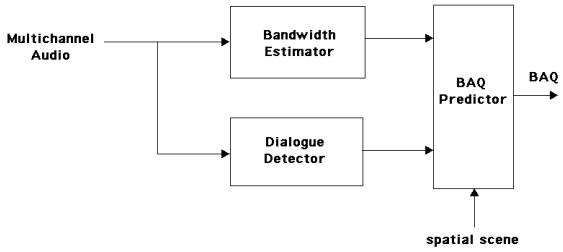

Figure 3: Schematic diagram of the Quality Adviser based predictor for BAQ

3. PREDICTION OF BASIC AUDIO QUALITY: METHOD 1

The method proposed here is inspired by the quality advisor proposed in [24]. There are some limitations for this model, but the results are promising. The algorithm used here directly uses the quality advisor model and has not been modified in any way for the purpose of this paper. Also, the quality advisor has already been validated with an independent experiment and the results have been previously reported in [24]. Hence, it was decided not to apply any sort of consistency test for this model. The algorithm extracts three features to predict the basic audio quality. They are bandwidth of the channels, presence/absence of dialogue in the centre channel

and the spatial characteristics. A schematic of the algorithm is given in Figure 3.

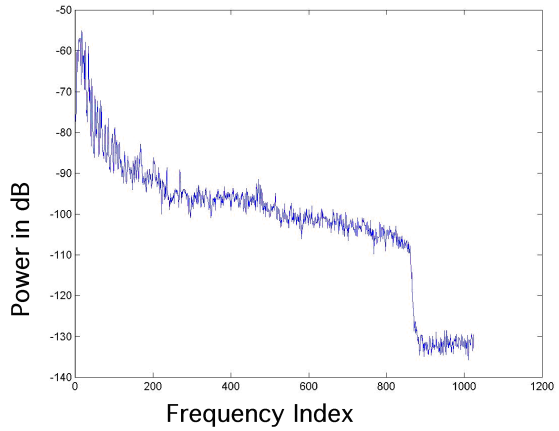

Figure 4a: An example of averaged spectra.

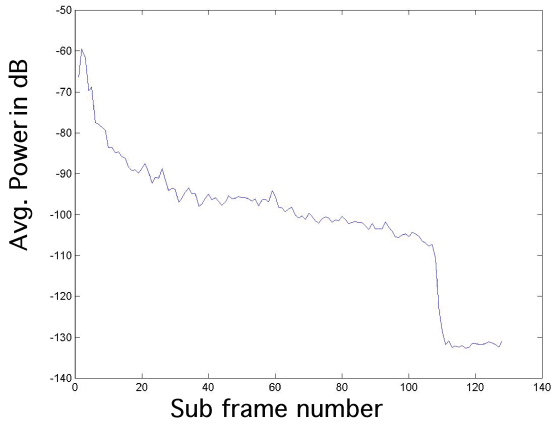

Figure 4b: Smoothed spectra

The bandwidth of each channel is determined by the bandwidth estimator. The averaged power spectrum of the audio signal is created and the bandwidth is estimated by dividing the entire signal into several windows. A window of size 42.67 ms is used and the bandwidth estimator determined the largest attenuated region of frequency. During the computation, the average power spectrum is smoothed by decimating the spectrum into different sub-frames (See Figure 4a and Figure 4b). The upper frequency limit of the bandwidth was considered to be the region at which the highest difference in power occurred. For the purpose of prediction, the averaged bandwidth of left and right channels, bandwidth of the centre channel, and averaged bandwidth of left and right surround channels are computed from the individual bandwidths of the channels. It was found that the channels with similar in bandwidth can be considered to be of equally bandlimited recording. Hence, if the magnitude of the differences between front, centre and surround channels did not exceed 1 kHz, recording was considered to be equally bandlimited, and was computed by averaging the three values.

The second feature is computed by the dialogue detector. For a given audio degradation process, the presence or absence of dialogue can affect the perceived audio quality in a significant way. It was assumed that for a recording with dialogue the centre channel contained a signal that had no or very low correlation with the other front channels. Also it was assumed that if the level of the centre channel was large compared to the other front channels, the recording could be considered as a recording with dialogue. The flow chart of the dialogue detector is given in Figure 5. The abbreviation CrCorr represents the procedure that was used to compute the cross correlation between two channels within 1 ms time window. The MCorr procedure finds the maximum from the correlation array obtained from the CrCorr procedure.

The BAQ estimator uses the regression equations (see Table 3) presented in [24] to compute the basic audio quality from the extracted features as described above.

The BAQ estimator was designed to deal with two types of audio scene feature. However, its full functionality has not been used here because of the computation complexity involved in estimating audio scene from multichannel audio recordings. Hence, in this model, the spatial scene was set to 'F-B', irrespective of the audio characteristics of the recordings. However, the effect of this assumption has been examined and found to be negligibly small.

3.1. Results

According to the results obtained using the method described above, the predicted scores for Basic Audio Quality, showed a very high correlation with the scores obtained during the listening test yielding a correlation coefficient of 0.971 and prediction error of 6%. The magnitude of the prediction error obtained in this experiment is considered to be small considering the fact that the error in the listening tests

due to inter-listener variance is often of the order of 10%. Figure 6 shows the scores obtained in the listening test plotted against the predicted scores.

