

An Objective Metric of Human Subjective Audio Quality Optimized for a Wide Range of Audio Fidelities

Charles D. Creusere, *Senior Member, IEEE*, Kumar D. Kallakuri, and Rahul Vanam, *Student Member, IEEE*

Abstract—The goal of this paper is to develop an audio quality metric that can accurately quantify subjective quality over audio fidelities ranging from highly impaired to perceptually lossless. As one example of its utility, such a metric would allow scalable audio coding algorithms to be easily optimized over their entire operating ranges. We have found that the ITU-recommended objective quality metric, ITU-R BS.1387, does not accurately predict subjective audio quality over the wide range of fidelity levels of interest to us. In developing the desired universal metric, we use as a starting point the model output variables (MOVs) that make up BS.1387 as well as the energy equalization truncation threshold which has been found to be particularly useful for highly impaired audio. To combine these MOVs into a single quality measure that is both accurate and robust, we have developed a hybrid least-squares/minimax optimization procedure. Our test results show that the min-max-optimized metric is up to 36% lower in maximum absolute error compared to a similar metric designed using the conventional least-squares procedure.

Index Terms—Audio quality metrics, metric optimization, objective metrics, perceptual audio analysis, quality evaluation, universal quality metrics.

I. INTRODUCTION

DIGITAL audio compression became truly practical with the work done by Johnston in the late 1980s incorporating a perceptual model of the human auditory system into the encoding process [1], [2]. Doing this allowed a compression algorithm to achieve significant bit-rate reductions while maintaining perceptually lossless quality, and the concept has lead directly to many of the most popular modern codecs (encoder/decoders) including MP3 (MPEG audio layer 3), MPEG-2 Advanced Audio Coder (AAC), MPEG-4, and Dolby Digital. In the late 1980s and early 1990s, many researchers recognized that traditional objective measures of codec performance such as segmental signal-to-noise ratio and mean squared error could not accurately assess the perceived quality

of audio produced by these more sophisticated algorithms. This led to the development of a number of objective measures of subjective audio quality, each with its own strengths and weaknesses [3]–[10]. In the early to mid 1990s, the International Telecommunications Union (ITU) formed a committee to study the problem and ultimately developed Recommendation BS.1387 (also called Perceptual Evaluation of Audio Quality, or PEAQ) which incorporated many of the previously developed metrics, defining them as model output variables (MOVs) [11].

The final version of ITU BS.1387 actually describes two different quality metrics: a lower complexity basic version and a more accurate advanced version. Henceforth, we refer to these here as ITU-basic and ITU-advanced, respectively. While these metrics were thoroughly validated using human subjective testing, it should be noted that they are both designed to operate on audio that is not significantly impaired, i.e., audio that is encoded to near perceptually lossless quality. They are not designed for nor have they been validated with the moderately to highly impaired audio that results when larger amounts of compression are required in an application. Thus, it should come as no great surprise that these metrics are highly inaccurate within this operating regime. Comparing the quality measure output by the BS.1387 metrics (called the ODG or objective difference grade) with the human subjective test data taken over a wide range of audio impairments as described in Section II, we find that the root mean squared prediction errors and correlation coefficients are 46.7 and 0.35 for ITU-basic while for ITU-advanced they are 44.5 and 0.59, respectively. Clearly, the ITU-recommended metrics are not universal in the sense that their outputs do not correlate strongly to quality as perceived by human test subjects over a wide range of audio fidelities or bit rates. It should be noted that the ODGs output by the BS.1387 metrics have been linearly rescaled to a range of 0 to 100 so that fair comparisons to the subjective ratings can be made.

To accomplish our goal of developing a more universal metric, we start with the MOVs used to create the ITU-basic and ITU-advanced quality estimates. Each of them quantifies some perceptual feature in an audio sequence, and their inclusion in the recommendation indicates that they have all proven to be useful in practice. We also consider as an additional MOV the energy equalization truncation threshold (EET) introduced by Creusere and shown to be especially effective by itself as a measure of quality in highly impaired audio [12]. To create a simple and robust universal metric for audio quality, we weight and linearly combine the outputs of a subset of MOVs that is

Manuscript received March 19, 2007; revised August 16, 2007. This work was supported by the National Science Foundation under Grant CCR 0133115. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

C. D. Creusere is with the Klipsch School of Electrical and Computer Engineering, New Mexico State University, Las Cruces, NM 88003 USA (e-mail: creusere@nmsu.edu).

K. D. Kallakuri is with Hughes Network Systems, Germantown, MD 20876 USA (e-mail: deepak_k_k@yahoo.com).

R. Vanam is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: rahulv@u.washington.edu).

Digital Object Identifier 10.1109/TASL.2007.907571

determined using a hybrid minimax/least-squares optimization procedure. It should be noted that results presented here are considerably different from those found in our preliminary research [13]–[16]. This is because all subjective testing in the earlier research was conducted using the differential comparison category rating (CCR) system [17] and only sequences encoded at 16 and 32 kb/s were evaluated. The Multi-Stimulus test with Hidden References and Anchor (MUSHRA) subjective testing protocol used here, on the other hand, is accurate over a wider range of audio impairments [18]. Also, audio compressed at four different bit rates has been considered here. Since the raw data used in the design process is more extensive and more accurate than that used in previous research presented by the authors [13]–[16], we expect the performance of the resulting metric to be more consistent with human subjective quality measurements. It should be noted that issues of stereo and multichannel perceptual quality are not directly addressed by the proposed metric. Furthermore, because the metric has been designed for and tested using music sequences, it may not be directly applicable to speech or general acoustic signals (e.g., sound effects).

