Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2011, Article ID 472185, 14 pages
doi:10.1155/2011/472185

Research Article

Gradient Ascent Subjective Multimedia Quality Testing

Stephen Voran and Andrew Catellier

United States Department of Commerce, National Telecommunications and Information Administration, Institute for Telecommunication Sciences, Telecommunications Theory Division, 325 Broadway, Boulder, CO 80305, USA

Correspondence should be addressed to Stephen Voran, svoran@its.bldrdoc.gov

Received 14 October 2010; Accepted 14 January 2011

Academic Editor: Vittorio Baroncini

Subjective testing is the most direct means of assessing multimedia quality as experienced by users. When multiple dimensions must be evaluated, these tests can become slow and costly. We present gradient ascent subjective testing (GAST) as an efficient way to locate optimizing sets of coding or transmission parameter values. GAST combines gradient ascent optimization techniques with subjective test trials. As a proof-of-concept, we used GAST to search a two-dimensional parameter space for the known region of maximal audio quality, using paired-comparison listening trials. That region was located accurately and much more efficiently than use of an exhaustive search. We also used GAST to search a two-dimensional quantizer design space for a point of maximal image quality, using side-by-side paired-comparison trials. The point of maximal image quality was efficiently located, and the corresponding quantizer shape and deadzone agree closely with the quantizer specifications for JPEG 2000, Part 1.

1. Introduction

Subjective testing is arguably the most basic and direct way to assess the user-perceived quality of image, video, audio, and multimedia presentations. Through careful selection of signals, presentation environments, presentation protocols, and test subjects, one can approximate a real-world scenario and acquire a representative sample of user perceptions for that scenario. Test protocols for audio [1, 2], video and still images [3], and multimedia [4] have been standardized. Subjective testing generally requires specialized equipment, software, laboratory environments, skills, and numerous human test subjects. These elements equate to significant expenses and weeks or months of work.

Objective estimators of perceived quality can reduce or eliminate many expenses and complications inherent in subjective testing [5–8]. But these savings come with a distinct cost—objective estimates can vary widely in their ability to track human perception and judgement. When new classes of visual or auditory distortions need to be evaluated, the limitations become crippling—there is no way to know how well an objective estimator will perform until there are subjective test results to compare it to. Yet once the subjective test is done, the question is answered for that class of distortions.

Between the subjective and objective testing lies another option: subjective testing with improved efficiency, that is, gathering more information using fewer experimental trials. Efficiency is critical when one needs to optimize a family of coding or transmission parameters that interact with each other.

For example, given a fixed available transmission bit-rate constraint (or storage file size constraint), one might seek to optimally partition those bits between basic signal coding and redundancy that improves robustness to transmission errors or losses (e.g., multidescriptive coding or forward error correction). Or one might wish to optimally allocate bits among several quantizers to produce a reduced-rate signal representation for an individual signal. And it may be necessary to find an optimal partitioning of bits between different signal components in a multimedia program. In each of these cases one is seeking a point in a multidimensional parameter space that produces maximal perceived quality. This can be a large and arduous quality assessment task.

One can design a subjective test to do an exhaustive search (ES) of a discretized version of the parameter space

using an absolute category rating (ACR) subjective test to evaluate each point in the space. But this can require the evaluation of a very large number of points, and it also requires one to guess at how to best discretize the parameter space.

In practice, if faced with the prospect of ES, one would likely iterate first testing a coarse sampling of the space using only a few subjects to roughly locate the region of maximal quality, and then further testing a finer sampling of that region using a larger number of subjects. This is an intuitive but ad hoc approach—at each iteration one must guess the appropriate discretization (both resolution and number of points) and the appropriate number of subjects to use. Or one might seek to iterate through a sequence of one-dimensional optimizations, but this approach will generally be very limiting and slow.

We present gradient ascent subjective testing (GAST) as an efficient alternative to ES ACR testing (and to ad hoc shortcuts). A preliminary version of this work and portions of this manuscript were previously published by the authors of [9]. GAST can efficiently and adaptively select a subset of points in the space to evaluate, eliminating any need to manually impose arbitrary discretizations on the space or to manually iterate testing protocols. GAST can incorporate the ACR approach but is particularly well matched to paired-comparison (PC) testing.