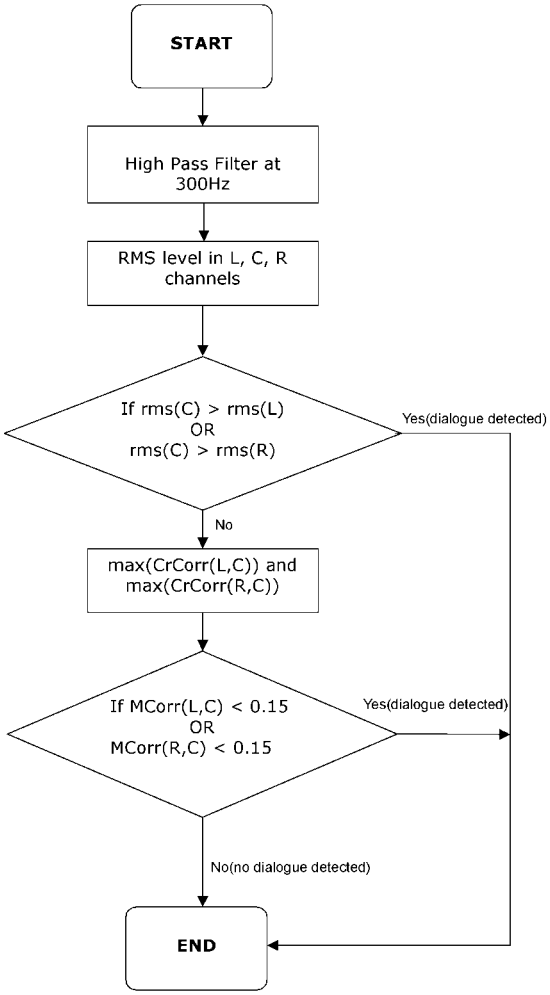

Figure 5: Flow chart of the dialogue detector

As it was mentioned above, in this model the spatial scene was set to 'F-B', irrespective of the audio characteristics of the recordings. It was found that this simplification did not cause much difference in the results. The results of prediction without this simplification (tested manually) showed a correlation of 0.974 with a similar standard error of estimate.

The prediction results of the model described above showed a very high correlation and low standard error of estimate. But, the employed algorithm has some limitations. Primarily, it is applicable only to bandlimited recordings. The model is not capable to predict the basic audio quality with down-mixed or other types of degradations. Hence it is necessary to develop an algorithm that is capable of predicting the BAQ with down-mixed or other types of degradations. The models presented in Sections 5 and 6 tried to predict the BAQ of down-mixed recordings as well..

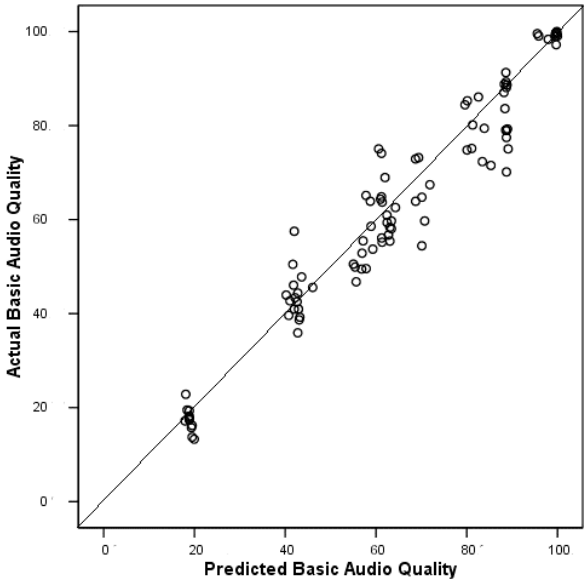

Figure 6 Scatter plot of the actual and predicted Basic Audio Quality scores

4. EXTRACTED FEATURES

A number of features were extracted from the recordings to predict the basic audio quality and other attributes of BAQ. Basically, there were two types of features extracted from the recordings: spectral features and spatial features. The spectral features represent the difference in the spectral content whereas the spatial features represent the difference in the spatial characteristics between the reference recording and test recording. All features were computed using MATLAB. A summary of the procedures followed to extract the features are given in the following paragraphs.

4.1. Spectral Features

Before extracting the spectral features, the multichannel recordings were downmixed to a mono signal by summing up all the channels. The mono versions were analysed in order to extract different features. The details of the spectral features are given below.

4.1.1. Spectral centroid and spectral rolloff based feature:

Spectral centroid is the centre of gravity of the magnitude spectrum of an audio signal. It is considered to be the objective representation of the subjective parameter brightness of an audio signal. The basic expression used for the computation of spectral centroid is given below [25]:

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]}, \qquad (1)$$

where Mt[n] is the magnitude of the spectrum for a given time frame and frequency index n.

The spectral rolloff also represents the spectral characteristics of an audio signal. The formation of the feature based on spectral rolloff is similar to that of spectral centroid feature. The basic calculation of spectral roll off is given as

$$\sum_{n=1}^{Rt} M_t[n] = 0.95 * \sum_{n=1}^{N} M_t[n], \qquad (2)$$

where R_t is the upper limit of summation where the 95% of the frame's energy is achieved [26].

For the purpose of prediction, the spectral centroid and rolloff are computed across different frames. The average value is computed for both the reference and test recording.

Based on the spectral centroid and spectral rolloff, three types of features were generated. The first one is the average value of spectral centroid and rolloff values calculated for each recording as given by Equations (1) and (2).

The second type of spectral feature was generated by computing the difference between the spectral centroid and spectral rolloff obtained for the reference and test recording, normalised with the reference value. The averaged feature B is given by

$$B = \frac{1}{M} \sum_{i=1}^{M} b_i , \qquad (3)$$

where i is the frame number, b the basic feature calculated for frame b_i and M the total number of frames present in the audio excerpt.

The third feature is generated by applying a rescaling in order to bring the computed features to be in the same range since the values vary from recording to recording. The following equation was used to calculate the rescaled features (B_{rsc}):

$$B_{rsc} = \frac{\mathrm{abs}(B_x) - B_r}{1 - B_r}, \qquad (4)$$

where B_r and B_x are the basic features computed for the reference and test recording respectively.

4.1.2. Centroid of coherence:

Another type of spectral feature is generated by finding the spectral coherence between the reference and the test recording. The function 'mscohere' available in the MATLAB was employed for the computation of the spectrum [27]. The spectral coherence of two audio signals are given as

$$C_{xy} = \frac{|Pxy(f)|^2}{Pxx(f)Pyy(f)}, \qquad (5)$$

The centroid of the coherence spectrum is used as a feature to predict the quality attributes.

4.2. Spatial Features

4.2.1. IACC based spatial features:

The purpose of the spatial features is to model the changes in spatial impression between the reference and the test recording. The spatial features were used as variables to predict the basic audio quality and its

attributes in the regression model. There were three types of spatial features extracted.

