OBJECTIVE MEASUREMENT OF PERCEIVED AUDITORY QUALITY IN MULTI-CHANNEL AUDIO COMPRESSION CODING SYSTEMS

INYONG CHOI[1], BARBARA G. SHINN-CUNNINGHAM[2], SANG BAE CHON[1], KOENG-MO SUNG[1]

[1] *Institute of New Media and Communications, School of Electrical Engineering and Computer Science, 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 have been used widely for evaluation of audio coding systems. However, even though many different competing 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 multi-channel audio compression coding systems. The method introduces three variables as measures of the degradations in spatial quality: interaural time difference distortion (ITD distortion), interaural level difference distortion (ILD distortion), and interaural cross-correlation coefficient distortion (IACC distortion). Simultaneously, ten Model Output Variables proposed in ITU-R recommendation BS.1387-1 are extracted from binaural signals that are synthesized using head related 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 the MPEG audio group. This new model, using the three interaural and ten non-spatial statistics, shows encouraging results in the prediction of 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. 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, such as those described in [1-7], have been proposed. Thanks to the efforts of the participants in International Telecommunication Union Radiocommunication Sector (ITU-R) to combine those methods and develop a single best method, ITU-R Recommendation BS.1387-1 [8] has been established and widely used. However, because its scope is restricted to evaluating degradations caused by known coding artifacts, the method described in [8] cannot predict perceived quality of newly developed audio coding technologies that import novel coding schemes to accomplish extremely high efficiency in compression with intermediate sound quality [9]. Moreover, the recommendation cannot be used to objectively assess multi-channel audio coding systems because it was designed only for monaural and stereo sounds [9].

Two recent models for the objective assessment of quality of multi-channel sound sources have been proposed [10, 11]. However, to date, satisfactory predictions of perceptual quality of newly developed low bit-rate multi-channel coding systems have not been reported. In this paper, a prediction model is introduced that can be used for the objective quality assessment of multi-channel audio compression coding systems, focusing on recently introduced, efficient-compression coding technologies.

An adequate predictive model for the perceived quality of multi-channel sound 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 our method, multi-channel signals are first converted into binaural signals using head related transfer functions (HRTFs) to simulate the signals that a listener would hear using a full, standard layout, multi-channel audio reproduction system. The binaural signals are processed to extract statistics thought to be important for spatial perception based on psychoacoustic principles.

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 and changes in perceived source width are taken into account in the current model.

For the localization of a sound source, it is generally accepted that the most robust and important spatial auditory cues are differences between left and right ear signals [12]. In the field of audio engineering, there have been successful applications based on those interaural features such as the binaural cue coding [13] systems, one of which is the standard multi-channel audio compression coding system also known as "MPEG Surround [14]." There are two such interaural differences that are important perceptually: interaural time differences (ITDs) and interaural level differences (ILDs) [12, 15]. Both ITDs and ILDs are important in localization, although they have different and complementary roles [12]. Thus, in our model, estimations for perceived changes in both ITDs and ILDs are implemented in different methods with different target frequencies.

The long-term IACC is another important spatial attribute, related to perceived source width, shose distortion is likely to be important in judging sound reproduction quality. The long-term IACC is influenced by the direction of incoming direct and reverberant sounds over a relatively long period of time, and thus will require computations that integrate over relatively long time frames.

To summarize, the statistics extracted for the evaluation of spatial quality are interaural time difference distortion (ITD distortion), interaural level difference distortion (ILD distortion [16, 17, 18]), and interaural cross-correlation coefficient distortion (IACC distortion [17, 18]).

Simultaneously, ten Model Output Variables (MOVs) in ITU-R BS.1387-1 are computed from the binaural signals for assessment of timbral quality. The prediction model based on these ten timbral features and three spatial features 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 [19, 20].

In Section 1, the implementation of the prediction model is illustrated. The prediction model is logically divided into three sequential parts: a binaural signal synthesis, 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. The proposed prediction model shows encouraging performance, with prediction-data correlation coefficients as large as 0.85. However, results suggest that prediction performance can be improved further by extending both coding analysis as well as signal analysis. Section 3 outlines a number of possible improvements and extensions, including the match between desired and realized high-frequency ITD cues for sounds with fluctuating high-frequency envelopes. 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 [21]. 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 [22]. 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.

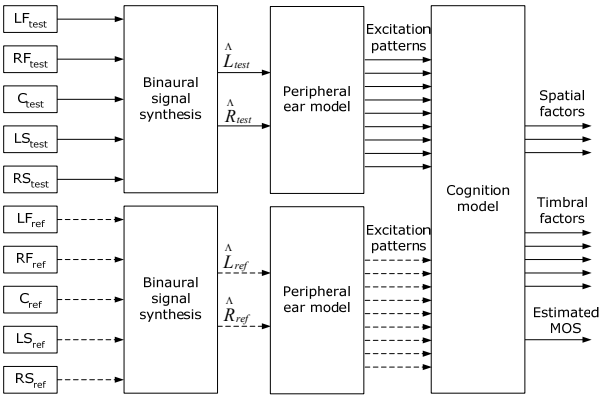


Figure 1: Overall structure of our prediction model

The overall process consists of a binaural signal simulator, a peripheral ear model, and a cognition model. The binaural signal simulator 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 [1]. 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 parts are described in following three sub-sections, respectively.

1.2 Simulation of binaural signals

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, given that 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.