This paper is organized as follows. Section II discusses the human subjective testing that we performed to gather the raw data required to design and validate our more universal quality metric. In Section III, the hybrid minimax/least-squares optimization strategy used to create the metric is described while comparative results are presented in Section IV. Finally, concluding remarks are made in Section V.

II. SUBJECTIVE TESTING

To design and validate an objective quality metric, it is necessary to perform human subjective testing. Numerous approaches to subjective audio testing have been used in the past, and a great deal of testing has been performed on the MPEG-4 suite of audio coding algorithms [19]–[27]. None of this work, however, is directly applicable here. In the 1999 MPEG-4 verification test [19], extensive perceptual testing was conducted on MPEG-4 audio compression algorithms including advanced audio coder (AAC), bit slice arithmetic coding (BSAC), and transform weighted interleaved vector quantization (TwinVQ). These tests are not applicable in the current context, however, for two reasons: 1) only rates higher than 64 kb/s were used in the BSAC comparisons, and 2) a testing methodology, ITU-R BS.1116, was used which has only been shown to be accurate for assessing the quality of near-perfect audio [28]. Because of differences in the testing procedures and the levels of impairment or due to the unavailability of audio test sequences, [20]–[27] are also similarly unhelpful to us.

Recently, the ITU has put forth recommendation ITU-R BS.1534-1 detailing a testing method called MUSHRA which has been shown to give reliable and accurate results for intermediate quality audio [18]. While it has only been fully validated for intermediate levels of impairment, the fact that the scoring of multiple sequences is performed interactively in a single trial and that it is done relative to the uncompressed reference audio leads one to expect that it should be reasonably accurate at all levels of impairment. This assumption is supported by the subjective data we have gathered. Therefore, we have adopted

this testing methodology for collecting the subjective data needed to design and validate a universal quality metric.

To perform the MUSHRA testing, we use a software package called System for Evaluation of Audio Quality (SEAQ) Version 2.03 produced by the Communications Research Centre, Ottawa, ON, Canada. This software allows us to present to a test subject as many as 12 test sequences plus the reference (unimpaired) sequence in a single trial, and it allows the subject to switch instantly between any of the sequences in a trial during playback. Thus, by swapping back and forth quickly between two test sequences, the subject can very accurately assess their relative qualities; by swapping back and forth between a test sequence and the reference sequence, the subject can assess the absolute quality. The subject can listen to a trial as many times as desired and can even force the audio to loop over small portions of the sequence in order to more accurately assess the differences. To complete a trial, the subject scores each test sequence in the range of 0 (extremely bad) to 100 (indistinguishable from the reference). Specifically, the range from 0–20 is labeled “bad,” the range from 20–40 is labeled “poor,” the range from 40–60 is labeled “fair,” the range from 60–80 is labeled “good,” and finally the range of 80–100 is labeled “excellent.” Since the authors do not know of a formal name for the units of the 0–100 point scale used by the MUSHRA protocol, we will define these units for the purpose of this work as MUSHRA Quality Units or MQU for short. Since the qualitative labels are mapped onto this 100-point scale in 20-point increments, we contend that a 20-point drop/rise in the MQU is equivalent to dropping/rising a whole quality grade. We will use this MQU-to-subjective quality mapping later in the paper to relate our numeric results back to human subjective opinion. Note that hidden reference and anchor signals can be used periodically as controls to validate the subjective data, allowing us to eliminate data collected from unreliable test subjects.

In total, we ran 20 trials with each of our 23 test subjects, and each trial had between two and four test sequences. A total of 48 different reconstructed sequences were tested, all of which were derived from five different monaural input audio sequences. These input sequences ranged from rock (Pat Benetar and Ronnie James Dio) to classical with durations of between 9 and 24 s. One of these sequences, *harpsichord*, is part of both the MPEG-4 and ITU-R BS.1387-1 audio test sets. We have chosen to use mostly nonstandard input sequences here since the standard sequences have usually been selected because they are very difficult to compress in a perceptually lossless manner and are thus useful in highlighting subtle deficiencies between high rate coding algorithms. A variety of coding algorithms are used in our trials, generating test sequences with encoding bit rates of 8, 16, 32, and 64 kb/s—well into the range of high audio impairment. Thus, in our subjective trials, our subjects are often being asked to gage how annoying different types of obvious audio impairments are relative to one another, and we felt that this could be done more accurately by the test subjects if the test sequences are taken from common and popular musical genres. Most of the coding algorithms used here are part of the MPEG-4 standard—specifically, AAC, BSAC, and TwinVQ [29]. We have also including sequences in these trials that were encoding using a newly developed fine-grained scalable compression algorithm derived from TwinVQ [30].

TABLE I
STATISTICS OF COLLECTED SUBJECTIVE DATA

<i>Bitrate</i>	<i>Mean MQU</i>	<i>Std. Dev.</i>	<i>Mean Conf.</i>	<i>Min. Conf.</i>	<i>Max. Conf.</i>
8 kb/s	43.7 ± 7.8	12.6	11.1	10.2	12.7
16 kb/s	51.7 ± 11.2	23.0	9.2	6.9	11.7
32 kb/s	64.1 ± 7.8	14.9	8.8	5.3	13.6
64 kb/s	80.4 ± 10.0	16.2	7.4	2.7	13.0

Our MUSHRA testing has resulted in a total of 48 subjectively scored audio sequences, each with a value between 0 and 100 that has been calculated by averaging over the scores of multiple test subjects and multiple presentations. The 95% confidence intervals for each of these average scores varies, of course, depending on the sample variance, but on the average over all 48 of them, it is ± 9.13 MQU. Relating this numeric value back human subjective quality, we can conclude that with 95% probability, the subjective quality estimated over all of the test sequences cannot be off by more than 1/2 a grade since 20 MQU separate each grade (e.g., “bad,” “fair,” “good,” etc.). Note that of these sequences, 10 were compressed at 8 kb/s, 15 at 16 kb/s, 14 at 32 kb/s, and 9 at 64 kb/s. Furthermore, of the 48 scored sequences, ten each were derived from three input audio sequences, while nine each were derived from the remaining two input sequences.

