PREDICTION OF PERCEIVED QUALITY IN MULTI-CHANNEL AUDIO COMPRESSION CODING SYSTEMS

IN YONG CHOI[1], BARBARA G. SHINN-CUNNINGHAM[2], SANG BAE CHON[3], KOENG-MO SUNG[4]

[1,3,4] *Institute of New Media and Communications, Seoul National University, Seoul, Korea*
{ciy,strlen,kmsung}@acoustics.snu.ac.kr
[2] *Hearing Research Center, Boston University, Boston, USA*
shinn@cns.bu.edu

Objective quality assessment methods, such as described in ITU-R Recommendation BS.1387-1 [1], have been widely used for evaluation of audio coding systems. However, even though many different multi-channel audio compression coding systems are being developed, most current quality assessment methods only predict results for monaural or stereo signals. In this paper, a prediction method is introduced that can be used for the objective quality assessment for multi-channel audio compression coding systems. The method introduces two novel variables, interaural level difference distortion (ILD distortion) and interaural cross-correlation coefficient distortion (IACC distortion) to predict degradations in spatial quality. Simultaneously, five Model Output Variables proposed in ITU-R BS.1387-1 are selectively extracted from binaural signals that are synthesized using binaural room transfer functions. The prediction model is trained and verified using results from subjective listening tests of multi-channel audio compression coding systems that were performed by participants in MPEG audio group. This new model, using the two interaural and five non-spatial statistics, shows encouraging results in prediction perceived quality.

INTRODUCTION

Low bit-rate audio coding technology now is being used in multi-channel audio compression technologies that manipulate the spatial impressions of the listener. Recently, ISO/IEC MPEG standardized a Binaural Cue Coding type [2] multi-channel audio coder that has low bit-rate but relatively high quality [3]. As the number of competing compression coding systems increases, reliable quality assessment becomes important for evaluating these systems. Because a good predictive or objective assessment model would enable easy comparison of the different compression schemes, numerous objective quality assessment methods have been proposed [4]. Two recent models for the objective assessment of spatial quality or quality of multi-channel sound sources have been proposed [5, 6]. However, to date, satisfactory predictions of perceptual quality of newly developed low bit-rate multi-channel coding systems have not been reported.

An adequate predictive model of sound quality must satisfy the following conditions. First, the listening environment for the multi-channel audio reproduction system must be modelled. Second, not only timbral degradations but also spatial degradations, such as sound localization errors, must be quantified. Lastly, the model must be trained and verified with reliable judgments of sound quality taken from listening tests using a large ensemble of different kinds of degradations in spatial and timbral quality.

In this paper, a prediction model is introduced that can be used for the objective quality assessment of multi-channel audio compression coding systems. In our method, multi-channel signals were first converted into binaural signals using binaural room transfer functions (BRTFs) measured in a listening room assuming a standard layout of multi-channel audio reproduction systems. After psychoacoustical processing of the binaural signals, interaural level difference distortion (ILD distortion [7, 8]) and interaural cross-correlation coefficient distortion (IACC distortion [8]) are computed in order to quantify degradations in spatial quality. Simultaneously, five Model Output Variables (MOVs) in ITU-R BS.1387-1 are selectively computed from the binaural signals for assessment of timbral quality. The prediction model is trained and verified using results of listening tests with multi-channel audio compression coding systems that were performed by participants in the MPEG audio group [9, 10].

In Section 1, the implementation of the prediction model is illustrated. The prediction model is logically divided into three sequential parts: a binaural hearing model, a peripheral ear model, and a cognition model. Those three parts are described in the three sub-sections of Section 1, respectively. The procedures for training and verification of the model are described in Section 2. The listening test database used in training and verification are also described in detail in Section 2. The verification results and future directions for this work

are discussed in Section 3. Finally, conclusions are given in Section 4.

1 MODEL IMPLEMENTATION

1.1 Overall process

In the field of perceived quality assessment for sound reproduction systems, Basic Audio Quality (BAQ) is commonly used [4]. The prediction model introduced in this paper also estimates BAQ, using a combination of interaural and spectral measures. BAQ is measured by presenting listeners with a pair of stimuli, a reference audio signal and the test signal (the reference signal processed by some coding scheme or other transmission channel) and asking them to report a single value that estimates the degradation of the test signal compared to its reference. In the database used for the training and verification of our model, the BAQ is represented by a 'Mean Opinion Score (MOS),' a value ranging from zero to one hundred points. The goal of our model is to predict the average MOS reported by listeners. In the current study, the input to the model is two multi-channel signals representing the test and the reference signals.