For creating the first set of spatial features, the recordings were converted to binaural signals by convolving the multichannel recordings with the HRTF database measured by Gardner and Martin [28]. The synthesised binaural recordings were created for head positions at 0, 30, 60, 90, 120, 150 and 180 degree head positions. From these binaural signals, three types of interaural cross correlation (IACC) based features were extracted. The first type is the broadband IACC, which is just the IACC calculated over the entire bandwidth of the binaural signal at 0 degree head position. The second type of IACC based feature is calculated from the broadband IACC feature, by applying the rescaling given by Equation (4). For creating the third type of IACC based feature, the low frequency band IACCs (at centre frequencies 500Hz, 1000Hz, and 2000Hz) were calculated and the maxima among these were taken. These maxima were rescaled to form a feature. These features were computed with binaural signals for all the head positions mentioned above. IACC is a useful physical correlate of source spaciousness or the subjective phenomenon of apparent source width and these IACC based features represent some spatial changes between the reference and test recording [29].

4.2.2. Back to Front Energy Ratio:

Morimoto describes a relationship between the spatial impression and the loudspeaker energies in a multichannel audio setup [30]. Morimoto and other researchers often used the Front-Back energy ratio as a descriptor of audio spaciousness. It is the representation of the energy distribution in the soundfield of the multichannel audio setup. The back to front energy ratio is taken as another type of spatial feature here for the prediction since the spatial impression is an attribute of the basic audio quality. The decision to use Back-to-Front energy ratio was to avoid a possibility of division by zero if there is a zero energy in the rear channels (as described earlier in this paper(see Table 2), in the recording database there is no zero energy in the front but there are zero energy in the back). The expression given below is used to compute this feature.

$$\text{BF ratio} = \frac{BK}{FT}, \qquad (6)$$

Two versions of BF ratio were computed. In the first version, BK represented the average energy in the rear channels and FT represented the average energy in the rear channels. In the second version, the BK and FT represented the sum of the energies of rear and front channels respectively.

4.2.3. Lateral Gain:

According to Soulodre *et al*, lateral gain (LG) is considered to be the objective measurement of the listener envelopment [31]. The purpose of this spatial feature was to represent the sense of spaciousness, one of the attribute of basic audio quality. It was computed using the following expression

$$\text{LG} = \frac{\int_{x}^{\infty} P_F^2(t)dt}{\int_{0}^{\infty} P_O^2(t)dt}[dB], \qquad (7)$$

Where P_F is the energy measured through a figure of eight microphone and P_O is the energy measured through an omni directional microphone. The value of x used for the integration was 80msec.

The feature based on LG was computed in three different ways. The first feature was computed directly using Equation (7) whereas the second feature was calculated as a difference of the LG computed for the reference and the test recording. The third type of feature based on LG was computed using the rescaling technique mentioned previously in this paper (Equation (4)).

The aforementioned features (see Table 4 in Appendix) and a set of selected interactions between them were used for the prediction of basic audio quality. The next two sections describe the prediction in detail. In the next section, a method which uses the aforementioned features to predict the basic audio quality directly in a regression model is described. Section 6 describes a method in which the basic audio quality is predicted indirectly by first predicting timbral fidelity, frontal spatial fidelity and surround spatial fidelity and then estimating basic audio quality using the regression model presented in [20].

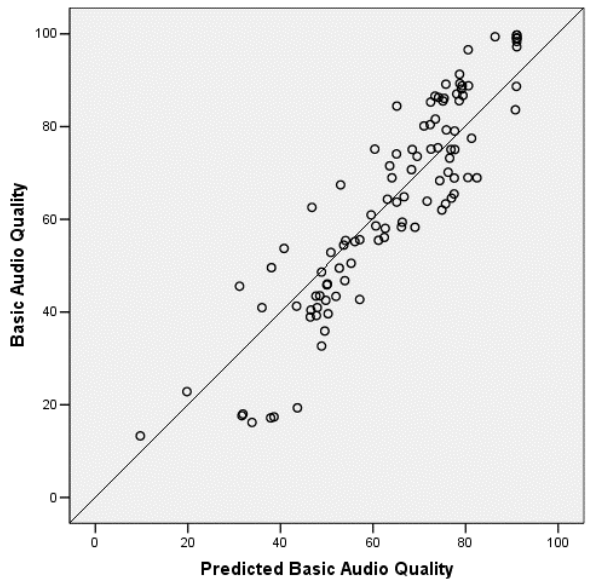

Figure 7 Scatter plot of the calibration results (BAQ prediction, method 2)

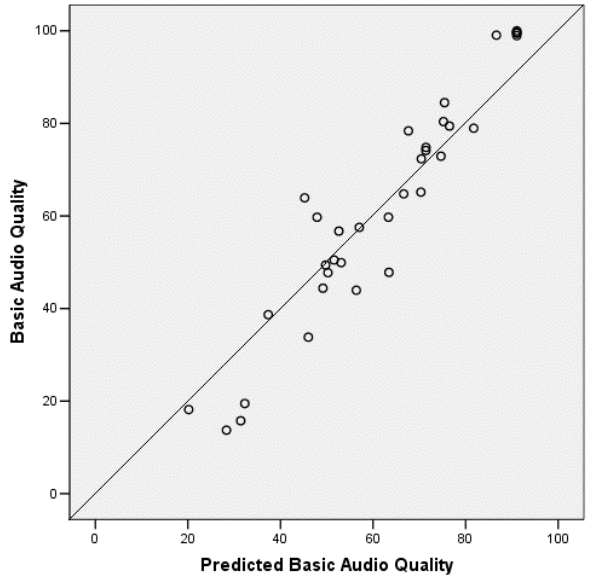

Figure 8 Scatter plot of the cross validated results (BAQ prediction, method 2)

5. PREDICTION OF BASIC AUDIO QUALITY: METHOD 2

The method 1 described in Section 3 is insufficient to predict the basic audio quality of multichannel recordings with degradations other than the band limitation of audio channels. The approach followed in the method presented here is quite straightforward and effective. As described in Section 4, a number of features were extracted from the audio recordings and used in a linear regression model to do the prediction of BAQ.