To further validate the accuracy and consistency of the collected data set, we have included various statistics estimated from it in Table I. Specifically, we have grouped the sequences by bit rate and have tabulated various statistics relating to the quality ratings given by the test subjects as well as the reliability of these ratings. Studying column 2, we first note that the mean MQU score increases with the bit rate as one would expect for any well-designed codec. Again, a 95% confidence interval is also estimated for each of the mean values. The increasing progression of mean MQU scores is important in two ways: it indicates that the collected data is reasonably accurate, and it supports our contention that the MUSHRA testing protocol is effective even for highly impaired audio (in this case, audio coded at 16 kb/s or less). For the 16- and 64-kb/s cases, the confidence intervals are larger than for the other two cases. While the standard deviation (column 3 in the table) is also higher in both of these cases, it is much higher in the 16-kb/s case. The reason for this is that the scalable BSAC codec performs very poorly at this bit rate with a mean MQU of about 20 while the mean MQU scores for the two TwinVQ-based codecs are around 50. The larger confidence interval at 64 kb/s is explained by the fact that we tested only nine sequences at this bit rate and that the TwinVQ-based codec performed consistently worse than the BSAC and ACC-based codecs.

In Table I, we also characterize the range of confidence intervals associated with each bit rate group. Specifically, we calculate the mean, minimum, and maximum of the confidence intervals for the test sequences that make up that group. Studying the table, we note that the mean confidence interval decreases as the bit rate increases. This indicates that the opinions of our test subjects about the quality of the audio are more consistent for less impaired audio than for highly impaired audio. This makes perfect sense. Four significantly different coding algorithms have been used to generate the test sequences, and at

low bit rates they each impair the reconstructed audio in unique ways. Thus, the rating provided by a given test subject will depend upon what form of impairment that subject finds to be most annoying. For example, the subject may have to choose between low bandwidth audio (i.e., audio that sounds like its coming out of a telephone) and higher bandwidth audio with noticeable harmonic distortion. Hence, there is an increase in the confidence interval as the level of impairment increases. It is interesting to note that the minimum of the confidence intervals also follows this trend while the maximum of the confidence interval appears to be more or less constant. This indicates that at every quality level (as indexed by the encoding bit rate), there is at least one test sequence for which there was less group consensus with respect to its quality. While the test sequences having the largest confidence intervals at a given bit rate are often from the classical music genre, in a few cases one of these same sequence encoded at a different bit rate or using a different algorithm will have a small confidence interval.

III. OPTIMIZATION METHODOLOGY

A. Least-Squares Weight Vector Design

As noted in Section I, ITU recommendation BS.1387 does not provide effective quality metrics over wide ranges of audio fidelity. The MOVs used in both the basic and advanced version of BS.1387, however, measure fundamental signal qualities that have been shown in numerous previous works to be related to the perception of audio quality [3]–[10]. Therefore, it makes sense to consider these MOVs as a starting point in developing a new and more versatile quality measure. Based on the results presented by Creusere [12], we also consider as an additional MOV the energy equalization threshold (EET) since it was found to be effective at high levels of audio impairment. These MOVs are summarized in Table II for both ITU-basic and ITU-advanced.

To create a scalar-valued estimate of audio quality, we use a least-squares procedure to find the optimal linear weighting for each MOV based on the human subjective testing detailed in Section II. It should be noted that both ITU-basic and ITU-advanced use three-layer neural networks to synthesize an objective difference grade (ODG)—a quality measure that varies between 0 (perfect) and -4 (highly impaired). We use instead a simple linear weighting here because we feel that such a solution is less susceptible to over training and thus likely to provide a more robust quality metric.

To calculate the weights, we find the standard least-squares solution to a system of linear equations, i.e.,

$$\mathbf{A}\mathbf{w} = \mathbf{p}. \quad (1)$$

Each element of vector \mathbf{p} is an average subjective score for a given audio sequence, determined using the MUSHRA testing

TABLE II
MODEL OUTPUT VARIABLES FOR BS.1387 BASIC AND ADVANCED

Number	ITU-Basic	ITU_advanced
1	BW_ref / Freq_ref	RMS Modulation Diff. A
2	BW_test / Freq_ref	RMS Noise Loudness
3	Total Noise Masking Ratio	Avg. Segmental NMR
4	Window Modulation Diff.	Avg. EHS
5	Avg. block Distortion	Avg. Linear Dist. A
6	EHS energy threshold	Energy Equalization Thresh.
7	Avg. Modulation Diff. 1	
8	Avg. Modulation Diff. 2	
9	RMS Noise Loudness	
10	Max. Filtered Prob. Detect.	
11	Relative Frame Distance	
12	Energy Equalization Thresh.	

procedure as outlined in Section II. Thus, \mathbf{p} is a 48×1 vector since 48 scored audio sequences have resulted from the subjective trials. Each row of matrix \mathbf{A} contains the MOV measurements calculated for the appropriate audio sequence. Consequently, \mathbf{A} is always a $48 \times N$ matrix where N is the number of MOVs used. The weight vector \mathbf{w} that minimizes the least-squares error is then found using the normal equations

$$\mathbf{w} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{p} \quad (2)$$

where $(\bullet)^T$ is the transpose operation and $(\bullet)^{-1}$ is the matrix inverse. The quality of an audio sequence is thus given by

$$q = \mathbf{w}^T \mathbf{m} \quad (3)$$

where \mathbf{m} is an $N \times 1$ vector containing the MOVs calculated for that sequence.