The overall structure of our prediction model is illustrated in Figure 1.

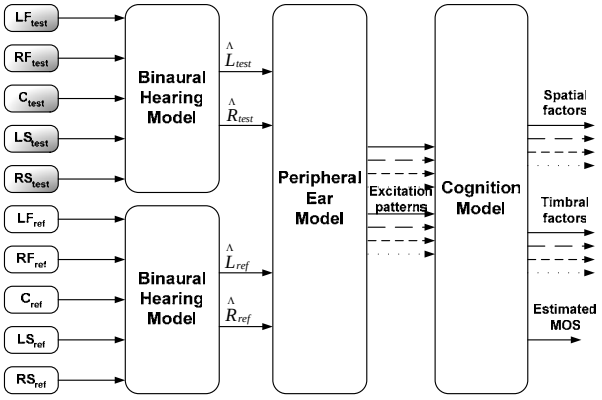

Figure 1: Overall structure of our prediction model

The process consists of a binaural hearing model, a peripheral ear model, and a cognition model. The binaural hearing model synthesizes the signals that a listener would receive if the multi-channel signal was played back to the listener in a standard, multi-speaker configuration in a standard listening space. The peripheral ear model transforms the binaural input signals into separate frequency channels, roughly approximating the excitation patterns that these signals would cause on the basilar membrane [11]. Lastly, the cognition model processes the excitation patterns to extract multiple interaural and spectral features from which the MOS is predicted. Through these stages, acoustic information is serially processed – the information flow from the multi-channel sound reproduction systems to the judgment by the central nervous system of the sound quality occurs in sequential order.

The implementations of above three models are described in following three sub-sections, respectively.

1.2 Binaural hearing model

In typical multi-channel audio cases, both the reference and test signals will consist of five signals for the five channels in the reproduction system. In our binaural hearing model, binaural signals representing the total left and right signals reaching the listener for the test and reference inputs (denoted by subscript *Test* and *Ref*, respectively) are synthesized by convolving each of the relevant five channel inputs with the pair of BRTFs corresponding to the appropriate loudspeaker location for that channel. The five resulting binaural signals then are summed to produce the total binaural signal that the listener would hear. Thus, the binaural test and reference signals are synthesized as shown in (1).

$$\begin{pmatrix} \hat{L}_{Test} & \hat{R}_{Test} \\ \hat{L}_{Ref} & \hat{R}_{Ref} \end{pmatrix} = \begin{pmatrix} H_{LfL} & H_{RfL} & H_{CL} & H_{LsL} & H_{RsL} \\ H_{LfR} & H_{RfR} & H_{CR} & H_{LsR} & H_{RsR} \end{pmatrix} \begin{pmatrix} LF_{Test} & LF_{Ref} \\ RF_{Test} & RF_{Ref} \\ C_{Test} & C_{Ref} \\ LS_{Test} & LS_{Ref} \\ RS_{Test} & RS_{Ref} \end{pmatrix} \quad \text{..... (1)}$$

H_{CL}, H_{LfL}, H_{RfL}, H_{LsL}, H_{RsL}, H_{CR}, H_{LfR}, H_{RfR}, H_{LsR}, H_{RsR} are the BRTFs representing ten hearing paths, such as center channel to left ear, left-front channel to left ear, and so on. $\hat{L}$ and $\hat{R}$ are the left ear input signal and the right ear input signal, respectively.

The ten BRTFs are recorded with a high quality head and torso simulator microphone placed in a multi-channel listening room in the Electronics and Telecommunications Research Institute in Korea, so that the transfer functions include not only the acoustic effects of the head and torso but also the characteristics of multi-channel sound reproduction systems and listening room responses. The geometric configuration of the multi-channel reproduction system was set up to match the standard recommendations in ITU-R BS.1116 [12]. This configuration has the center channel loudspeaker located at zero degrees, the left-front channel and right-front channel loudspeakers at -30 degrees and +30 degrees, respectively, and the left-subsequent channel and the right-subsequent channel loudspeakers at -110 degrees and +110 degrees, respectively.