Some prior work towards more efficient subjective testing exists. It has been proposed that in some cases a range of values for a single video coding parameter can be searched for a quality maximum by setting up an interactive control (e.g., a slider) and allowing subjects to adjust it at will until a maximal level of video quality is perceived [10]. One might seek to extend this to multiple parameters, in which case subjects could be facing very difficult and lengthy tasks. GAST naturally searches multiple dimensions while test subjects interact with the same simple univariate PC or ACR test protocol.

A quality matching scheme that uses an interactive control is described in [11]. Here, the control is adjusted until a quality match between two side-by-side video players is perceived. This takes advantage of the power of paired-comparisons for quality matching in one dimension but does not apply to multidimensional optimization.

The adaptive psychometric testing method in [12] uses subject responses to modify stimulus levels so that they efficiently converge to the threshold of perception. This is a powerful univariate threshold locating technique but it does not address multidimensional optimization.

In Section 2, we describe the GAST algorithm. Section 3.1 details a proof-of-concept experiment using the GAST algorithm to identify a known region of maximal audio quality in a two-dimensional parameter space. In this experiment the region of maximal audio quality was identified accurately and efficiently. In Section 3.2 we describe an image-quality experiment. Here, we used GAST to identify values of two related wavelet coefficient quantization parameters (dead-zone and shape) that maximize image quality. Discussion and observations are provided in Section 4.

2. Gradient Ascent Subjective Testing Algorithm

Finding the point in n-dimensional space that approximately maximizes (or minimizes) an objective function defined on that space is a classic problem and many different avenues to its solution have been offered over the years. Such background is far beyond the scope of this paper, but numerous texts provide detailed expositions of the development of these approaches, their relative strengths and weaknesses, and the relationships among them [13–16].

A unifying key idea is to evaluate the objective function at a small number of intelligently selected points, use those results to select more points, and thus continue to better locate the desired maximal point. This may involve only function values (direct-search methods), first derivatives of the function (gradient methods), or both first and second derivatives (second-order methods). Key performance attributes that differentiate the various methods are convergence and efficiency.

We wish to optimize perceived quality on an n-dimensional parameter space—the objective function is perceived quality, and it will be evaluated by human subjects. Thus, a GAST algorithm implementation platform includes a computer and one or more human subjects. Software calculates a pair of points in the parameter space where the objective function (perceived quality) should be evaluated and then facilitates the presentation of stimuli associated with this pair of points. The subject evaluates the two stimuli relative to each other, and the software uses the response to then calculate the next pair of points to evaluate. The software and the subject continue this interplay until termination criteria indicate that it is likely that a point of maximum quality has been located.

Our approach could be built on any number of optimization algorithms. We have elected to use a basic gradient ascent algorithm because it seems well matched to expected properties of our actual applications (i.e., smooth, slowly varying objective functions with fairly broad maxima that can only be imprecisely evaluated). The GAST algorithm iterates between two main steps: finding the direction that produces maximum quality increase (direction of steepest ascent), and then exploring that direction to the maximum extent by performing a line search for a quality maximum. Each of these steps requires subjective scores from a test subject.

2.1. Subjective Scores. The GAST algorithm requires subjective scores to find directions and to search lines. Ultimately these scores must describe perceived quality at one point in the parameter space relative to a second point. Almost any subjective testing scale could be used and scores could be appropriately processed to get this relative quality information.

But paired-comparison (PC) testing scales are particularly well suited to the GAST algorithm. Here, the testing protocol directly extracts relative quality information. Examples of PC (sometimes called "forced choice") protocols can be found in [1–3]. Two stimuli are presented, and a subject indicates any preference between the two. For visual

stimuli, either sequential or side-by-side presentations are possible. Another option is to employ an A/B switch that allows the subject to switch between the two stimuli at will. For auditory stimuli, the options are sequential presentation and A/B switching.

PC testing has the added benefit that comparing two stimuli can often be an easier task for subjects than providing absolute ratings for two stimuli presented in isolation from each other. An easier task can result in reduced variation in individual performance of that task, thus reducing undesired variation in subjective test results.

The assignment of the two signals to the two presentation positions (first or second, left or right, A or B) can be randomized on a per-trial basis, as long as the resulting score is processed to compensate for that randomization. Outside of this processing, PC scores can be used directly. If other testing scales are used, then pairs of scores can be additionally processed (e.g., subtracted) to conform with this convention.