B. Minimax-Optimal MOV Selection

The solution in (2) is guaranteed always to minimize the mean squared error (mse) between the vector \mathbf{p} and a vector \mathbf{q} containing as its elements the respective q found using (3). It also has the effect of maximizing the correlation coefficient between the objective metric in (3) and the true subjective data. Thus, within the training set used to design the weight vector \mathbf{w} , the metric given by (3) is in every way optimal. Unfortunately, it is not possible create a training set that is large enough to represent both the incredible diversity of audio sequences (genre, artists, songs, etc.) and the large range of audio fidelities required for a universal metric. Consequently, it is worthwhile to sacrifice performance within the training set if we can achieve better performance outside of it and thus create a more robust metric.

Our approach to this problem is to perform a minimax optimization over the *set* of MOVs but use conventional least-squares to calculate the optimal weight vector for that particular set. In this optimization procedure, our cost function is the maximum absolute error (ME) between the objective metric output q and the subjective quality measured in the trials with the maximum being taken over the 48 test sequences. In minimizing this maximum error, we are searching for a solution which is not horribly bad for any given test sequence. In order to make this

TABLE III
OPTIMIZATION RESULTS FOR ITU-BASIC MODEL OUTPUT VARIABLE SET

Case	RMSE	ME	Corr. Coef.	Retained MOVs
LS Optimization	8.5	30.2	0.91	NA
LS Opt., no EET	8.8	32.4	0.91	NA
LS Opt., dual rate	6.3	25.1	0.95	NA
LS Opt., DR, no EET	6.4	25.9	0.95	NA
Minimax, HO	10.9	21.0	0.85	3, 8, 10, 12
Minimax	9.8	19.4	0.88	All but 5, 6, 11
Minimax, HO, no EET	12.1	22.6	0.82	2, 3, 7, 8, 9, 10
Minimax, no EET	11.5	21.4	0.84	2, 3, 4, 7, 8, 9, 10
Minimax, DR, HO	9.1	18.6	0.89	2, 3, 8, 10, 12
Minimax, DR	9.1	18.6	0.89	2, 3, 8, 10, 12
Minimax, DR, HO, no EET	10.3	20.2	0.85	3, 4, 7, 8, 10
Minimax, DR, no EET	10.3	20.2	0.85	3, 4, 7, 8, 10

solution even more robust outside of the training set, we also consider as a cost function the maximum absolute error with holdout. To calculate this, we design the weights of our metric for a selected MOV set using (2) but without one of the audio sequences and then apply the metric to the held-out sequence to calculate the ME. We do this for all 48 sequences and take the largest ME value as the cost. By minimizing over this cost function, we expect the resulting weight vector to be more robust since the cost is being evaluated (in a successive fashion) outside of the training set.

The minimax optimization algorithm described in the paragraph above can be summarized as follows.

- 1) Select a subset of MOVs.
- 2) Use (2) to find the least-squares optimal weight vector \mathbf{w} .
- 3) Compute maximum squared error (with or without holdout) over the 48 test sequences: i.e., $e = \max_{1 \leq n \leq 48} (q_n - p_n)^2$ where n indexes the test sequence while q_n and p_n are, respectively, the objective [from (3)] and subjective quality measurements for the n th test sequence.
- 4) Repeat 1 to 3 until all possible subsets of MOVs have been evaluated. Select the subset of MOVs that minimizes e .

In the Section IV, we apply this optimization to the subjective data from Section II and characterize its performance relative to the conventional least-squares approach.

IV. RESULTS AND DISCUSSION

The results of the minimax optimization introduced in Section III are summarized in Table III for ITU-basic and Table IV for ITU-advanced. In each table, six different cases are considered as follows:

- 1) conventional least squares optimization over the entire data set;

TABLE IV
OPTIMIZATION RESULTS FOR ITU-ADVANCED MODEL OUTPUT VARIABLE SET

<i>Case</i>	<i>RMSE</i>	<i>ME</i>	<i>Corr. Coef.</i>	<i>Retained MOVs</i>
LS Optimization	19.9	53.9	0.47	NA
LS Opt., no EET	23.2	53.3	0.43	NA
LS Opt., dual rate	15.7	46.7	0.68	NA
LS Opt., DR, no EET	16.3	45.8	0.66	NA
Minimax, HO	24.3	54.5	0.38	1, 2, 4, 5
Minimax	22.4	52.7	0.43	1, 3, 4, 5, 6
Minimax, HO, no EET	24.3	54.5	0.38	1, 2, 4, 5
Minimax, no EET	23.3	52.8	0.42	1, 3, 4, 5
Minimax, dual rate, HO	16.3	42.7	0.66	1, 3, 4, 5, 6
Minimax, dual rate	16.3	42.7	0.66	1, 3, 4, 5, 6
Minimax, DR, HO, no EET	17.3	45.0	0.62	1, 3, 4, 5
Minimax, DR, no EET	17.3	45.0	0.62	1, 3, 4, 5

- 2) separating the low and high bit rate audio sequences and determining separate sets of least-squares optimal weights for each bit rate;
- 3) minimizing the maximum absolute error as applied to the holdout case;
- 4) minimizing the maximum absolute error without holding out any sequences;
- 5) the same as (3) but separately optimized at low and high bit rates as in (2);
- 6) the same as (4) but with high and low bit rates as in (2).