Most previous quality evaluation models, such as ITU-R BS 1387-1, are designed for monaural sound. When they are used to evaluate stereo signals, these systems separately compare the left and right channels of the test signal to the corresponding channels of the

reference signal. The sound quality is objectively judged separately for the two channels, and these two judgments are averaged to estimate the perceived sound quality. However, that matching scheme is not appropriate for multi-channel signals, since a listener hearing a multi-channel reproduction does not listen to the five signals in isolation, but rather to their combination. Multi-channel signals are generally played in multi-channel reproduction systems with multiple loudspeakers. Thus, the resulting total binaural signals should be compared when listeners judge sound quality. We have found that the MOVs of BS.1387-1 are only weakly correlated (correlation coefficients were in the range between 0.03 and 0.40) with the subjective evaluation data when the sound quality for each of the five channels measured separately and then averaged. However, when perceived quality of the total resultant binaural test signal is judged against the binaural reference signal, quality predictions are much better, with correlations ranging from 0.45 to 0.60.

1.3 Peripheral ear model

Synthesized binaural signals are processed by a peripheral ear model. The peripheral ear model converts ear input signals to a representation like the signals exciting hair cells in the human basilar membrane, which translate mechanical vibrations from acoustic inputs into neurally firing in the auditory nerve fibers.

The peripheral ear model is implemented by taking a Discrete Fourier Transform (DFT), level scaling, filtering to simulate the ear canal resonance, cochlea filter-bank smoothing, adding internal noise, and then spreading to account for temporal and simultaneous masking. In our model, a 2048-point DFT is used. A bank of twenty equivalent rectangular bandwidth (ERB) filters is used to simulate a cochlea filter-bank model. The rest of the peripheral ear model is identical to the "FFT based peripheral ear model" of ITU-R BS.1387-1. See [1] for detailed information.

The output of the peripheral ear model is referred to as the excitation pattern. The excitation pattern encodes loudness patterns, modulation patterns, spectral content, and short-term time-frequency content of the inputs.

1.4 Cognition model

The cognition model extracts multiple factors that have high correlations with human judgments of sound quality. These various factors are computed from the excitation pattern output of the peripheral ear model. For convenience, the factors are conceptually separated into spatial factors and timbral factors in the following sub-sections.

1.4.1 Calculation of factors for spatial quality

Even though the BAQ yields only a single value for one test signal, the sound quality itself has many attributes that contribute to the overall perceived sound quality. For this reason, most prediction models measure several features to quantify the relevant attributes that influence perceived quality.

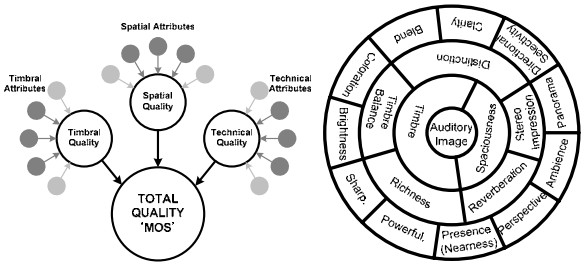

Figure 2: Conceptual illustration of Total Auditory Quality [13] and MUlti-level auditoRy Assessment Language (MURAL) [14]

Figure 2 illustrates the attributes of sound quality used in the MUlti-level auditoRy Assessment Language (MURAL) [14] model. Attributes are divided into two groups, affecting either 'timbre' or 'spatial impression.' More recently, Berg and Rumsey [13] classified the attributes of sound quality into three categories: timbral quality, spatial quality, and technical quality. No matter what kinds of classification are considered, spatial quality is an important part of the perceived sound quality. Especially for multi-channel coding systems, spatial quality is very important.

Degradations of spatial quality can come about from distortion of many different perceptual attributes, including changes in perceived source location, perceived source width, diffuseness, etc. Of these possible spatial degradations, errors in perceived location are taken into account first. It is generally accepted that the most robust and important spatial auditory cues are computed by calculating differences between left and right ears [15]. There are two such interaural differences that are important perceptually: interaural time differences (ITDs) and interaural level differences (ILDs). Although both ITDs and ILDs are important localization cues, those two cues play different roles and have different importance in different frequency regions. However in this initial implementation of our model, ITD distortions are not yet being used. Instead, ILD distortions and IACC distortions are used.