We use $S(\mathbf{x},\mathbf{y})$ to represent the (possibly processed) subjective score resulting from the presentation of the signal parameterized by the vector $\mathbf{x}$ (representing a point in n-dimensional space) and the signal parameterized by the vector $\mathbf{y}$. Positive values of $S(\mathbf{x},\mathbf{y})$ indicate that the $\mathbf{y}$ signal was preferred to the $\mathbf{x}$ signal, negative values indicate the opposite, and zero indicates that there was no preference.

2.2. Direction Finding. Consider a point in an n-dimensional space represented by a column vector $\mathbf{x}$. We seek to find the direction in which the objective function increases most rapidly. The direction-finding algorithm finds an approximate solution using between n and $2 \cdot n$ finite differences. Let

$$\mathbf{x}_k^{\pm} = \mathbf{x} \pm \Delta_d \cdot \mathbf{I}^k, \quad k = 1, 2, \ldots, n, \tag{1}$$

indicate a point near $\mathbf{x}$ differing from $\mathbf{x}$ in only the kth dimension. In (1), Δ_d is a fixed scalar direction-finding step size, and $\mathbf{I}^k$ is the kth column of the $n \times n$ identity matrix. Δ_d needs to be large enough to cause detectable changes in perceived quality, but small enough to provide accurate localized information about those changes.

The direction-finding algorithm gathers subjective scores $S(\mathbf{x}, \mathbf{x}_k^{\pm})$ for each dimension k, as allowed. If the parameter space is bounded, $\mathbf{x}_k^+$ or $\mathbf{x}_k^-$ could be outside the parameter space, the corresponding signal would not exist, and the corresponding subjective score would not exist. If only one subjective score exists for dimension k, then the corresponding element $\delta_k(\mathbf{x})$ of the direction vector $\boldsymbol{\delta}(\mathbf{x})$ is given by

$$\delta_k(\mathbf{x}) = \frac{S\left(\mathbf{x}, \mathbf{x}_k^{\pm}\right)}{\pm\Delta_d}. \tag{2}$$

For dimensions where both subjective scores exist, $\delta_k(\mathbf{x})$ is given by

$$\delta_k(\mathbf{x}) = 0, \quad \text{when } S\left(\mathbf{x}, \mathbf{x}_k^-\right) < 0,\ S\left(\mathbf{x}, \mathbf{x}_k^+\right) < 0\ , \tag{3}$$

$$\delta_k(\mathbf{x}) = \frac{S\left(\mathbf{x}, \mathbf{x}_k^+\right) - S\left(\mathbf{x}, \mathbf{x}_k^-\right)}{2\Delta_d}, \quad \text{otherwise.} \tag{4}$$

Equation (3) treats the special case where $\mathbf{x}$ is located at a maximum in dimension k. Equation (4) treats the general case where two subjective scores are available and uses them together to approximate an average local slope in dimension k. Finally, if $\mathbf{x}$ is on the boundary of the parameter space and $\delta_k(\mathbf{x})$ points outside the space, the search terminates.

Once $\delta_k(\mathbf{x})$ has been calculated for all n dimensions, the resulting direction vector $\boldsymbol{\delta}(\mathbf{x})$ is scaled to have unit norm:

$$\hat{\boldsymbol{\delta}}(\mathbf{x}) = \frac{\boldsymbol{\delta}(\mathbf{x})}{|\boldsymbol{\delta}(\mathbf{x})|}. \tag{5}$$

The result is a unit-norm vector $\hat{\boldsymbol{\delta}}(\mathbf{x})$ that provides an approximate indication of the direction in which the objective function increases most rapidly. It is an approximate result because it is based on finite differences in the parameter space, and because the subjective scores are constrained to five distinct values. The impact of this approximation will depend on the specific context in which GAST is used. Our proof-of-concept experiment was unhindered by this approximation.

2.3. Golden Section Line Search. Given an arbitrary line segment in parameter space, the iterative line search algorithm in GAST finds the point on that line segment that approximately maximizes the objective function. The algorithm is initialized by a point represented by the column vector, $\mathbf{x_0}$, a unit-norm direction vector, $\hat{\boldsymbol{\delta}}(\mathbf{x_0})$, and a boundary definition for the parameter space. The first step is to find the line segment (or "line" for brevity) that runs in the direction $\hat{\boldsymbol{\delta}}(\mathbf{x_0})$ from $\mathbf{x_0}$ to the boundary of the parameter space. We call the second end of this line $\mathbf{x_3}$.