For each of these cases, we also evaluate the impact of the EET MOV in isolation by either allowing its selection in the optimization process or not. The root mean squared error (rmse), ME, correlation coefficient, and the list of MOVs retained by the optimization process are tabulated in the columns. The correlation coefficient is a dimensionless quantity with a value between -1 and 1 and is defined as

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (4)$$

where $\text{cov}()$ is the cross-covariance between the two arguments and the σ 's are the standard deviations. In our case, random variable X is the perceptual rating predicted by the objective metric, while random variable Y is the subjective rating determined through testing, and the moments required to evaluate (4) are estimated directly from the test data. The value of ρ thus characterizes the predictive power of the objective metric: if ρ is very close to 1 , then the objective metric is providing very good estimates of the subjectively measured quality, while if it is close to zero, our metric is doing no better than random guessing (negative values of ρ indicate the degree of anticorrelation which is not an issue here since we can always reverse the sign on the predicted value). Note that the conventional least squares optimization results summarized in the first four rows of both tables do not use the minimax optimization procedure to form subsets of MOVs and thus column number 5 does not apply.

Studying these tables, we clearly see that the optimization based on ITU-basic MOVs is superior to that based on ITU-advanced MOVs in every possible respect. For example, in terms of rmse, the ITU-advanced results are almost twice as large as those of the ITU-basic derived metric. To give more insight into these numeric results, we must relate them back to the original subjective testing criteria as defined in the MUSHRA protocols and as discussed in Section II. Specifically, the rmse, ME-H, and ME results given in Tables III and IV are all scaled in terms of MQU, and they can therefore be related directly back to perceived audio quality. As an example, we compare the minimax optimization case with holdout (labeled “Minimax, HO”) in Tables III and IV: the rmse values are 10.9 MQU and 24.3 MQU for ITU-basic and ITU-advanced, respectively. Given that 20 MQU separate each quality grade (e.g., “fair” to “good”), we see that the quality predicted by the ITU-advanced metric is, on the average, off by more than a full quality grade while that predicted by the ITU-basic metric is off by half a grade. In terms of maximum absolute error, the comparison between the two metrics is similar in most cases. While it may seem surprising at first glance that the ITU-advanced metric performs so poorly, it should be noted that both algorithms are designed to operate at fidelities close to perceptually losslessness. In this regime, ITU-advanced is undoubtedly superior to ITU-basic, but neither of them is useful over a wide range of audio fidelities as was mentioned in Section I. Thus, there is no reason to expect that linear combinations of ITU-advanced MOVs should necessarily result in a metric that is any more accurate in this scenario than combinations of ITU-basic MOVs. Given the poor performance of the metric derived from the ITU-advanced MOV set, we focus primarily on metrics derived from the ITU-basic MOV set for the remainder of this section.

Most of the results summarized in Table III are exactly as one would expect. Specifically, forming our metric using an optimal linear combination of all 12 MOVs always results in the lowest mse and the largest correlation coefficient. Furthermore, designing two sets of optimal weights—one for the 8- and 16-kb/s sequences and the other for the 32- and 64-kb/s sequences—results in the absolute lowest rmse (6.3) and the highest correlation coefficient (0.95). An rmse of 6.3 implies that the average quality prediction is accurate to approximately a third of a quality grade. Note that such a metric is only feasible when the bit rate of the compressed sequence is available. If we have access to the compressed representation of the audio, then this information is either directly available or can be easily derived. On the other hand, if we only have access to the reconstructed audio sequence, we cannot use a bit rate-conditioned metric.

When the minimax procedure outlined in Section III is applied, the rmse and correlation coefficient degrade somewhat, but the ME is significantly reduced. For example, the ME for the single-rate case drops 36% when minimax optimization is applied: from 30.2 to 19.4 MQU. In perceptual terms, this means that the worst case prediction error is 1.5 quality grades for LS optimization versus one quality grade for minimax optimization. For example, an audio sequence whose quality is predicted using standard least-square optimization to be in the middle of the “fair” range at 50 MQU might have an actual subjective quality anywhere between 20 MQU (“bad” range) and 80 MQU

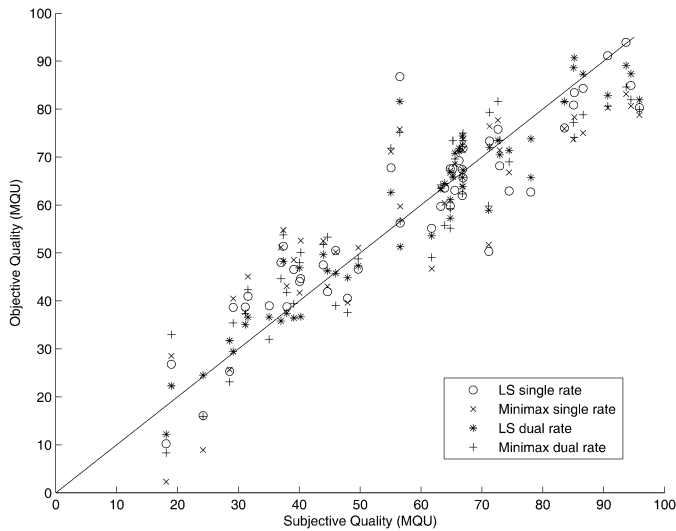


Fig. 1. Scatter plot of objective versus subjective quality measurements for 48 test sequences. Minimax and least-squares optimization are considered for the single- and dual-rate metrics. A 45° line has been superimposed on the plot to indicate the ideal case.