ILD is calculated as ten times the logarithm of the intensity ratio between the left ear input X_L and right ear input X_R from the time-frequency segments in the k^{th} ERB band in the n^{th} time frame.

$$ILD[k,n] = 10\log_{10}\left(\frac{\sum_{l} X_{\mathrm{L}}[l]X_{\mathrm{L}}[l]^{*}}{\sum_{l} X_{\mathrm{R}}[l]X_{\mathrm{R}}[l]^{*}}\right) \quad (2)$$

Accordingly, we can define the distortion of inter-aural level difference as shown in the following equations. $ILD_{test}[k,n]$ and $ILD_{ref}[k,n]$ are the ILD of test signal and original signal, respectively. *ILDDist* is the computed distortion of the ILD:

$$ILDDist[k,n] = w[k, ILD_{ref}[k,n]] \cdot \left| ILD_{test}[k,n] - ILD_{ref}[k,n] \right| \quad (3)$$

$$ILDDist[n] = \frac{1}{Z} \sum_{k=0}^{Z-1} ILDDist[k,n] \quad (4)$$

$$ILDDist = \frac{1}{N} \sum_{n=1}^{N} ILDDist[n] \quad (5)$$

$w[k, ILD_{ref}[k,n]]$ is a non-linear weighting function that varies with frequency-band index k and the ILD of reference signal. The goal of the weighting is to take into account the different contributions (relative importance) of the ILD distortion in each frequency band on the perceived location, which also varies with the original location of the sound source. While this weighting is likely to be important to fine-tune our results, all weights are fixed to equal 1 in our initial implementation.

The IACC is calculated as shown in (6).

$$IACC[k,n] = \mathrm{Re}\left\{ \frac{\sum_l X_{\mathrm{L}}[l] X_{\mathrm{R}}[l]^*}{\sqrt{\sum_l X_{\mathrm{L}}[l] X_{\mathrm{L}}[l]^* \sum_l X_{\mathrm{R}}[l] X_{\mathrm{R}}[l]^*}} \right\} \quad (6)$$

We can derive $IACC_{test}[k,n]$ and $IACC_{ref}[k,n]$, which are the IACC of test signal and original signal for the k^{th} ERB band and the n^{th} time frame, respectively.

IACCDist is calculated in an analogous way to the ILD distortion.

$$IACCDist[k,n] = w[k, IACC_{ref}[k,n]] \cdot \left| IACC_{test}[k,n] - IACC_{ref}[k,n] \right| \quad (7)$$

$$IACCDist[n] = \frac{1}{Z} \sum_{k=0}^{Z-1} IACCDist[k,n] \quad (8)$$

$$IACCDist = \frac{1}{N} \sum_{n=1}^{N} IACCDist[n] \quad (9)$$

IACC and ILD are known to be independent of each other and have different roles in spatial perception [16]. Thus, theoretically, both variables are needed as they measure different aspects of spatial degradation.

The computed measures of ILD and IACC distortion are highly correlated with subjective quality judgments of signals processed through different spatial audio compression codecs, using a wide range of sound sources [9, 10].

1.4.2 Calculation of factors for timbral quality

Five Model Output Variables (MOVs; see [1] for details, e.g. equations and numerical data.) are selected from ITU-R BS.1387-1 to quantify spectral degradations in our model. There are originally sixteen MOVs in BS.1387-1, but only five MOVs were used in order to reduce redundancy and to improve the effectiveness and prediction performance. In our correlation analysis, the five selected MOVs from BS.1387-1 yield consistently high correlation coefficients with quality judgments from the multi-channel listening test database (from 0.45 to 0.60), when measured in binaural signals.

The selected MOVs are described briefly in Table 1.

MOV	Description
ADB	Averaged distortion block. Ratio of total distortion to the total number of distorted blocks.
NMRtotB	Logarithm of the averaged total noise to masker energy ratio
EHS	Harmonic structure of the error
AModDif1B	Averaged modulation difference
NLoudB	Averaged noise loudness

Table 1: MOVs of ITU-R BS.1387-1 that were used as factors for timbral degradations

1.4.3 Estimation of MOS

All of the features explained in above sub-sections are used as inputs of a single layer feed forward neural network for the estimation of MOS. Initially, a pure linear function is used as an activation function of the neural network.

2 TRAINING AND VERIFICATION OF MODEL

2.1 Listening test database

As yet, the data in the listening test database of low bit-rate multi-channel compression coding systems is not widely distributed. However, a valuable database from listening tests of the ISO/IEC MPEG audio group from 2004 to 2005 [9, 10] is available. The MPEG listening tests were performed by volunteers in order to evaluate the sound quality of several low bit-rate multi-channel compression coding systems. The listening tests followed the procedures set out in ITU-R BS.1534 "Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) [17]." Listeners were asked to give Mean Opinion Scores (MOS) of the test signal quality using a

scale from 0 to 100. A score of 100 means the test signal quality is equal to the quality of the reference signal.