In typical multi-channel audio cases, both the reference and test signals consist of five signals for the five channels in the reproduction system. In our binaural signal simulator, 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 head related impulse responses (HRIRs) 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).

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

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

The HRTFs are recorded in a carefully designed listening chamber that satisfies the conditions of a reference listening room as recommended in the ITU-R BS.1116 [23], using high-quality sound reproduction systems that satisfy the conditions of the reference monitor loudspeakers, described in [23]. The directions of the ten HRTFs correspond to those of loudspeakers in a standard layout of multi-channel audio reproduction systems. The geometric configuration of the standard layout 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-surround channel and the right-surround channel loudspeakers at -110 degrees and +110 degrees, respectively [23].

Whereas the listeners are free to move their head directions and positions during listening tests, the binaural signal simulator using a single set of HRTFs and assumes that the listener is at a fixed position with his/her head oriented forward It may be possible to extend the evaluation by simulating multiple listener positions and extracting multiple sets of quality measures from those various binaural signals. Such an approach will result in a large increase in the computation complexity of the model, and so is not yet included. However, recent studies have measured listeners' head movement while listening to spatial sound sources [24], so it is possible to explore how head movements could be incorporated in the future.

Timbral features – the MOVs – are also calculated from the simulated binaural signals. 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 was 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 as high as 0.68. This result demonstrates the importance of modeling the listening environment when evaluating spatial sound reproduction systems.

1.3 Peripheral ear model

Synthesized binaural signals are processed by a peripheral ear model. The peripheral ear model converts ear input signals into a representation like the signals exciting hair cells in the human basilar membrane, which translate mechanical vibrations from acoustic inputs into neural firing in the auditory nerve fibers. In our proposed method, the peripheral ear model includes computation units identical to those of BS.1387-1, since the MOVs are used as timbral features. However, for

the coding of excitation patterns for the computation of spatial – interaural – features, a different filter-bank structure is used to obtain different temporal and frequency resolutions against those of the peripheral ear model for the MOVs.

1.3.1 Peripheral ear model for timbral features (MOVs)

ITU-R BS.1387-1 introduces two peripheral ear models, "FFT-based ear model" and "filter bank-based ear model," that fully take account of generally accepted concepts of psychoacoustics. The outputs of those two ear models are used to calculate different sets of MOVs. In our experiment, the MOVs obtained from the outputs of an FFT-based ear model ("the basic MOVs") gave slightly better predictions of perceived quality than the MOVs from a filter-bank-based ear model ("the advanced MOVs"). Thus, the set of the basic MOVs was selected as timbral features. In other words, the peripheral ear model in the proposed method for the timbral features is identical to the FFT-based model of BS.1387-1.

In the FFT-based ear model of ITU-R BS.1387-1, the peripheral ear outputs includes stages of 1) computation of the discrete time Fourier transform (DFT), 2) level scaling, 3) frequency weighting to simulate outer and middle ear transfer characteristics, 4) grouping into critical bands, 5) adding internal noise, and then 6) simulation of temporal and simultaneous masking in both the time and frequency domains. The DFT is computed using time frames of 21ms duration with 50% overlap (see [8, 25] for details).

1.3.2 Peripheral ear model for spatial features

For an appropriate computation of interaural features, temporal resolution must be increased over that of an FFT-based ear model. Moreover, because the perceptual sensitivity to temporal fluctuations in ITD, ILD, and IACC cues differs, these variables should be computed with different temporal resolution. Thus, an additional filter-bank structure is needed for the computation of interaural cues.

The 4th-order Patterson-Holdsworth filter bank, also known as the Gammatone filter bank [26], is selected as the additional filter bank. Although a better temporal resolution also can be obtained by the modification of the DFT size or by the use of "filter bank-based ear model" in [8], the Gammatone filter bank brings some benefits, including low complexity, since it can be implemented by infinite impulse response filters, and relatively accurate simulation of the cochlea filter outputs.

The filter bank has 24 bands with the center frequencies determined by the ERB scale [27, 28]. Binaural signals are converted into band-pass filtered signals by the filter banks, weighted by the outer and middle ear weighting factors, and divided into time segments.

The peripheral ear outputs toward inputs to the ITD cognition model take 7/8 overlapping 20-ms rectangular windows. For the ILD distortion computation, the band-pass filtered signals are segmented by 3/4 overlapping 10-ms rectangular windows, while the signals are divided by 3/4 overlapping 50-ms rectangular windows for the IACC computation.

1.4 Cognition model

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

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.

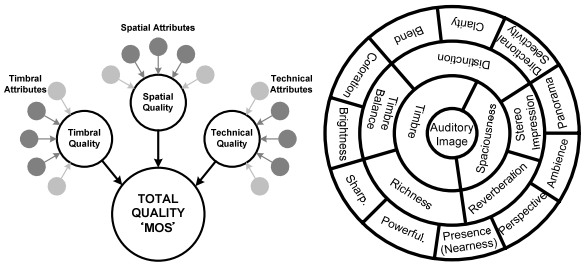


Figure 2: Conceptual illustration of Total Auditory Quality and MUlti-level auditoRy Assessment Language (MURAL)

Figure 2 illustrates the attributes of sound quality used in the MUlti-level auditoRy Assessment Language (MURAL) [29] model. Attributes are divided into two groups, affecting either 'timbre' or 'spatial impression.' More recently, Berg and Rumsey [30] 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.