(“excellent” range). Using the minimax optimized metric, this range narrows to 31 MQU (“fair”) to 69 MQU (“good”). Similar results are achieved for the dual-rate case with a reduction in the ME of 26% for the minimax optimized case relative to conventional least-squares, equating to an MQU reduction of 6.5 or almost a third of a subjective quality grade.

These results can also be interpreted graphically by means of a scatter plot. Such a plot is shown in Fig. 1 where the objective quality estimate from ITU-basic MOVs is plotted versus the subjective data for each of the 48 test sequences. Specifically, we consider both the single and dual rate metrics, comparing the proposed minimax optimized metrics with the conventional least-squares metrics. Examining the figure, it does appear that the data points are generally clustered closer to the optimal 45° for both the single- and dual-rate least-squares metrics. This result is, of course, guaranteed by the least-squares optimization cost function. One also clearly notes from the figure, however, that the outliers for both the single- and dual-rate least-squares optimizations are further from the optimal line than their respective minimax optimized solutions. Again, this is exactly what we would expect to see.

Referring to Table III, we note that the results for minimax optimization with holdout are never better (and, in fact, are worse for the single-rate case) than similar results that do not remove the data sample being predicted from the predictor design process. Recall that we held this sample out of the least-squares design process to force the final solution to be more robust to data points outside of the training set. Thus, it is impossible for the minimax with holdout solution to result in better rmse or ME performance than the conventional minimax solution over this training set. The solution, however, should be more robust outside of this training set since the prediction error that is calculated for each training sample with a metric which excludes that sample. Note that we only postulate this robustness: we cannot prove or guarantee it.

Studying the set of MOVs retained by the minimax optimization procedure, it is interesting to note that four of them—3, 8,

10, and 12—are retained in all four cases. Referring to Table II, we note that these four MOVs correspond to Total Noise Masking Ratio, Average Modulation Difference 2, Maximum Filtered Probability of Detection, and the Energy Equalization Truncation Threshold. Note again that this last MOV is not part of the ITU-basic set but rather was introduced by Creusere [12]. For the single-rate case with holdout, using a least-squares optimal linear weighting of just these four MOVs provides the optimal minimax solution. Both dual-rate minimax optimization cases add just a single MOV to this set—2 or the bandwidth of the test sequence divided by the frequency reference—while the single-rate case, optimized for ME without holdout, uses all but three of the original MOVs.

To quantify the significance of the EET MOV, we have also performed all of the optimizations in Tables III and IV without including it. For the ITU-basic results of Table III, the largest increase in the rmse incurred by eliminating the EET MOV is 1.7 for the single-rate minimax optimization case while the largest ME increase incurred is 2.2 for single-rate least-squares optimization. Clearly, such small changes are not, in and of themselves, perceptually relevant. It is worth noting, however, that if we remove *any* single MOV from the optimization process, the impact on the rmse is small. Specifically, for the single-rate least-squares case, removing MOV number 10 (Maximum Filtered Probability of Detection) will have the most impact with an increase in rmse of 1.28—a perceptually insignificant amount. Furthermore, removing any one of eight other MOVs causes a smaller increase in the rmse than the 0.3 increase incurred by removing the EET MOV. One could claim that EET is more relevant to the accuracy of the metric than all but three of the other MOVs, but a more reasonable way to view these results is that no single MOV parameter truly dominates the metric. Instead, a group of these MOVs together are responsible for this metric’s predictive accuracy.

At this point, one might ask why certain MOVs are more relevant to the metric than others. We certainly do not have the complete answer to this question, but we can postulate as to why some MOVs might be more useful in predicting audio quality across a wide range of fidelities than others. Studying the last column of Table III, we note that MOVs 3, 8, 10, and 12 are selected in every optimization where EET (MOV 12) is allowed. MOV 3 is the total noise masking ratio (TNMR) which was developed by Brandenburg originally as a stand-alone quality metric [4]. The TNMR basically quantifies the amount by which the estimated noise power exceeds a signal-dependent masking threshold, averaged across the entire audio sequence. For a given audio sequence and at any specific time instant within that sequence, the masking threshold is fixed since it depends only on the reference (i.e., uncompressed, original) audio. Thus, when comparisons are made between versions of a specific audio sequence reconstructed to differing levels of fidelity, the result of this MOV should mirror the increasing or decreasing fidelity of the reconstruction, although not necessarily in a linear fashion.

The average modulation difference 2 (MOV 8) measures the introduction of signal modulations by the codec into passages where the reference audio contains few or no modulations. Given the earlier observation by Creusere [12] that the BSAC codec introduces a large number of spurious modulations into the reconstructed audio at low bit rates, it is easy to believe