In the listening tests, eleven different broad-band sound sources were used. All the sound sources are multi-channel (5.1 channel) signals with durations of twenty seconds. They are carefully selected to represent a broad range of various kinds of sounds, including classical music, popular music, a movie sound with a monologue, percussive ambience sounds, etc. The contents of the sound sources are described briefly in Table 2.

Material Name	Category
BBC Applause	Pathological & Ambience
ARL Applause	Pathological & Ambience
Chostakovitch	Music (back: direct)
Fountain music	Pathological & Ambience
Glock	Pathological & Ambience
Indie2	Movie sound
Jackson1	Music (back: ambience)
Pops	Music (back: direct)
Poulenc	Music (back: direct)
Rock concert	Music (back: ambience)
Stomp	Movie sound

Table 2: Sound Sources included in the listening test database

The eleven sound sources were encoded and decoded using eleven different multi-channel compression coding systems. Thus, there are 11 X 11 = 121 items in the database.

The effectiveness of the compression is shown in Table III, which gives the bit-rate achieved by the tested multi-channel compression coding systems when their codec indexes were set randomly.

CODEC INDEX	BITRATE	CODEC INDEX	BITRATE
α	182 kb/s	H	97 kb/s
β	177 kb/s	Θ	109 kb/s
γ	177 kb/s	I	172 kb/s
δ	189 kb/s	K	92 kb/s
ε	102 kb/s	Λ	160 kb/s
ζ	97 kb/s		

Table 3: Low bit-rate multi-channel audio compression coding systems that were evaluated

The MOS for each signal was judged by 42 ~ 128 listeners and averaged. The averaged MOS judgments for all signals and coding schemes lie in the range between 42.87 and 89.76. The standard deviations and the numbers of the listeners are used for the calculation of 99% confidential intervals and 95% confidential intervals for the MOS for each signal and coding scheme. Those intervals are used as tolerance values for the analysis of prediction failure rate. The 95% confidential intervals fall in the range between 1.64 and 6.32, with a mean value of 3.93. The 99% confidential intervals have a mean value of 5.17 and lie in the range between 2.16 and 8.32.

2.2 Training of the prediction model

From the 121 items, 61 items were randomly selected and used to train our prediction model. The seven predictive factors – two spatial factors and five timbral factors – were computed for each of the 61 items. These values were used as input elements of a feed-forward neural network whose output was an MOS value. Training of the network set the network weights so that the network output best matched the average MOS judgments of the training items for the appropriate inputs. Our network model is initially developed with a single layer and pure linear activation function (i.e., in this initial implementation, the prediction is based on a regression model that weights the seven input factors).

2.3 Verification of the prediction model

The remaining 60 items not used to train the network weights are used for the verification of the prediction model. The trained network then predicts the MOS of each item from the extracted factors.

Figure 3 shows the relation between the average perceived MOS and the estimated MOS, with the first order regression line. The correlation coefficient between measured and predicted MOS is 0.77.

Estimation error is computed as the difference of the predicted MOS minus the perceptual average MOS. The mean of the absolute values of estimation errors is 5.53, and the standard deviation of absolute errors is 4.33.

A prediction for an item is called a "success" if estimation error is within some tolerance range; otherwise, the prediction is called a failure. Using the 99% confidential interval for the tolerance, the prediction failure rate is 23 / 60 or 38 %. The mean value of the absolute errors for the prediction-failed items is 9.15 with standard deviation equal to 4.69. For the prediction-failed items, Figure 3 shows the 99% confidential intervals (tolerance ranges).

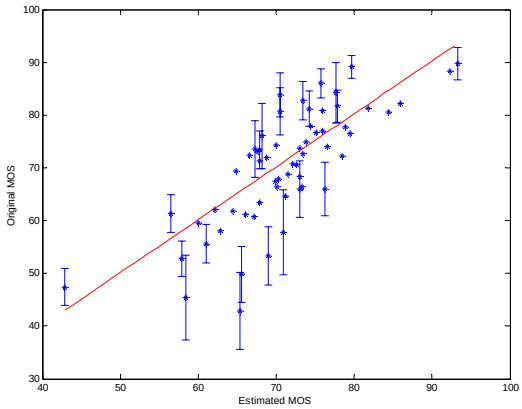

Figure 3: Relation between the average perceived MOS and predicted MOS. The correlation coefficient between perceived and predicted MOS is 0.77. For the prediction-failed items, tolerance ranges (based on 99% confidence intervals) are given.