As mentioned in the introduction session, among the many different perceptual attributes related to the spatial quality, errors in perceived location based on ITD and ILD distortions and changes in perceived source width based on IACC distortions are taken into account in the current model.

Although both ITDs and ILDs are important localization cues, they are processed in different brain

nuclei and play different and complementary roles in spatial perception [12, 31]. ITDs are computed by coincidence-detecting circuits with various interaural time delays in the medial superior olive (MSO) [32, 33]. ILDs are extracted in a neighbouring region, the lateral superior olive (LSO). ITDs provide strong spatial cues for low frequency sounds (below 1500Hz), while ILDs are most important in high frequency sounds (above 2500Hz). Although ITDs in the envelopes of high-frequency sounds are perceptually salient [34, 35], the importance of these cues depends on characteristics of the stimulus [36] and is not as important in perception as the low-frequency ITDs. Thus, in this initial development of the cognition model, ITDs and ILDs are implemented in different methods with different target frequencies, i.e. low frequency spatial distortions are based on measurement of ITD changes, while high frequency spatial distortions use ILD changes. Section 3 discusses how high-frequency envelope ITDs might be incorporated into future models.

The computational models for measuring these psychoacoustical factors are described in the next sub-sections.

1.4.1 Computation of errors in low frequency sound directions based on ITD distortion

ITD, derived by coincidence detection neurons in the brain (MSO), is an important cue for sound source localization, especially for a low-frequency sound. Thus, differences in the low-frequency ITD between the test and reference signal is likely to be important in predicting sound quality. However, because of the non-linear nature of human neural systems, computation of a perceptual distance between two ITDs requires multiple stages of computation.

First, ITD can be computed from the following time-window-based normalized cross-correlation function (*NCF*), where $X_{L,k,n}$ and $X_{R,k,n}$ are peripheral ear model outputs of the left ear and the right ear, respectively. d is the time lag represented in samples. k and n are the frequency band and time frame indices. The cross-correlation is calculated over 7/8 overlapping rectangular time windows with the length approximately equivalent to 20ms.

$$NCF_{k,n}[d] = \frac{\sum_l X_{L,k,n}[l] X_{R,k,n}[l+d]}{\sqrt{\sum_l X_{L,k,n}^2[l] X_{R,k,n}^2[l]}} \tag{2}$$

The interaural cross-correlation coefficient (IACC) is defined as the maximum value of the *NCF* over all d, and the ITD is the value of d giving this maximum. These values are denoted as $IACC[k,n]$ and $ITD[k,n]$ in (3) and (4), to indicate their frequency and time indexes:

$$IACC[k,n] = \max \left| NCF_{k,n}[d] \right|_{d=-N}^{d=+N} \tag{3}$$

$$ITD[k,n] = \arg\max_d \left| NCF_{k,n}[d] \right|_{d=-N}^{d=+N} \tag{4}$$

Parameter N is the range of d, covering all theoretically possible ITD values, represented in sample numbers. ITD is measured in both the test and reference signals, and is denoted as $ITD_{test}[k,n]$ and $ITD_{ref}[k,n]$ in the next computation stage.

Second, inspired by a computational model for predicting source direction based on ITD (in which the interaural phase difference is the phase and IACC is the amplitude of a vector represented in polar coordinates [37]), the perceptual change of the source direction can be appropriately calculated as the Euclidian distance between two positions on a unit circle. Since the distance between two different azimuth angles (θ_1 and θ_2, with the same radius of 1) can be calculated as (5), the perceptual distance between two source directions due to the ITD difference can be modeled as (6).

$$\sqrt{(\cos\theta_1 - \cos\theta_2)^2 + (\sin\theta_1 - \sin\theta_2)^2} = \sqrt{2 - 2\cos(\theta_1 - \theta_2)} \quad \text{.....(5)}$$

$$\Delta ITD[k,n] = \sqrt{2 - 2\cos\pi \, {f_s}/{N} \, (ITD_{test}[k,n] - ITD_{ref}[k,n])} \quad \text{.....(6)}$$

In this formulation, parameter f_s is the sampling rate and N is the maximum ITD represented in sample numbers. Thus, f_s/N can be regarded as a normalizing factor that restricts the input of the cosine function of (6) to be in the range from 0 to π.

Lastly, it must be considered that ITD detection probably fails in some cases. Perceived source direction can be ambiguous if the IACC is too low to produce reliable percepts of source direction. Thus, we need to apply a decision factor that takes into account the certainty of the calculated ITD. In our computation method, this certainty is modelled by a nonlinear transformation of the IACCs through an approximate tangential sigmoid function, as in (7) and (8). Parameters S and T_k are constants for steepness and threshold. The tangential sigmoid function is shown in figure 3 for the slope parameter (S) set to 50 and the threshold (T_k) set to 0.5. Note that, in the model, T_k takes on different values in different frequency bands, in order to account for different sensitivity to ITD in different frequency bands.