The initial predictions using multiple linear regression revealed that the extracted features were highly correlated to each other since they attempted to model the same properties of the audio signals. The values of the variance inflation factor (VIF) from multiple linear regression models were very high indicating the multicollinearity problem. Multicollinearity is the situation in which two or more predictors are strongly correlated to each other, making it difficult to interpret the strength of the effect of each predictor in the regression model. Hence it was decided to use ridge regression as an alternative method to tackle the multicollinearity problem. The 'RIDGE' macro available in the SPSS programme was used as a tool to perform the ridge regression. An iterative regression analysis strategy was adopted in order to get a stable and simplified model, which is described in the following paragraphs.

In the first iteration, all the extracted features were included in the ridge regression model. A set of features, exhibiting high Beta values (those features with Beta values greater than 0.075 were selected, since the Beta values of other features were comparatively very low) were selected as a first step of simplification, since they were important and found to be statistically significant at $p < 0.05$. The selected features are given in Table 5 (see Appendix).

In the second iteration, only the features found to be important in the first iteration were included in the regression model. Table 6 shows the results from the ridge regression. From the table it is seen that the interaction between the centroid of coherence and the back to front ratio (COHBFR), and the difference of spectral centroid (CENT_DIF) were found to be relatively unimportant since they showed a low Beta values compared to that of other features. Hence, it

was decided to remove these two features from the regression model. Finally, a simplified model (see Table 7) with rescaled spectral centroid (CENT_RSC), rescaled rolloff (ROLL_RSC), difference rolloff (ROLL_DIF), broadband IACC at 0 degree head position (CORRBB0), centroid of coherence(COH), maxima of low frequency band IACCs at 30 and 150 degree head positions(IACC30 and IACC150) were used for the prediction and it showed a correlation of 0.91 with a calibration error of 9.7 %. The scatter plot of the calibration is given in Figure 7.

The validity of the model mentioned above was tested using a subset of the database as described earlier in this paper. The cross validation showed that the scores were highly correlated (R = 0.96 and SE = 8.6%). The scatter plot of the cross validated results are shown in Figure 8.

6. PREDICTION OF BASIC AUDIO QUALITY: METHOD 3

In this method, an indirect approach is used to predict the basic audio quality. Rumsey et al. describe the influence of timbral fidelity, frontal spatial fidelity and the surround spatial fidelity on basic audio quality [20]. They come up with a regression equation to predict the basic audio quality using the three attributes as independent variables. The regression equation is as follows

$$BAQ = 0.80\ Timbral + 0.30\ Frontal + 0.09\ Surround - 18.7, \quad (8)$$

where BAQ is the basic audio quality, Timbral is the timbral fidelity, Frontal is the frontal spatial fidelity and Surround is the surround spatial fidelity obtained from the listening tests.

The method described here predicts these three attributes independently using the extracted features and uses the regression Equation (8) to predict the BAQ.

The description of the prediction model used for timbral fidelity is given in the following paragraphs. The model used for the prediction of frontal spatial fidelity and surround spatial fidelity were highly correlated to listening test scores and the accuracy was tested with a separate validation experiment (not presented in this paper). A detailed coverage of the prediction results of frontal spatial fidelity and surround spatial fidelity are given in [32].

6.1. Prediction of timbral fidelity

For predicting the timbral fidelity, the extracted features mentioned in Section 4 were applied to a regression model. Informal prediction using the multiple linear regression showed that the extracted features were highly correlated to the listening test scores. Hence it was decided to use ridge regression, as a more robust solution to act against the multicollinearity problem.

The iterative process followed for the prediction of basic audio quality described in the previous section has showed that the features rescaled spectral centroid feature (CENT_RSC), spectral centroid difference feature (CENT_DIFF), rescaled rolloff (ROLL_RSC) rolloff difference feature (ROLL_DIFF), centroid of coherence (COH), and the interaction between the centroid of coherence and a BF ration (COHBFR) were found to be important and statistically significant. It can be seen that all features except COHBFR represent spectral characteristics of the recordings. The interaction feature COHBFR also can be considered as a feature that represents spectral characteristics since it is an interaction of the spectral feature COH. None of the spatial features were found to be statistically important. In fact, as already reported here, there are recordings other than bandlimited type of degradation in the listening test database. Hence, a simple informal prediction using linear regression has been done in order to check whether any of the spatial features show any relations with the listening test scores. This revealed that the broadband IACC has high correlation with the scores.

Another informal prediction has come up with a model that has a correlation of 0.95 with a standard error of 7.7 %. The model used the rescaled spectral centroid (CENT_RSC), difference spectral centroid (CENT_DIFF), rescaled rolloff (ROLL_RSC) difference rolloff (ROLL_DIFF), centroid of coherence (COH) and rescaled broadband IACC at 0 degree head position (CORRBB0) as independent variables. Also, the features used for the prediction were found to be statistically significant at $p < 0.05$. But, high correlation among the features and the high VIF values of revealed that the model is exhibiting a high multicollinearity problem [33]. Hence a model has been built by ridge regression with the

aforementioned features used for the informal prediction. The model is given the Table 8. It showed a correlation of 0.92 with a standard error or 9.5 % for the calibration. It also showed a correlation of 0.94 with a standard error of 8.9% for the cross-validation database as described in section 1. The scatter plots for the calibration and the cross validation are presented in Figures 9 and 10 respectively.

After estimating the three attributes, the results are used to predict the basic audio quality using Equation (9). The prediction showed a correlation of 0.92 with a standard error of 9.5 %. Figure 11 shows the scatter plot.

In the next section, the comparison of the models presented in Sections 3, 5 and 6 is presented.