If we use a more strict tolerance scheme of the 95% confidential interval, the prediction error rate increases to 36 / 60 or 60%. In that case, the mean value of the absolute errors for prediction-failed items is 7.72 and the standard deviation is 4.24.

In Figure 4, estimation errors are shown as a function of the perceived MOS. In general, the errors are positive for low MOS and negative for high MOS. This tendency can inform future approaches for improving the network model.

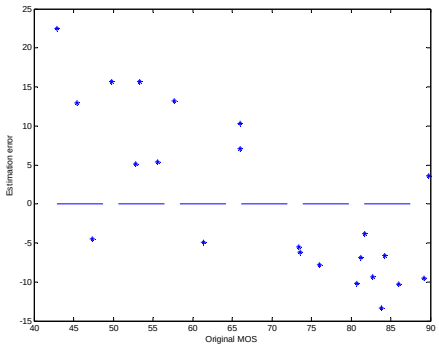

Figure 4: Estimation errors for each "failed" item as a function of the perceived MOS

3 DISCUSSION

Results from our quality prediction model, compared to the prediction performance of different versions of BS.1387-1, are encouraging. The comparison of correlation coefficients is shown in Figure 5. Note that, in this comparison, correlation coefficients of current BS.1387-1 versions are representing their prediction performances in stereo – not multi-channel – databases ("DB-1" and "DB-2"), since the current BS.1387-1 versions cannot behave in multi-channel situations. However, at least, this comparison shows that the proposed model is on a par with the old models in its prediction performance.

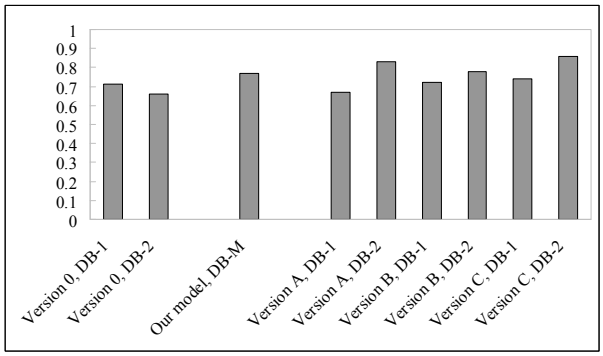

Figure 5: Comparison of correlations for several versions of ITU-R BS.1387-1 with two different stereo databases (represented as "DB-1" and "DB-2"), and our model with the multi-channel database ("DB-M"). "Version 0" is the early version of BS.1387-1, and Version A, B, C are final versions of BS.1387-1. Specific compression versions and databases are not reported, for anonymity.

The final versions of BS.1387-1 produced correlation coefficients between predictions and perceived MOS that ranged from 0.67 to 0.86 for the different databases and different versions. The early version of BS.1387-1 gave correlations of 0.71 and 0.66 for two different databases. Our model predictions give a correlation coefficient of 0.77 with the perceived MOS.

Since our prediction model implements the monaural (timbral) factors used in BS.1387-1, one can view our model as an extension of the BS.1387-1. Figure 6 illustrates this way of envisioning our model.

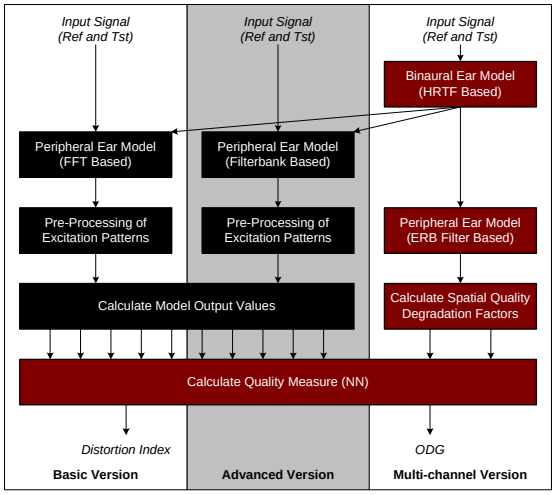

Figure 6: One example approach for extending ITU-R BS.1387-1 to multi-channel use.

Although performance of this initial implementation of our model is encouraging, there is room for improvement. Below, we consider some issues that could be incorporated into future work to try to improve the model's performance.