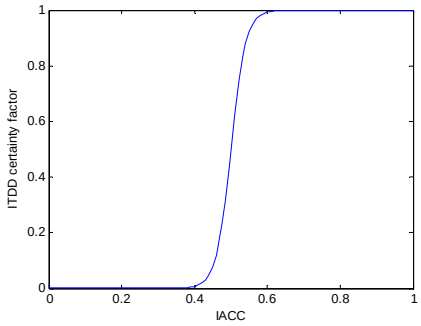


Figure 3: ITD certainty factor model, with the steepness parameter of 50 and and threshold of 0.5. X-axis represents the absolute value of IACC.

$$p_{test}[k,n] = \{1 + e^{-S(|IACC_{test}[k,n]| - T_k)}\}^{-1} \quad (7)$$

$$p_{ref}[k,n] = \{1 + e^{-S(|IACC_{ref}[k,n]| - T_k)}\}^{-1} \quad (8)$$

After applying those certainty factors to the $\Delta ITD[k,n]$, the ITD distortion is given by (9).

$$ITDDist[k,n] = \frac{1}{2}(p_{test}[k,n] + p_{ref}[k,n]) \cdot \Delta ITD[k,n] \quad .....(9)$$

The resulting ITD distortions are averaged over frequency bands and time frames, as in (10) and (11). *ITDDist* is used to represent the average ITD distortion, which measures perceptual distance between the test and reference source directions due to differences in their ITDs.

$$ITDDist[n] = \frac{1}{Z}\sum_{k=0}^{Z-1} w_1[k] \cdot ITDDist[k,n] \quad (10)$$

$$ITDDist = \frac{1}{N}\sum_{n=1}^{N} ITDDist[n] \quad (11)$$

$w_1[k]$ is a non-linear weighting factor, which takes into account the relative importance of the ITD distortion in each frequency band for perceived location.

1.4.2 Computation of perceived degradation in high frequency sound direction

ILD, or interaural level difference between the left and right ear inputs, is an important cue for perception of sound direction of high-frequency sounds. Thus, the computation of ILD difference between the test and reference signals reflects degradations in the perceived source direction for a high-frequency sound.

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} frequency band in the n^{th} time frame. The intensity levels are calculated in 3/4 overlapping rectangular time windows that have the sizes approximately equivalent to 10ms (half of the length of the window used in ITD computation).

$$ILD[k,n] = 10\log_{10}\left(\frac{\sum_l X_{L,k,n}^2[l]}{\sum_l X_{R,k,n}^2[l]}\right) \quad (12)$$

ILD is extracted in the LSO of the brain through a different physiological mechanism from that extracting ITD. Thus, the computation method for the distortion in inter-aural level differences differs from that of ITD distortion. First, the perceptual distance between two source locations due to the ILD difference is estimated by linear subtraction of the ILD in dB. Second, the sound localization judgment based on ILD is weighted by the intensity of the signal in a given time-frequency segment. Thus, the ILD distortion is modelled as:

$$ILDDist[k,n] = w_2[k] \cdot \log_{10}\left(\sum_l X_{L,k,n}^2[l]\right) \cdot \left|ILD_{test}[k,n] - ILD_{ref}[k,n]\right| \quad .....(13)$$

where $ILD_{test}[k,n]$ and $ILD_{ref}[k,n]$ are the ILDs of the test and reference signals, respectively. $w_2[k]$ is a non-linear weighting factor, which mirrors the relative importance of the ILD distortion in each frequency band.

By averaging over frequency bands and time frames (as in (14) and (15)), we get *ILDDist*, which is a measure of perceptual distance between the ILD-based source direction of the test and reference signals.

$$ILDDist[n] = \frac{1}{Z}\sum_{k=0}^{Z-1} ILDDist[k,n] \quad (14)$$

$$ILDDist = \frac{1}{N}\sum_{n=1}^{N} ILDDist[n] \quad (15)$$

1.4.3 Computation of perceived degradation in apparent source width

IACCDist is used as a measure of degradations in the apparent source width, and it is calculated as:

$$IACCDist[k,n] = w_3[k] \cdot \left|IACC_{test}[k,n] - IACC_{ref}[k,n]\right| \quad .....(16)$$

$$IACCDist[n] = \frac{1}{Z}\sum_{k=0}^{Z-1} IACCDist[k,n] \quad (17)$$

$$IACCDist = \frac{1}{N}\sum_{n=1}^{N} IACCDist[n] \quad (18)$$

Note that the IACC for this *IACCDist* computation is calculated the same way as in (3), but the cross-

correlation here uses a longer time window (approximately 50ms).

1.4.4 Calculation of factors for timbral quality

Model Output Variables (MOVs) from ITU-R BS.1387-1 are used to quantify spectral degradations in our model. There are two versions in BS.1387-1: the Advanced Version with six MOVs and the Basic Version using eleven MOVs. In the current method, the ten MOVs from the Basic Version were used, except that one Basic Version MOV, which produces saturated values for the multi-channel sound sources used here, was discarded; the MOVs used are briefly described in Table 2 (see [8, 25] for details, e.g. equations and numerical data).

1.4.5 A neural network for multi-dimensional perception of sound: Estimation of MOS

Estimation of MOS, the measure of the one dimensional attribute "BAQ," is performed by two network models: a linear estimator and an artificial neural network model with multiple inputs and one output. Through the computations explained in above sub-sections, a total of thirteen attributes – three interaural features and ten MOVs [8, 25] – were extracted and used as the inputs to the network models.

The three spatial features and ten MOVs are summarized in table 1 and table 2, respectively.

Features	Description
ITDDist	Perceptual distance between source directions of the signal under test and the original signal due to the ITD difference. Computed for low frequency sounds (below 1500Hz).
ILDDist	Perceptual distance between source directions of the signal under test and the original signal due to the ILD difference. Computed for high frequency sounds (above 2500Hz).
IACCDist	Perceptual difference between apparent source widths of the signal under test and the original signal due to the IACC difference.