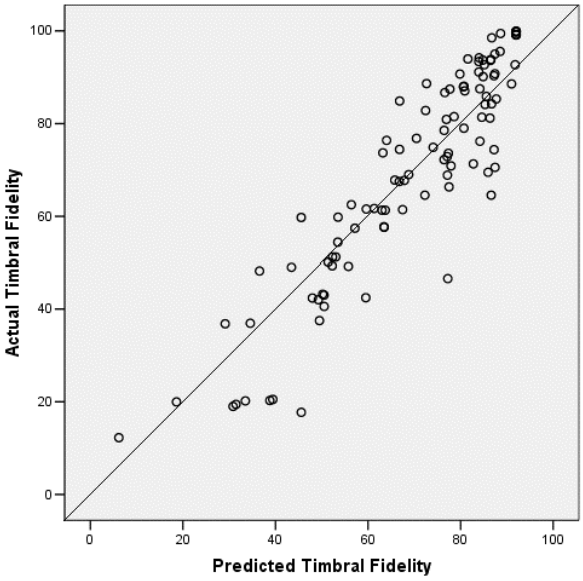

Figure 9 Scatter plot of the calibration results (Timbral fidelity prediction, method 3)

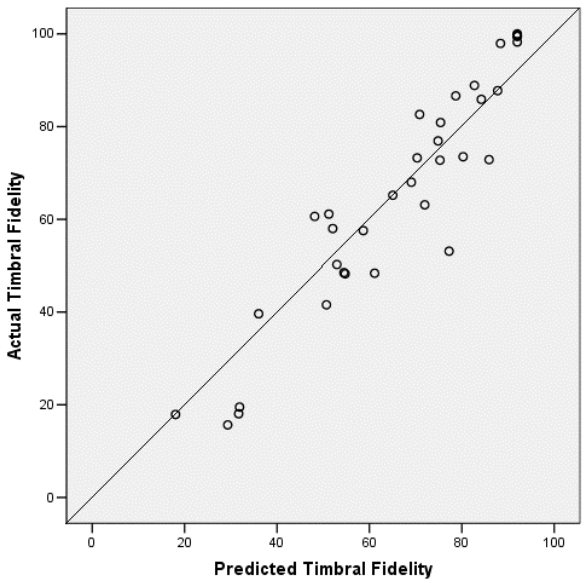

Figure 10 Scatter plot of the cross-validated results (Timbral fidelity prediction, method 3)

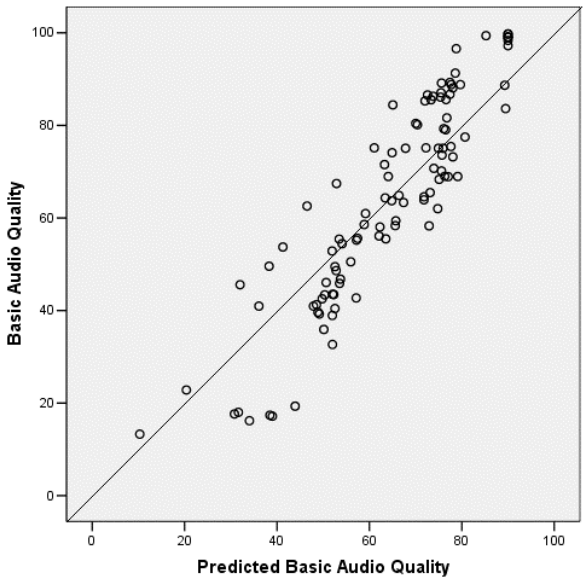

Figure 11 Scatter plot of the calibration results (BAQ prediction, method 3))

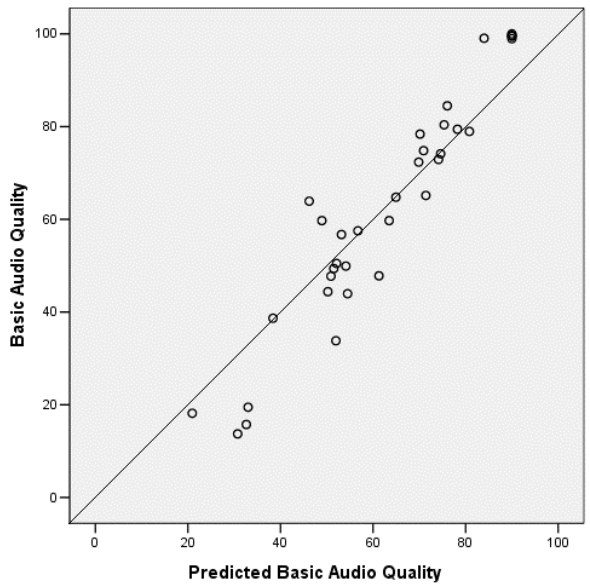

Figure 12 Scatter plot of cross-validation results (BAQ prediction, method 3))

7. DISCUSSION

The three methods describe models to predict the basic audio quality. Table 9 gives the summary of results from the three models.

	Calibration		Cross-validation	
	Correlation R	Standard Error SE(%)	Correlation R	Standard Error SE
Model 1	0.97	5.92	-	-
Model 2	0.91	9.67	0.96	8.64
Model 3	0.92	10.00	0.96	9.11

Table 9: Summary of BAQ prediction models

The first model is limited in its capability to predict the BAQ of all sorts of degradations. The model was only capable to predict the BAQ of the recordings with bandlimited items. In real situations, any type of degradation can happen such as down-mixing. But, the model showed a very high correlation with a small standard error of prediction to the listening test scores. This model is suitable for broadcasters to monitor the BAQ in real time although its capability is limited (it cannot be used when other types of degradations are employed in such situations). Hence, this model needs to be modified so that it can detect quality degradation due to other types of degradations such as down-mixing.

The second and third models tried to predict the BAQ of downmixed recordings as well to a certain extent. These models compare the test recording with a reference recording and created features to do the prediction.