First, psychophysical knowledge about spatial auditory perception should be included in the model. For example, the minimum audible angle [18] and the minimum audible movement angle [19] vary both with the frequency and location of a source. However, this change in spatial sensitivity is not yet incorporated in the current computations of the ILD and IACC distortions. The weighting factors in equations (3) and (7) for these features are set up to allow the weights to vary with these parameters; however, such a frequency- and location-depending weighting is not yet implemented. In addition, the ITD cue, which is not considered in the current model, should be also included as a feature in future implementations.

Second, the multiple factors in the cognition model need to be verified, to see whether or not they can be treated as independent principal components. Simultaneously, the network function that estimates the MOS also can be improved by using a different activation function (e.g. sigmoid functions) that provides some nonlinearity in the predictions.

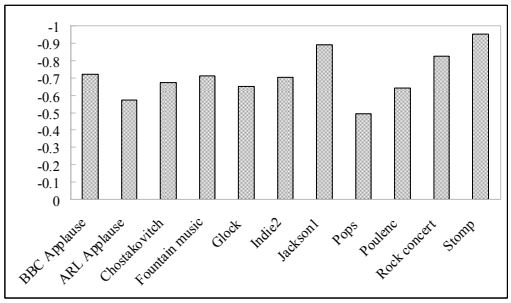

Figure 7: Correlation coefficients between ILD distortion and eleven different kinds of sound sources.

Third, one should consider non-linear effects of the selection of a reference signal on human judgments of sound quality. Using objective assessment methods, we try to evaluate the quality of "devices" such as a compression codec, broadcasting systems, transmission lines, etc. To evaluate the device, a reference signal is passed through the device under test, and the signal at the output of the device is compared to the reference signal. However, these judgments can be affected by the kind of reference signal that is used. Moreover, devices under test generally show different types and amount of quality degradation for different kinds of sound sources. These effects are also found in our experiments. In the database used for training our model, there are eleven different sound sources. From the correlation analysis performed separately for each of different sound sources, the extracted factors (ILD distortion, IACC distortion, and MOVs) have different amounts of influence on the subjective evaluation data (seen as differences in the correlation between the factor of interest and the perceived MOS).

As an example, correlation coefficients of ILD distortion are shown in Figure 7 for different kinds of sound sources. The correlations varied for different sound sources across a range from -0.49 to -0.95. The highest correlation value occurs for the "Stomp" source, which contains various percussion instruments moving around a listener. Perception of sound source location is more sensitive for impulsive sounds like these. Thus, the distortion of interaural cues has a larger effect on perceived sound quality for this kind of signal. In contrast, if a sound has few temporal fluctuations (unlike in the percussive "Stomp" source), location cues are less important for sound quality.

In Figure 8, the waveforms of the binaural signals for "Stomp" and "Pops" are compared. ILD distortion has the lowest correlation with the subjective evaluation results for "Pops."

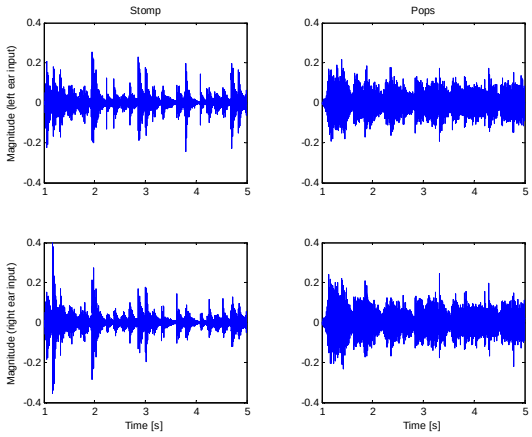

Figure 8: Waveform comparison between binaural signals for "Stomp" and "Pops." The two panels on the left show the waveforms for "Stomp" while the panels on the right show "Pops." The top row shows the left-ear signal and the bottom row shows the right-ear signal. The x-axis represents time in seconds. Magnitude is represented in a relative scale.

The temporal character of the two sources is very different. The "Stomp" source contains many more impulsive sounds with more frequent changes in interaural magnitude ratio than "Pops." In this direct comparison between those two extreme cases, it is easy to envision why interaural cues have a greater impact on a judgment of sound quality for "Stomp," with its impulsive structure, than, for "Pops." The quantification of those relationships between the temporal structure of the reference sound and the importance of interaural cues in sound quality judgments may lead to new

methods for improving the model, to take into account characteristics of the source in determining how to weight spatial features in the prediction of sound quality.