This line is the input to the iterative portion of the algorithm. Each iteration results in a new, shorter line that is evaluated on the next iteration. This evaluation is based on the comparison of the objective function at two interior points that lie on this line. These points are called $\mathbf{x_1}$ and $\mathbf{x_2}$ and are ordered as shown in Figure 1. If $S(\mathbf{x_1}, \mathbf{x_2}) < 0$ (consistent with the example of the solid line), then the new line to search on the next iteration is the line between $\mathbf{x_0}$ and $\mathbf{x_2}$. If $0 < S(\mathbf{x_1}, \mathbf{x_2})$ (consistent with the example of the broken line), then the new line to search is the line between $\mathbf{x_1}$ and $\mathbf{x_3}$.

Motivated by a desire for predictable convergence, we add the constraint that each iteration must scale the line down by a constant value $0 < \gamma < 1$, regardless of which interval is chosen as the new interval. This means that

$$|\mathbf{x_2} - \mathbf{x_0}| = |\mathbf{x_3} - \mathbf{x_1}| = \gamma|\mathbf{x_3} - \mathbf{x_0}|, \tag{6}$$

$$|\mathbf{x_1} - \mathbf{x_0}| = |\mathbf{x_3} - \mathbf{x_0}| - |\mathbf{x_3} - \mathbf{x_1}| = (1 - \gamma)|\mathbf{x_3} - \mathbf{x_0}|. \tag{7}$$

Regardless of the subjective score, the new shorter line (between $\mathbf{x_0}$ and $\mathbf{x_2}$ or between $\mathbf{x_1}$ and $\mathbf{x_3}$) always inherits an interior point from the longer line ($\mathbf{x_1}$ in first case and $\mathbf{x_2}$ in the second case). Motivated by a desire to use paired comparisons efficiently, we add the constraint that this inherited (from iteration i) interior point must be one of the two interior points evaluated in iteration $i + 1$.

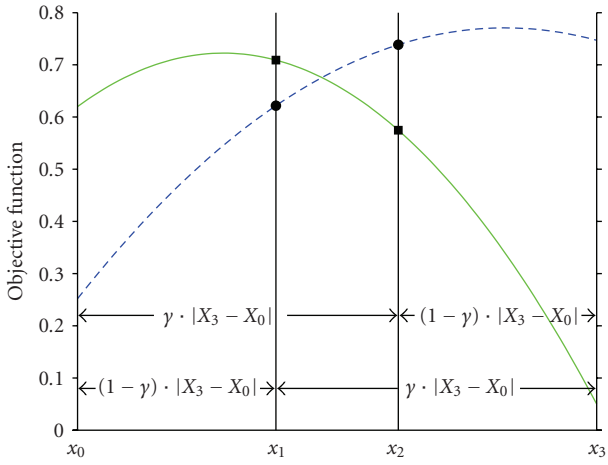


FIGURE 1: Example relationships for four points in the line search.

Consider the case where the result of iteration i is the line between $\mathbf{x_0}$ and $\mathbf{x_2}$ (consistent with the solid line in the example of Figure 1). That new shorter line inherits the interior point $\mathbf{x_1}$. In iteration $i+1$ a second interior point must be added. If this new point is inserted to the left of $\mathbf{x_1}$, then $\mathbf{x_1}$ would now (iteration $i+1$) serve the role that $\mathbf{x_2}$ played in iteration i. Using (6) we conclude that

$$|\mathbf{x}_1 - \mathbf{x_0}| = \gamma^2 |\mathbf{x_3} - \mathbf{x_0}|. \tag{8}$$

Comparing (7) and (8) we conclude that

$$\gamma^2 = (1-\gamma) \quad \text{so } \gamma = \frac{-1+\sqrt{5}}{2}. \tag{9}$$

Finally,

$$\frac{1}{\gamma} = \gamma + 1 = \frac{1+\sqrt{5}}{2} = \varphi \approx 1.618\,. \tag{10}$$

If the new point is inserted to the right of $\mathbf{x_1}$, then $\mathbf{x_1}$ would now (iteration $i+1$) serve the same role that it played in iteration i. Using (6) and (7) we conclude that

$$|\mathbf{x_1} - \mathbf{x_0}| = (1-\gamma)|\mathbf{x_3} - \mathbf{x_0}| = (1-\gamma)\gamma|\mathbf{x_3} - \mathbf{x_0}|, \tag{11}$$

but this can only be solved by $\gamma = 1$, which violates the allowed range on γ. Thus the new point must be inserted to the left of $\mathbf{x_1}$.