In the second method, it can be seen that most of the direct spectral features were important. In the work of Rumsey et al. [20] it is reported that the most important attribute of BAQ is timbral fidelity. The importance of the spectral features in this model supports this fact. The next important attribute of basic audio quality reported in [20] is frontal spatial fidelity. From Section 5, it can be seen that the IACC based features at 0 and 30 degree (CORRBB0 and IACC30) head positions were statistically important for the prediction of BAQ. These two features can be considered to be a basic representation of the frontal spatial fidelity attribute since they were computed at the head position inside the frontal arc (see Figure 1), although the angle 30 degree is just the border of the frontal arc. However the importance of IACC30 is rather less compared to that of IACC0. Also, two other IACC features computed at 90 and 150 degree head positions(CORRBB90, IACC150) were found to be important and statistically significant for this model (See Section 5) The importance of these two features correspond to the third attribute of BAQ, surround spatial fidelity as reported in [20]. In addition, among the spectral features, COH showed the highest importance according to the regression model, which means that it can be used as an effective feature to represent the spectral changes.

The second model showed a high correlation with the listening test scores. The cross-validation also showed a high correlation and low error of estimation, which means that the significant features found in the model, can be used for the prediction of basic audio quality. However, it would be ideal to use the model after testing the consistency with the data obtained from an independent listening test.

The third method also showed a similar trend to that of the model presented in Section 5. A similar set of features were found to be important for the three attributes of basic audio quality as seen in the previous paragraphs. For timbral fidelity, all spectral features without any interactions were found to be

important. It was also found that one of the spatial features was important in the prediction of timbral fidelity. This is due to reason that there were recordings with degradations causing the spatial changes. The important features that are statistically significant for the prediction of frontal spatial fidelity and surround spatial fidelity are described in [32].

In summary, model 1 showed a very high correlation with a small stand error of estimate. However it needs to be modified so as to predict the BAQ of recordings with other broader range of audio quality degradations. The models 2 and 3 showed a very similar trend in the prediction because of the very close values of correlations and the standard error of estimate obtained for the calibration and the cross validation. To generalise and to verify these findings, the models need to be tested with a separate validation experiment.

8. CONCLUSIONS AND FUTURE WORK

In this paper, three methods were proposed to predict the basic audio quality of multichannel audio recordings. The first one can be classified as an unintrusive method as it does not require comparison of tested recording with its reference. Hence, it can be used in broadcasting applications (with certain modifications) where a real-time monitoring of audio quality is needed. The other two methods performed in a very similar way in terms of accuracy. These models revealed the underlying relationship of different spectral features and IACC based features between the basic audio quality and its three attributes, timbral fidelity, frontal spatial fidelity and surround spatial fidelity. However, the obtained regression models need to be checked for its consistency with a separate validation experiments in order to verify how generalisable the models are. Since the audio quality degradations used in the listening tests were of basic types (band-limiting and down-mixing), in order to generalise the models for its universal applicability, more degradations need to tested and features that represent those degradations still have to be found. This may lead to the realisation of the multichannel "PEAQ" model to predict the basic audio quality of multichannel audio. Although the experiments described in this paper were incorporated only 5.1 surround recordings, the prediction capability of the obtained regression models can be extended for more complex multichannel audio setups such as WFS, with some modifications to support them.

9. REFERENCES

[1] M. R. Schroeder, B. S. Atal, J. L. Hall "Optimising digital speech coders by exploiting masking properties of human ear" in J. Acoust. Soc. Am. Vol. 66 pp 1647-1652 Dec. 1979.

[2] J. Karjalainen, "A new auditory model for the evaluation of sound quality of audio system", in Proceedings of the ICASSP, Tampa, Florida, pp 608-611, March 1985.

[3] K. Brandenburg, "Evaluation of quality for audio encoding at low bit rates", in 82nd AES Convention, London, preprint 2433, 1987.

[4] T. Thiede and E. Kabot, "A New Perceptual Quality Measure for the Bit Rate Reduced Audio," presented at the 100th Convention of the Audio Engineering Society, J. Audio Eng. Soc. (Abstracts), vol. 44, p. 653 (1996 July/Aug.), Preprint 4280.

[5] T. Sporer, "Objective Audio Signal Evaluation--Applied Psychoacoustics for modelling the Perceived Quality of Digital Audio," presented at the 103rd Convention of the Audio Engineering Society,

[6] J. G. Beerends, J. A. Stemerdink, "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation", in J. Audio Eng. Soc., vol. 40, pp. 963-978, Dec. 1992.

[7] B. Paillard, P. Mabilleau, S. Morisette, J. Soumagne, "Perceval: Perceptual Evaluation of the Quality of Audio Signals", in J. Audio Eng. Soc., vol. 40, pp. 21- 31, Jan. 1992.

[8] C. Colomes, M. Lever, J. B. Rault, Y. F. Dehery, "A Perceptual Model Applied to Audio Bit-Rate Reduction", in J. Audio Eng. Soc., vol. 43, pp. 233- 240, April 1995.

[9] ITU-R Recommendation BS.1387-1, "Method for Objective Measurements of Perceived Audio Quality", 1998.

[10] T. Thiede et al. "PEAQ- The ITU Standard for Objective Measurement of perceived Audio Quality". in J. Audio Eng. Soc., vol. 48, No. 1/2, January/February 2000).

[11] P. Kozlowski and A.B. Dobrucki "Proposed Changes to the methods of Objective, Perceptual Based Evaluation of Compressed Speech and Audio signals" presented at AES convention 116th convention, Berlin, Germany, 2004 May 8-11 Paper 6085.

[12] P. Kozlowski and A.B. Dobrucki "Adjustment of Parameters Proposed for the Objective, Perceptual Based Evaluation Methods of Compressed Speech and Audio Signals." Presented at AES convention 117th convention, San Francisco CA, 2004 October 28-31 Paper 6286.

[13] B. Fieten et. al, "Audio Adaptation According to Usage Environment and Perceptual Quality Metrics." in IEEE transactions on Multimedia, Vol. 7, No. 3 June 2005

[14] J.G.A. Barbedo and A. Lopes "Strategies to Increase the Applicability of Methods for Objective Assessment of Audio Quality" presented at AES 116th Convention, Berlin, Germany, 2004 May 8–11 Paper 6080.

[15] S. Torres-Guijarro et. al. "Coding Strategies and quality measure for multichannel audio". Presented at AES 116th Convention, Berlin, Germany, 2004 May 8–11 Paper 6114.