Lastly, the listening test database needs to be enlarged. Recently, ITU-R is collecting new data from listening tests with various multi-channel compression coding systems [20].

4 CONCLUSIONS

In this paper, an objective method is introduced that can be used to predict perceived quality in multi-channel audio compression coding systems. The method takes into account degradations in both spatial quality and timbral quality, extending previous approaches by incorporating a binaural hearing model from which interaural features are computed. After training our model with the listening test database that includes perceptual evaluation of various low bit-rate multi-channel audio-coding systems, our model gives encouraging results. In particular, predictions of perceived quality are comparable to results from other evaluation models. Still, there is room for improvement in our model's performance. Future efforts will incorporate knowledge from psychophysical research, particularly from spatial and binaural experiments, as mentioned in the discussion.

ACKNOWLEDGEMENT

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, KRF-2006-612-D00068).

REFERENCES

[1] ITU-R Recommendation BS.1387-1, "Method for Objective Measurement of Perceived Audio Quality," International Telecommunication Union, Geneva, Swiss, 1999.

[2] Frank Baumgarte and Christof Faller, "Binaural Cue coding. Part I: Psychoacoustic fundamentals and design principles," IEEE transactions on speech and audio processing, vol.11, no.6, pp.509-519, 2003.

[3] ISO/IEC JTC1/SC29/WG11 (MPEG) Document N6691, "Tutorial on MPEG Surround Audio Coding," Poznan, Poland, July 2005.

[4] Søren Bech and Nick Zacharov, "Perceptual Audio Evaluation - Theory, Method and Application," John Wiley & Sons, Chichester, 2006.

[5] S. Torres-Guijarro, J. A. Beracoechea-Alava, F. J. Casajus-Quiros, and I. Perez-Garcia, "Coding Strategies and quality measure for multichannel audio," Audio Eng. Soc. 116th Convention, Berlin, Germany, 2004.

[6] Sunish George, Slawomir Zielinski, and Francis Rumsey, "Initial developments of an objective method for the prediction of basic audio quality for surround audio recordings," Audio Eng. Soc. 120th Convention, Paris, France, 2006

[7] ISO/IEC JTC1/SC29/WG11 (MPEG) Document M12265, "Objective Measurement of Total Auditory Quality of Spatial Audio Coding," Poznan, Poland, July 2005

[8] In Yong Choi, Sang Bae Chon, and Koeng-Mo Sung, "Measuring spatial attributes of multi-channel audio coding systems," 9th Western Pacific Acoustics Conference (WESPAC), Seoul, Korea, 2006

[9] ISO/IEC JTC1/SC29/WG11 (MPEG) Document N6813, "Report on Spatial Audio Coding RM0 Selection Tests," Palma de Mallorca, Oct. 2004.

[10] ISO/IEC JTC1/SC29/WG11 (MPEG) Document N7138, "Report on MPEG Spatial Audio Coding RM0 Listening Tests," Busan, April 2005.

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

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

[13] Jan Berg and Francis Rumsey, "Systematic Evaluation of Perceived Spatial Quality," Audio Eng. Soc. 24th International Conference on Multichannel Audio, Banff, Canada, June 2003

[14] Tomasz Letowski, "Sound Quality Assessment: Concepts and Criteria," Audio Eng. Soc. 87th Convention, New York, Oct. 1989.

[15] Jens Blauert, "Spatial Hearing: The Psychophysics of Human Sound Localization," MIT Press, Boston, 1983.

[16] Barbara G. Shinn-Cunningham, "Learning Reverberation: Considerations for Spatial Auditory Displays," International Conference of Auditory Display, April 2000.

[17] ITU-R Recommendation BS. 1534-1, "Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)", International Telecommunication Union, Geneva, Swiss, 2001.

[18] A.W. Mills, "On the minimum audible angle," J. Acoust. Soc. Am., 30, pp.237-246, 1958.

[19] David R. Perrott, and Juliana Tucker, "Minimum audible movement angle as a function of signal frequency and the velocity of the source," J. Acoust. Soc. Am., Vol. 83, Issue 4, pp.1522-1527, April 1988

[20] ISO/IEC JTC1/SC29/WG11 (MPEG) Document M12151, "Liaison Statement from ITU-R TG 6/9 to ISO/IEC MPEG, SMPTE, and EBU," Poznan, July 2005.