Table 1: BAQ estimator inputs I - Interaural features for measuring degradations in spatial quality

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
BWRef	Bandwidth of the reference signal
BWTest	Bandwidth of the signal under test
AModDif1B	Averaged modulation difference
AModDif2B	Averaged modulation difference with emphasis on the modulation changes where the reference contains little modulations.
WinModDifB	Windowed averaged modulation difference
RDF	Relative fraction of frames with significant noise component
NLoudB	Averaged noise loudness

Table 2: BAQ estimator inputs II - MOVs of ITU-R BS.1387-1 that were used as factors for timbral degradations

The artificial neural network model is developed as a two-layer feed-forward network and trained using the "backwards propagation of errors (backpropagation)" method. The first layer has five nodes with the tangent sigmoid transfer function, and the second layer has one linear node. The procedures and results of network training are shown in the next section.

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 [19, 20] 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) [22]." 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 test signals were used. All the test signals are multi-channel (5.1 channel) signals with durations of twenty seconds, selected to represent a broad range of various kinds of sounds (e.g., classical music, popular music, a movie sound with a monologue, percussive ambience sounds, etc.). The contents of the test signals are described briefly in Table 3.

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 3: Test excerpts included in the listening test database

The eleven test excerpts 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 4, 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	Η	97 kb/s
β	177 kb/s	Θ	109 kb/s
γ	177 kb/s	Ι	172 kb/s
δ	189 kb/s	Κ	92 kb/s
ε	102 kb/s	Λ	160 kb/s
ζ	97 kb/s		

Table 4: 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. Confidence intervals are used as tolerance values for the analysis of prediction failure rate. The 99% confidential intervals fall in the range between 2.16 and 8.32, with a mean value of 5.17.

2.2 Training of the prediction model

From the 121 items, 61 items were randomly selected and used to train the prediction model. The thirteen predictive factors – three spatial factors and ten timbral factors – were computed for each of the 61 items. These values were used as the input elements of a linear estimator and 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.

2.3 Verification of the prediction model

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

Three types of network functions were developed: a linear estimator with only ten MOV inputs ("10-input LE"), a linear estimator with thirteen input parameters including ten MOVs and three interaural features ("13-input LE"), and a two-layer feed-forward neural network ("13-input NN"). The linear estimation network models were trained to attain the least square error, and the artificial neural network model was trained by the backpropagation method. The best model can be derived by comparing the three different approaches. First, comparison between 10-input LE and 13-input LE establishes the value of the newly proposed interaural features. Second, comparison between 13-input LE and 13-input NN shows the improvement obtained by using a neural network instead of a linear estimator.

Figures 4 (a), (b), (c) show the relations between the average perceived MOS and the estimated MOS that are predicted by the above three methods, with the first order regression lines. Table 5 shows a comparison of several diagnostic attributes that evaluate the performance of the three models. The correlations coefficients are largest using the non-linear network model with the new interaural features; the correlation coefficients between measured and predicted MOS are 0.71, 0.79, and 0.85 for 10-input LE, 13-input LE, and 13-input NN, respectively.

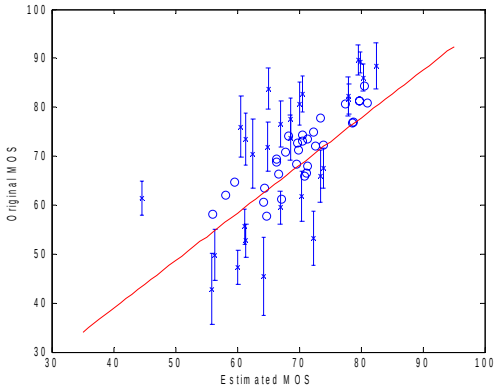
Estimation error is computed as the difference between the predicted MOS and the perceptual average MOS. The mean of the absolute values of estimation errors was 6.18 for the 10-input LE. This error was reduced to 5.44 when the three spatial features were included as inputs, and to 5.09 when the non-linear neural network was used instead of the LE.

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% confidence intervals as the tolerance, the prediction failure rate was 28 / 60 or 47 % when the 10-input LE was used. The failure rates for the LE and the non-linear neural network using spatial features were lower than for the 10-input LE (21 / 60 = 35% and 18 / 60 = 30%, respectively). The mean values of the absolute errors for the prediction-failed items are 9.94, 9.79, and 9.02, with standard deviations equal to 4.50, 4.41, and 4.19. For the prediction-failed items, Figure 4 shows the 99% confidence intervals (tolerance ranges). In summary, correlation is greatest and the mean number of errors

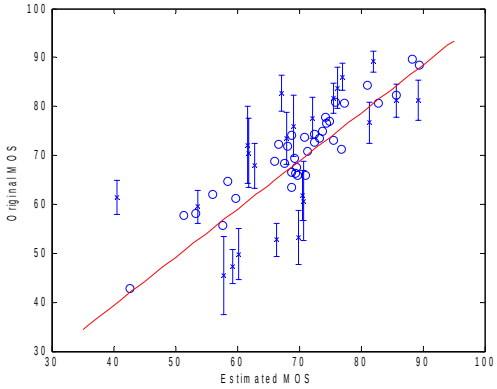
smaller when the three, newly proposed interaural features are used as classifier inputs. Use of a non-linear neural network also improves the performance of MOS prediction over a LE using the same input parameters.