[16] R. Vanam and C. D. Creusere "Evaluating Low Bitrate scalable audio quality using advanced version of PEAQ and Energy Equalisation Approach" in Acoustics, Speech, and Signal Processing, 2005. Proceedings (ICASSP '05). IEEE International Conference on Volume 3, March 18-23, 2005 pp189 – 192.

[17] S. George, S. Zielinski, F. Rumsey "Prediction of Basic Audio Quality for multichannel audio recordings: Initial developments" presented at Digital Music Research Network Workshop and Roadmap Launch, 21 December 2005.

[18] T. Letowski, "Sound Quality Assessment: Concepts and criteria" Presented at the 87th convention 1989 October 18-21 NewYork- Preprint No. 2825.

[19] ITU-R BS 1534 - MUSHRA-EBU, "Method for Subjective Listening Tests of Intermediate Audio Quality," Draft EBU Rec. B/AIM 022 (Rev. 8)/BMC 607rev, European Broadcasting Union (2000 Jan.).

[20] F. Rumsey, S. Zielinski, R. Kassier and S. Bech. "On the relative importance of spatial and timbral fidelities in judgements of degraded multichannel audio quality." in J. Acoust. Soc. Am. 968-976, Vol. 118, August 2005.

[21] S. Zielinski, F. Rumsey, R. Kassier and S. Bech. "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" in J. Audio Eng. Soc., Vol. 53, No. 3, 2005 March.

[22] S. Zielinski, F. Rumsey, R. Kassier and S. Bech. "Comparison of quality degradation effects caused by limitation of bandwidth and by down-mix algorithms in consumer multichannel audio delivery systems," presented at 114th AES Convention, Amsterdam, 22–25 March, Paper 5802.

[23] S. Zielinski, F. Rumsey, R. Kassier and S. Bech. "Effects of down-mix algorithms on quality of surround sound," 780–798 in J. Audio Eng. Soc., Vol. 51, No. 9, 2003 September.

[24] S. Zielinski., F. Rumsey, R. Kassier, and S. Bech. "Development and Initial Validation of a Multichannel Audio Quality Expert System." J. Audio Eng. Soc. Vol. 53, 1/2, pp. 4-21, (January/Feburary 2005).

[25] G. Tzanetakis, P. Cook, "Musical Genre Classification of Audio Signals" in IEEE Transactions On Speech And Audio Processing, Vol. 10, No. 5, July 2002 pp 293-302.

[26] G. Tzanetakis, P. Cook, "Multifeature Audio Segmentation For Browsing And Annotation" in Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999.

[27] Description of 'mscohere' function in Matlab Help ver. 7.0

[28] B. Gardner, K. Martin "HRTF Measurements of a KEMAR Dummy-Head Microphone " Available: http://sound.media.mit.edu/KEMAR.html, Janurary 10, 2006

[29] F. Rumsey, Spatial Audio, (Focal Press, Oxford, UK, 2001),

[30] M. Morimoto "The Role of Rear Loudspeakers in Spatial Impression", Presented at the 103rd AES Convention, 1997 September 26-29 New York.

[31] G. Soulodre, M. Lavoie and S. Norcross, "Objective Measures of Listener Envelopment in Multichannel Surround Systems" in J. Audio Eng.Soc., Vol.51, No.9, 2003 September.

[32] S. George "Report on the prediction of frontal and surround spatial fidelity", Institute of Sound Recording, University of Surrey January 2006.

[33] A. Field "Discovering statistics using SPSS", (SAGE publication UK, second edition, 2005).

[34] D. Montgomery et. al, Introduction to Linear Regression Analysis (Willey Interscience 2001, third edition).

[35] N. R Draper, H. Smith "Applied Regression Analysis" Second edition, Wiley series 1981

[36] K. H. Esbensen, Multivariate analysis in practice, (Camo Process A S, second edition, 2002).

[37] "Methods for subjective assessment of small impairments in audio systems including multichannel sound systems". International Telecommunications Union (1994), ITU-R Recommendation BS. 1116.

10. APPENDIX

No.	Name	Low pass filter cut-off frequency	Used for
1	All 3500Hz	L, R, C, LS, RS – 3.5kHz	F-B, F-F
2	All 8000Hz	L, R, C, LS, RS – 8kHz	F-B, F-F
3	All 12000 Hz	L, R, C, LS, RS – 12kHz	F-B, F-F
4	Hybrid A	L, R – 20kHz; C – 10kHz; LS, RS – 5kHz	F-B
5	Hybrid B	L, R – 20kHz; C – 13kHz; LS, RS – 3.5kHz	F-B
6	Hybrid C	L, R – 18.25kHz; C – 3.5kHz; LS, RS – 10kHz	F-F
7	Hybrid D	L, R – 14.125kHz; C – 3.5kHz; LS, RS – 14.125kHz	F-F
8	Hybrid E	L, R – 13kHz; C – 7kHz; LS, RS – 3.5kHz	F-B
9	Hybrid F	L, R – 10kHz; C – 13kHz; LS, RS – 3.5kHz	F-B
10	Hybrid G	L, R – 11.25kHz; C – 3.5kHz; LS, RS – 7kHz	F-F
11	Hybrid H	L, R – 9.125kHz; C – 3.5kHz; LS, RS – 9.125kHz	F-F

Table 1: Bandlimited versions (for calibration experiment)

No.	Degradation name	Algorithm	Used for
1	3/0	Down mixed to 3 channels in front	F-B, F-F
2	2/0	Downmixed to 2 channels in front	F-B, F-F
3.	1/0	Downmixed to 1 channel in front	F-B, F-F
4	1/2	Downmixed the L and R to C and LS and RS kept unchanged.	F-F

Table 2: Down-mixed versions (for calibration experiment)