TABLE V
RESULTING MOV WEIGHTS FOR ITU-BASIC OPTIMIZATIONS IN TABLE III

<i>MOV</i>	<i>LS</i>	<i>LS-DR-LBR</i>	<i>LS-DR-HBR</i>	<i>MM-HO</i>	<i>MM</i>	<i>MM-DR-LBR</i>	<i>MM-DR-HBR</i>
1	-0.003	-0.027	1.156	~	0.018	~	~
2	0.018	0.014	-0.018	~	0.031	0.022	0.076
3	4.034	1.188	0.989	0.900	2.771	2.796	1.034
4	0.386	-3.436	1.331	~	-0.296	~	~
5	-26.86	-7.417	-68.48	~	~	~	~
6	52.58	-119.71	274.36	~	~	~	~
7	-0.249	6.099	-2.401	~	0.339	~	~
8	-0.113	-0.776	0.244	-0.145	-0.149	-0.196	-0.098
9	-7.914	-13.97	1.364	~	-4.977	~	~
10	134.18	146.34	-656.02	93.855	74.971	97.920	58.430
11	26.29	-28.92	-149.36	~	~	~	~
12	-0.540	-0.500	0.081	-1.200	-1.070	-1.461	-0.982

that this MOV is helping to quantify the magnitude of these distortions. Note that the EET MOV was originally designed to quantify such distortion as well [12], so there may well be some redundancy in the information provided by these two MOVs. The final ITU-basic MOV that is consistently selected in the minimax optimization is the maximum filtered probability of detection or MFPD (MOV 10 in Table II). This MOV attempts to quantify the probability that a human listener will be able to detect a difference between the reference and test audio signals. In particular, it is a temporal average that weights the detection probabilities for later time frames in the sequence more heavily than for earlier time frames. This is consistent with the psychoacoustic observation that listeners remember more clearly the auditory differences in the portion of the signal that they have heard most recently. The MFPD is ultimately derived from a model of the hair-cell excitation pattern induced inside of the inner ear by a given sound, and it is highly non-linear with respect to superpositions of input sounds. While one would expect the probability that a listener detects an audible difference between the test and reference sequences to increase as the fidelity of the test sequence drops, there is really no reason to expect the output of this MOV to track perceived relative quality when the level of impairment is so high that a listener cannot help but notice it. Thus, further study is needed to determine exactly why the MFPD is particularly useful as part of an objective quality metric that is optimized to operate well outside of the low impairment range.

As noted in Table III, optimization with holdout results in only four retained MOVs while optimization without holdout results in nine retained MOVs. Projecting the minimax solution into the 4-D space of the minimax-with-holdout solution, we find the the Euclidean distance between these two solutions is 18.98. Interestingly, the weights for MOVs 8 (average modulation difference 2) and 12 (EET) are very close for both solutions, while the weight for MOV 3 has the greatest percentage difference. While we previously noted a relationship between the former two MOVs, it is not clear how this relationship might induce the relative invariance observed in the weights selected for the two different optimization cost functions. Table V lists the sets of weights that results from each of the EET-inclusive

optimizations summarized in Table III (the tilde indicates that a weight is not used).

V. CONCLUSION

We have developed a new and more universal audio quality metric using MOVs from ITU-recommended BS.1387 combined with the previously developed energy equalization truncation threshold. Specifically, this metric is designed to characterize audio quality over a wide range of fidelities, from highly impaired to perceptually lossless. To both design and validate the proposed metric, we have collected human subjective test data using the MUSHRA protocol, and we have developed a hybrid least-squares/minimax optimization procedure in an effort to maximize its robustness outside of our relatively limited set of training data.

An obvious area of future research is to better understand psychoacoustically why a small subset of MOVs appears to be particularly useful in developing a more general quality metric. This is clearly a very difficult research problem that would require some very clever human subjective testing to determine exactly how the level of impairment being measured by a given MOV is perceived by human listeners. Of course, the human perception of audio quality is truly dynamic in time: we might perceive one part of an audio sequence to be good but another to be highly impaired. Thus, a dynamic metric for audio quality which could provide a measure of the audio quality versus time would also be highly desirable. Such a metric could be used to optimize bit allocations in the encoder, and it would allow for a worst case analysis of the quality of audio sequences. While the proposed optimization framework can easily support a dynamic metric, the process of validating it versus the human perception of audio quality would be very challenging because human subjective quality would need to be monitored dynamically as the test sequence was being played back.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their many constructive suggestions which greatly improved the final version of this paper.