	10-input LE	13-input LE	13-input NN
Correlation Coeffcient	0.71	0.79	0.85
Mean of Estimation Errors	6.18	5.44	5.09
Standard Deviation of Estimation Errors	4.84	4.39	4.04
Prediction Failure Rate	28	21	18
Mean of Estimation Errors in failed items	9.94	9.79	9.02
Std. dev. of Est. Errors in failed items	4.50	4.41	4.19

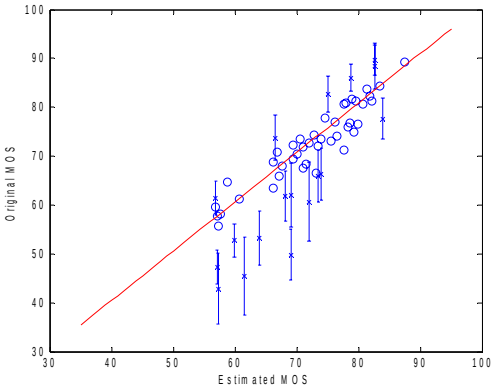
Table 5: Comparison of various criteria for prediction performance in three types of the MOS estimators: ten input linear estimator, thirteen input linear estimator, and thirteen input neural network.



(a) Estimated MOS: predicted by a linear estimator with 10 inputs.



(b) Estimated MOS: predicted by a linear estimator with 13 inputs.



(c) Estimated MOS: predicted by a neural network with 13 inputs.

Figure 4: Relation between the average perceived MOS and predicted MOS with three types of estimators. The correlation coefficient between perceived and predicted MOS is 0.71 (a), 0.79 (b), and 0.85 (c). For the prediction-failed items, tolerance ranges (based on 99% confidence intervals) are given.

3 DISCUSSION

The proposed interaural difference distortion variables are highly correlated with listening test results. Correlation coefficients for the thirteen attributes, ten MOVs and three spatial features, are compared in Figure 5. The correlation coefficients are calculated using binaural signals from all 121 items in the listening test database. In the correlation analysis, the computed measures of ITD, ILD and IACC distortions and the ten selected MOVs from the Basic Version of BS.1387-1 yield consistently high correlation coefficients. ILD distortion, IACC distortion, and ITD distortion have correlation coefficients of -0.78, -0.62, and -0.61, respectively.

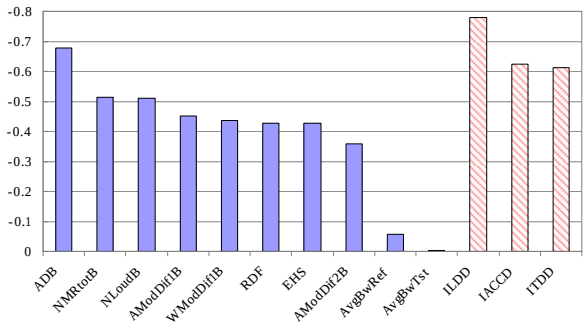


Figure 5: Comparison of correlations for ten MOVs of ITU-R BS.1387-1 and proposed spatial features with the listening test results. Because these values are negatively correlated with the spatial feature measures, the y-axis shows the negative correlation coefficient, so that better predictions yield taller bars in the plot. Correlation coefficients are calculated using binaural signals from all 121 items in the listening-test database.

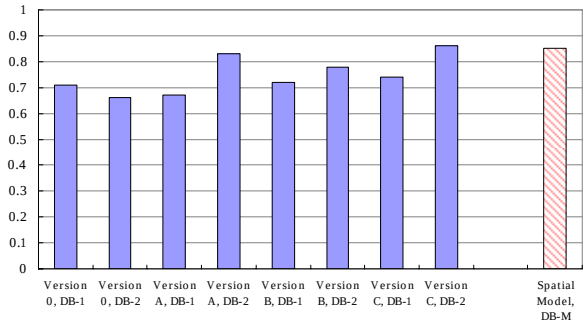


Figure 6: Comparison of correlations for several versions of ITU-R BS.1387-1 with two different stereo databases (represented as "DB-1" and "DB-2"), and our model with the multi-channel database ("DB-M"). "Version 0" is the early version of BS.1387-1, and Version A, B, C are final versions of BS.1387-1. Specific compression versions and databases are not reported, to preserve anonymity.

Results from our MOS prediction model, compared to the prediction performance of different versions of BS.1387-1, are also encouraging. The comparison of correlation coefficients is shown in Figure 6. Note that, in this comparison, correlation coefficients of current BS.1387-1 versions represent their prediction performance for stereo – not multi-channel – databases ("DB-1" and "DB-2"), since the current BS.1387-1 versions cannot be used to evaluate multi-channel conditions. Nonetheless, this comparison shows that the proposed model is on a par with the old models in its prediction performance.

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 (reported in [8]). The early version of BS.1387-1 gave correlations of 0.71 and 0.66 for two different databases. Our model predictions have a correlation of 0.85 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 7 illustrates this way of envisioning our model.

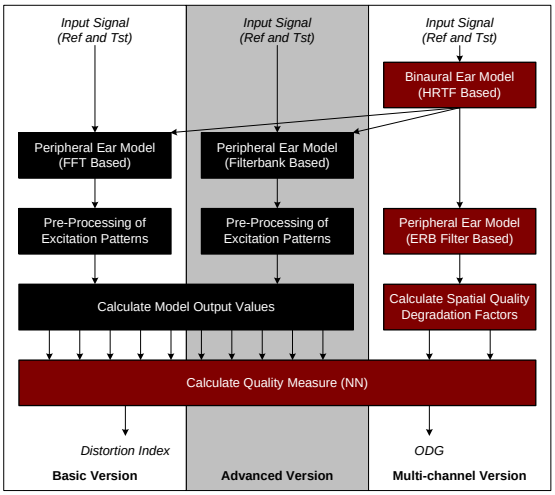


Figure 7: One example approach for extending ITU-R BS.1387-1 to multi-channel use.