Sub-set No.	Spatial Scene	Dialogue in the centre channel	Equal Bandwidth in all channels	Regression Equation
1	F-B	No	No	Q = 3.26lr + 0.24c + 0.38sur + 18.9
2	F-B	Yes	No	Q = 0.31lr + 3.51c - 0.4sur + 27.7
3	F-F	No	No	Q = 2.38lr + 0.15c + 1.43sur + 17.8
4	F-F	Yes	No	Q = 0.24lr + 3.09c + 0.32sur + 26.1
5	Any	either	Yes	Q = 5f

Table 3: Regression equations used for BAQ prediction

No.	Acronym	Description
1	CENTROID	Average spectral centroid (ASC)
2	CENT_RSC	Rescaled ASC
3	CENT_DIF	Normalized ASC difference between reference and test recording
4	ROLLOFF	Average spectral rolloff (ARO)
5	ROLL_RSC	Rescaled ARO
6	ROLL_DIF	Normalized ARO difference between reference and test recording
7	CORRBB0	Broadband IACC at 0 degree head position
8	CORRBB90	Broadband IACC at 90 degree head position
9	BFRATIO	Back-to-front energy ratio
10	BFR_MEAN	Ratio of average energies in the rear channels to front channels
11	CORR	Unscaled IACC
12	COH	Spectral coherence
13	IACC0	Maxima of IACC at 0 degree head position
14	IACC30	Maxima of IACC at 30 degree head position
15	IACC60	Maxima of IACC at 60 degree head position
16	IACC90	Maxima of IACC at 90 degree head position
17	IACC120	Maxima of IACC at 120 degree head position
18	IACC150	Maxima of IACC at 150 degree head position
19	IACC180	Maxima of IACC at 180 degree head position
20	LG	Unscaled Lateral gain(LG)
21	LG_DIFF	LG difference between reference and test recording
22	LG_RSC	Rescaled LG

Table 4: List of extracted features

Feature Name	B	SE(B)	Beta	t	95 % Confidence intervals	
					Upper Limit	Lower Limit
CENT_RSC	-17.0728	3.24713	-0.11716	-5.2578	-10.7084	-23.4372
CENT_DIF	-15.3167	3.38775	-0.10273	-4.52118	-8.67667	-21.9567
ROLL_RSC	-17.5996	3.24670	-0.11403	-5.42076	-11.236	-23.9631
ROLL_DIF	-17.6144	3.36113	-0.11097	-5.24063	-11.0266	-24.2023
CORRBB0	-5.93685	2.10783	-0.07657	-2.81657	-1.8055	-10.0682
CORRBB90	-8.44324	2.89496	-0.09574	-2.91653	-2.76912	-14.1174
COH	0.00293	0.00030	0.390968	9.726523	0.003521	0.00234
IACC30	-7.18019	1.72041	-0.09004	-4.17353	-3.80818	-10.5522
IACC150	-6.75215	2.11823	-0.08113	-3.18762	-2.6004	-10.9039
COHBFR	0.000684	0.00019	0.104093	3.489505	0.001068	0.0003

Table 5: Basic audio quality (first iteration features)

Feature Name	B	SE(B)	Beta	t	95 % Confidence intervals	
					Upper Limit	Lower Limit
CENT_RSC	-15.0638	3.30477	-0.10338	-4.5582	-8.58645	-21.5412
CENT_DIF	-14.0792	3.38251	-0.09443	-4.16234	-7.44944	-20.7089
ROLL_RSC	-20.63	3.29628	-0.13366	-6.25856	-14.1693	-27.0907
ROLL_DIF	-21.0931	3.51433	-0.13289	-6.00203	-14.205	-27.9812
CORRBB0	-11.6893	2.53875	-0.15076	-4.60433	-6.71332	-16.6652
CORRBB90	-9.02736	3.24467	-0.10236	-2.78221	-2.66779	-15.3869
COH	0.00334	0.00028	0.446479	11.78456	0.003903	0.00279
IACC30	-9.62087	2.04357	-0.12065	-4.70786	-5.61546	-13.6263
IACC150	-11.9577	2.43561	-0.14368	-4.90954	-7.18394	-16.7315
COHBFR	-1.4E-05	0.00023	-0.00214	-0.05932	0.000451	-0.00048
Constant	50.9350	3.25210	0	15.66218	57.30919	44.56094

Table 6: Basic audio quality (second iteration features)

Feature Name	B	SE(B)	Beta	t	95 % Confidence intervals	
					Upper Limit	Lower Limit
CENT_RSC	-20.5064	4.44381	-0.14072	-4.61459	-11.7965	-29.2163
ROLL_RSC	-23.7293	3.12149	-0.15374	-7.6019	-17.6111	-29.8474
ROLL_DIF	-24.4299	3.38811	-0.15391	-7.21046	-17.7892	-31.0706
CORRBB0	-11.8011	2.49141	-0.1522	-4.73673	-6.91797	-16.6843
CORRBB90	-9.3678	3.19895	-0.10622	-2.92839	-3.09784	-15.6378
COH	0.00341	0.00027	0.455351	12.29367	0.003957	0.002869
IACC30	-9.42847	1.99226	-0.11824	-4.73254	-5.52363	-13.3333
IACC150	-11.9392	2.38412	-0.14346	-5.00782	-7.26636	-16.6121
Constant	50.0981	3.03076	0	16.52983	56.03835	44.15775

Table 7: Basic audio quality (third iteration features)

Feature Name	B	SE(B)	Beta	t	95 % Confidence intervals	
					Upper Limit	Lower Limit
cent_rsc	-17.9649	3.16291	-0.11949	-5.67988	-11.7656	-24.1643
cent_dif	-16.4782	3.29091	-0.10712	-5.00719	-10.028	-22.9284
roll_rsc	-21.4166	3.21940	-0.13449	-6.65234	-15.1065	-27.7266
roll_dif	-23.5711	3.36295	-0.14394	-7.00904	-16.9797	-30.1625
coh	0.00314	0.00027	0.406149	11.55177	0.003673	0.002608
corrbb0	-10.816	2.65636	-0.13521	-4.07172	-5.60949	-16.0224
Constant	54.3078	2.99468	0	18.13471	60.17738	48.43821

Table 8: Timbral fidelity model