REFERENCES

- [1] J. D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1988, pp. 2524–2527.
- [2] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, no. 2, pp. 314–323, Feb. 1988.
- [3] T. Thiede and E. Kabot, "A new perceptual quality measure for bit rate reduced audio," in *Proc. Contribution 100th AES Convention*, Copenhagen, Denmark, 1996, preprint 4280.
- [4] K. Brandenburg, "Evaluation of quality for audio encoding at low bit rates," in *Proc. Contribution 82nd AES Convention*, London, U.K., 1987, preprint 2433.
- [5] T. Sporer, "Objective audio signal evaluation—Applied psychoacoustics for modeling the perceived quality of digital audio," in *Proc. 103rd AES Convention*, New York, Oct. 1997, preprint 4280.
- [6] J. G. Beerends and J. Stemerdink, "A perceptual audio quality measure based on a psycho-acoustical representation," *J. Audio Eng. Soc.*, vol. 40, pp. 963–978, Dec. 1992.
- [7] B. Paillard, P. Mabilieu, S. Morissette, and J. Soumagne, "Perceval: Perceptual evaluation of the quality of audio signals," *J. Audio Eng. Soc.*, vol. 40, pp. 21–31, 1992.
- [8] C. Colomes, M. Lever, J. B. Rault, and Y. F. Dehery, "A perceptual model applied to audio bit-rate reduction," *J. Audio Eng. Soc.*, vol. 43, pp. 233–240, Apr. 1995.
- [9] E. Zwicker and R. Feldtkeller, *Das Ohr als Nachrichtenempfänger*. Stuttgart, Germany: Hirzel Verlag.
- [10] D. R. Guard, M. P. Hollier, and M. O. J. Hawksford, "Objective perceptual analysis: Comparing the audible performance of data reduction schemes," in *Proc. 96th Convention Audio Eng. Soc.*, 1994, preprint 3797.
- [11] "Methods for objective measurement of perceived audio quality," Rec. ITU-R BS.1387-1, 1998–2001 [Online]. Available: www.itu.int
- [12] C. D. Creusere, "Understanding perceptual distortion in MPEG scalable audio coding," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 422–431, May 2005.
- [13] K. D. S. Kallakuri, "An improved quality metric for highly to moderately impaired audio," M.S. thesis, New Mexico State Univ., Las Cruces, NM, Dec. 2004.
- [14] R. Vanam, "Scalable perceptual metric for evaluating audio impairment," M.S. thesis, New Mexico State Univ., Las Cruces, NM, Jun. 2005.
- [15] R. Vanam and C. D. Creusere, "Evaluating low bitrate scalable audio quality using advanced version of PEAQ and energy equalization approach," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, vol. 3, pp. 189–192.
- [16] R. Vanam and C. D. Creusere, "Scalable perceptual metric for evaluating audio quality," in *Proc. Conf. Rec. 39th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Nov. 2005, pp. 319–323.
- [17] "Subjective performance assessment of telephone-band and wide-bandwidth digital codecs," Rec. ITU-R P.830, 1996 [Online]. Available: www.itu.int
- [18] Method for the subjective assessment of intermediate quality level of coding systems," Rec. ITU-R BS.1534-1, 2001–2003 [Online]. Available: <http://www.itu.int>
- [19] *Report on the MPEG-4 Audio Version 2 Verification Test: N3075*, ISO/IEC JTC1/SC29/WG11, Int. Org. Standardization, Dec. 1999.
- [20] *MPEG-4 Audio: Results of AAC and Twin VQ Tool Comparative Tests*, ISO/IEC JTC1/SC29/WG11/N2011, Feb. 1998.
- [21] *MPEG-4 Audio Verification Test Results: Audio on Internet*, ISO/IEC JTC1/SC29/WG11/MPEG98/N2425, October 1998 [Online]. Available: <http://www.tnt.uni-hannover.de/project/mpeg/audio/public/w2425.pdf>
- [22] A. Aggarwal and K. Rose, "A conditional enhancement-layer quantizer for the scalable MPEG advanced audio coder," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 1833–1836.
- [23] A. Aggarwal and K. Rose, "Approaches to improve quantization performance over the scalable advanced audio coder," in *Proc. 112th Convention AES*, 2002, preprint 5557.
- [24] G. Stoll and F. Kozamernik, "EBU listening tests on internet audio codecs," EBU Tech. Rev., No. 283, Jun. 2000 [Online]. Available: http://www.ebu.ch/trev_home.html
- [25] "EBU subjective listening tests on low-bitrate audio codecs," Tech. 3296, Jun. 2003.
- [26] G. Soulodre, T. Grusec, M. Lavoie, and L. Thibault, "Subjective evaluation of state-of-the-art 2-channel audio codecs," in *Proc. 104th AES Convention*, May 1998, preprint 4740.
- [27] G. Soulodre, T. Grusec, M. Lavoie, and L. Thibault, "Subjective evaluation of state-of-the-art two-channel audio codecs," *J. Audio Eng. Soc.*, vol. 46, no. 3, pp. 164–177, Mar. 1998.
- [28] "Methods for subjective assessment of small impairments in audio systems including multichannel sound systems," Rec. ITU-R BS.1116, 1994 [Online]. Available: <http://www.itu.int>
- [29] K. Brandenburg, O. Kunz, and A. Sugiyama, "MPEG-4 natural audio coding," *Signal Process.: Image Commun.*, vol. 15, no. 1, pp. 423–444, Jan. 2000.
- [30] S. Kandadai and C. D. Creusere, "Perceptually-weighted audio coding that scales to extremely low bitrates," in *Proc. Data Compression Conf.*, Snowbird, UT, Mar. 2006, pp. 382–391.



Charles D. Creusere (SM'04) received the B.S. degree in electrical and computer engineering from the University of California, Davis, in 1985 and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, in 1990 and 1993, respectively.

After receiving the B.S. degree, he went to work for the Naval Weapons Center, China Lake, CA. In 1989, he was awarded a Fellowship from the DoD to attend graduate school in Santa Barbara where he worked with Prof. S. Mitra in the area of multirate filter banks. He has been an Associate Professor in the Klipsch School of Electrical and Computer Engineering, New Mexico State University, Las Cruces, since 2000. His current research interests include image, video, and audio processing for variety of purposes including remote sensing and signal understanding.

He was an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2002 to 2005.



Kumar D. Kallakuri was born in Rajahmundry, Andhra Pradesh, India, in 1981. He received the B.E. degree in electronics and instrumentation from Andhra University, Visakhapatnam, India in 2002 and the M.S. degree from New Mexico State University, Las Cruces, in 2004.

He was a DSP Software Engineer with Floreat, Inc., Saratoga, CA, from 2005 to 2006. Since 2006, he has been a DSP Engineer at Hughes Network Systems, Germantown, MD.



Rahul Vanam (S'07) received the B.E. degree in electronics and communication engineering from Bangalore University, Bangalore, India, in 2000, and the M.S.E.E. degree from the New Mexico State University, Las Cruces, in 2005. He is currently pursuing the Ph.D. degree in electrical engineering at the University of Washington, Seattle.

He worked from 2000 to 2003 for the DSP and Multimedia group of Wipro Technologies, Bangalore, India. He was an Intern with Nvidia, Inc., Santa Clara, CA, and Thomson Corporate Research, Princeton, NJ. His research interests include video coding and perceptual audio coding.