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, the multiple factors used in the cognition model need to be verified, to see whether or not they can be treated as independent principal components. Furthermore, the network function that estimates the MOS also may be improved by using a different structure.

Second, high-frequency envelope ITDs should be considered in the next development. Although the "duplex theory [15, 31]" of localization has great explanatory power, it was developed to describe results for sinusoidal stimuli. It is well known that high-frequency envelope ITDs are perceptually important [34, 35], although their salience depends strongly on the temporal characteristics of the input stimuli [36]. If it is possible to incorporated knowledge about the detectability of envelope ITDs for different high-frequency stimuli, a high-frequency ITD measure would likely provide new information about spatial sound quality. In our present ITD-cognition model, the probability factor that models sensitivity to changes of ITDs is implemented as a function of the magnitude of the interaural cross-correlation coefficient. A sensitivity factor similar to this may be important when developing a high-frequency, envelope ITD metric, with different

input parameters such as the steepness of the envelope curve, etc.

Now we can naturally move on to the third topic of future works: signal analysis. One should consider (non-linear) effects in 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, since devices under test generally show different types and degree of quality degradation for different kinds of test signals.

These effects are also found in our experiments. In the database used for training our model, there are eleven different test signals. From the correlation analysis performed separately for each of different test signals, the extracted factors (ITD distortion, 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 ITD, ILD, and IACC distortions are shown in Figure 8 for different kinds of test excerpts. The correlations varied for different test excerpts across a range from -0.48 to -0.96. The highest correlation occurs for test signals like "Glock" and "Stomp," which contain percussion instruments moving around a listener. Sensitivity to spatial cues is higher for impulsive sounds like these than for more continuous sounds. 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, location cues are less important for sound quality.

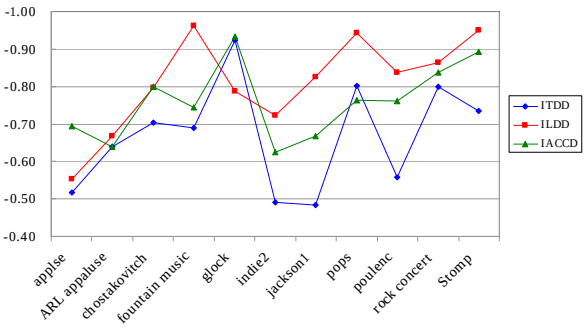


Figure 8: Correlation coefficients between ILD distortion and eleven different kinds of test excerpts.

In Figure 9, the waveforms of the binaural signals for "Indie 2" and "Stomp," which are examples of a non-percussive and an extremely percussive stimulus, respectively, are compared. Distortions of interaural differences yield low correlations with the subjective evaluation results for "Indie 2," but high correlation with "Stomp," which contains many transients.

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 "Indie 2." Moreover, the "Stomp" is created by an advanced recording technology that uses multi-channel microphones equipped in the space of a sound event, whereas the "Indie 2" is an audio clip of a movie soundtrack created by a conventional method in a studio. 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 and realism of spatial information, than for "Indie 2." Finding a way to quantify these differences in signal quality is likely to lead to new methods for improving the model by taking into account characteristics of the source in determining how to weight spatial features in the prediction of sound quality.

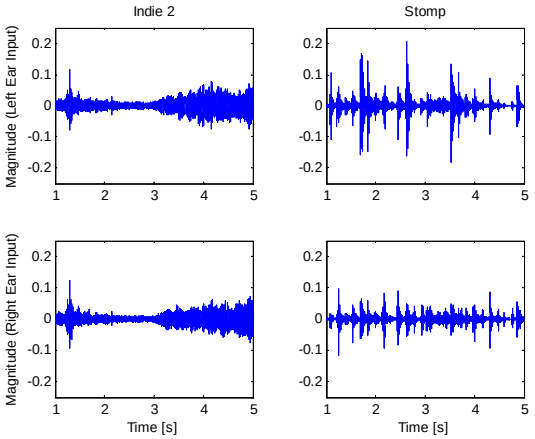


Figure 9: Waveform comparison between binaural signals for "Indie 2" and "Stomp." The two panels on the left show the waveforms for "Indie 2" while the panels on the right show "Stomp." The top row shows the left-ear signal and the bottom row shows the right-ear signal. The x-axis represents time in seconds. Magnitude is represented in a relative scale.

Signal analysis could also improve how the other, timbral variables are weighted in the model. The current system only considers a small set of known coding artifacts (changes of modulation patterns, frequency-band limitation, adding of noise, etc.) and it uses a fixed network function to extract a predicted MOS from the multiple variables regardless of the type of signal used. A method that quantifies different signal attributes and adjusts the weighting and scaling of the different factors in the prediction, based on the type of input signal, is likely to produce better results.

Lastly, the listening test database needs to be enlarged. ITU-R is currently collecting new data from listening tests with various multi-channel compression coding systems [38], which will provide additional tests of the current algorithm and other new approaches.

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 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 or better than results from other evaluation models.

5 ACKNOWLEDGEMENT

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, KRF-2006-612-D00068). Authors are grateful to Virginia Best for reading and discussions.

REFERENCES

[1] 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, 1992.

[2] J. Herre, E. Eberlein, H. Schott, and C. Schmidmer, “Analysis Tool for Real Time Measurements Using Perceptual Criteria,” Audio Eng. Soc. 11th Conference, Portland, USA, 1992.

[3] J. G. Beerends and J. A. Stemerdink, “A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation,” J. Audio Eng. Soc., Vol. 40, pp. 963-978, 1992.

[4] C. Colomes, M. Lever, J. B. Rault, and Y. F. Dehery, “A Perceptual Model Applied to Audio Bitrate Reduction,” J. Audio Eng. Soc., Vol. 43, pp. 233-240, 1995.

[5] T. Thiede and E. Kabot, “A New Perceptual Quality Measure for the Bit Rate Reduced Audio,” Audio Eng. Soc. 100th Convention, Copenhagen, Denmark, 1996.

[6] W. C. Treurniet, “Simulation of Individual Listeners with an Auditory Model,” Audio Eng. Soc. 100th Convention, Copenhagen, Denmark, 1996.

[7] T. Sporer, “Objective Audio Signal Evaluation – Applied Psychoacoustics for Modeling the Perceived Quality of Digital Audio,” Audio Eng. Soc. 103rd Convention, New York, USA, 1997.

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

[9] ITU-R Question 122/6, “Objective Perceptual Audio Quality Measurement Methods,” International Telecommunication Union, Geneva, Swiss, 2006.

[10] 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.

[11] S. George, S. Zielinski, and F. 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.

[12] J. Blauert, "Spatial Hearing: The Psychophysics of Human Sound Localization,” MIT Press, Boston, 1983.

[13] F. Baumgarte and C. Faller, “Binaural Cue Coding. Part I: Psychoacoustic Fundamentals and Design Principles,” IEEE Transactions on Speech and Audio Processing, Vol. 11 (6), pp. 509-519, 2003.

[14] ISO/IEC JTC1/SC29/WG11 (MPEG) Document 23003-1, 2006.

[15] J. W. Strutt, “On Our Perception of Sound Direction,” Philos. Mag. Vol. 13, pp. 214-232, 1907.

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

[17] I. Choi, S. B. Chon, and K.-M. Sung, "Measuring Spatial Attributes of Multi-channel Audio Coding Systems," Western Pacific Acoustics 9th Conference, Seoul, Korea, 2006.

[18] I. Choi, B. G. Shinn-Cunningham, S. B. Chon, and K.-M. Sung, "Prediction of Perceived Quality in Multi-channel Audio Compression Coding Systems," Audio Eng. Soc. 30th Conference, Saariselkä, Finland, 2007.

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

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

[21] S. Bech and N. Zacharov, "Perceptual Audio Evaluation - Theory, Method and Application," John Wiley & Sons, Chichester, 2006.

[22] ITU-R Recommendation BS. 1534-1, "Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)”, International

Telecommunication Union, Geneva, Swiss, 2001.

[23] 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.

[24] C. Kim, R. Mason and T. Brookes, "An Investigation into Head Movements Made When Evaluating Various Attributes of Sound," Audio Eng. Soc. 122nd Convention, Vienna, Austria, 2007.

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

[26] R. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "The Auditory Filter Bank," MRC-APU Report 2341, Cambridge, England, 1991.

[27] B. C. J. Moore, "An Introduction to the Psychology of Hearing," Academic Press, London, 1997.

[28] 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., vol. 45, pp. 224-240, 1997.

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

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

[31] E. A. Macpherson and J. C. Middlebrooks, "Listener Weighting of Cues for Lateral Angle: The Duplex Theory of Sound Localization Revisited," J. Acoust. Soc. Am. Vol. 111 (5), Pt. 1, pp. 2219-2236, 2002.

[32] L. A. Jeffress, "A Place Theory of Sound Localization," J. Comp. Physiol. Psychol. Vol. 41, pp. 35-39, 1948.

[33] P. X. Joris, P. H. Smith, and T. C. T. Yin, "Coincidence Detection in the Auditory System: 50 years after Jeffress," Neuron, Vol. 21, pp. 1235-1238, 1998.

[34] E. R. Hafter and R. H. Dye, "Detection of Interaural Differences of Time in Trains of High-frequency Clicks as a Function of Interclick Interval and Number," J. Acoust. Soc. Am. Vol. 73, pp. 644-651, 1983.

[35] L. R. Bernstein and C. Trahiotis, "Enhancing Sensitivity to Interaural Delays at High Frequencies by Using Transposed Stimuli," J. Acoust. Soc. Am., Vol. 112 (3), Pt. 1, pp. 1026-1036, 2002.

[36] G. C. Stecker, "Rate-limited, but Accurate, Central Processing of Interaural Time Differences in Modulated High-frequency Sounds. Focus on: Neural Sensitivity to Interaural Envelope Delays in the Inferior Colliculus of the Guinea Pig," J. Neurophysiology, Vol. 93, pp. 3048-3049, 2005.

[37] B. G. Shinn-Cunningham and K. Kawakyu, "Neural Representation of Source Direction in Reverberant Space," IEEE Workshop on Application of Signal Processing to Audio and Acoustics, New Paltz, USA, 2003.

[38] 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.