If iteration i produces the line between $\mathbf{x_1}$ and $\mathbf{x_3}$ (consistent with the broken line in the example of Figure 1), an analogous set of results will follow. Thus, $\gamma = 1/\varphi$ is the only value to use in (6) and (7) to locate $\mathbf{x_1}$ and $\mathbf{x_2}$ so that the uniform-scaling-per-iteration constraint and the interior-point-reuse constraint are satisfied. The line to search scales by $\gamma = 1/\varphi$ at each iteration. The irrational number φ is called the golden section or golden mean. It defines an aesthetically pleasing rectangle that has been used widely in architecture and art and also lends its name to this line search algorithm [16].

In GAST this golden section line search iterates until $S(\mathbf{x_1}, \mathbf{x_2}) = 0$ and $|\mathbf{x_2} - \mathbf{x_1}| < \Delta_t$, where Δ_t is a termination parameter. This condition indicates that there is no preference between two signals whose parameterizations are sufficiently close to each other. The algorithm returns $(1/2)(\mathbf{x_2} + \mathbf{x_1})$ as the approximation to the point on the original line where the objective function is maximized. Our proof-of-concept experiments indicate that the approximation is a good one. If $S(\mathbf{x_1}, \mathbf{x_2}) = 0$ when $\Delta_t \le |\mathbf{x_2} - \mathbf{x_1}|$, then $\mathbf{x_1}$ and $\mathbf{x_2}$ are moved apart in increments until a nonzero vote is returned. This is a special case that breaks from the golden section constraints.

2.4. Entire Algorithm. To start the GAST algorithm, one must select a starting point, $\mathbf{x_0}$, in the n-dimensional parameter space. We have successfully used both deterministic points on the boundary of the space and randomly selected interior points. The direction-finding algorithm is applied to find $\hat{\boldsymbol{\delta}}(\mathbf{x_0})$, indicating the direction of steepest ascent from $\mathbf{x_0}$. Next, $\mathbf{x_0}$ and $\hat{\boldsymbol{\delta}}(\mathbf{x_0})$ are provided to the line search algorithm, which searches in the direction $\hat{\boldsymbol{\delta}}(\mathbf{x_0})$ from $\mathbf{x_0}$ to the boundary of the search space and returns the maximizing point $\mathbf{x_1}$.

The direction-finding algorithm is then used to find $\hat{\boldsymbol{\delta}}(\mathbf{x_1})$, which shows the direction of steepest ascent from $\mathbf{x_1}$. Line searching and direction finding continue to alternate in this fashion until a terminating condition is satisfied. At any iteration, the output of the last line search is the best approximation to the point in the parameter space that maximizes the objective function.

One terminating condition is $\hat{\delta}(\mathbf{x_i}) = 0$, since this indicates that there is no direction to move from $\mathbf{x_i}$ to increase the objective function. Equations (2) through (4) show that this could be due to subjective scores of zero (no differences detected), a local maximum, or a local minimum that is judged to be perfectly symmetrical in all n dimensions. Terminating in a local minimum is not desirable; so if this is deemed a possibility, one should test for it (the test is analogous to the one in (3)) and restart the GAST algorithm from a new starting point as necessary. The algorithm also terminates if the distance between the input and output points of a line search is less than Δ_t, since future iterations will be unlikely to move the result outside that neighborhood.

The GAST algorithm climbs the surface of the objective function to find a maximal value. If multiple local maxima exist, the algorithm will find one of them but there is no guarantee that it will be the global maximum. If multiple local maxima are suspected, then multiple trials using multiple starting places will help to identify them.

2.5. GAST Algorithm Implementation. The direction finding and the golden section line search algorithms were coded inside objects called "tunes" (since our first experiment involved musical excerpts) such that all calculations take place transparently to an outer algorithm that facilitates subject interaction. The outer algorithm needs only to instantiate said tunes by specifying $\mathbf{x_0}$, Δ_d, and Δ_t, request parameter pairs associated with the signal pairs that are

presented, submit subjective scores, and keep track of all tune objects that it instantiated.

The outer algorithm is also responsible for drawing a graphical user interface to be used by the subject, as well as instantiating, polling, and updating necessary tune objects, presenting signals to subjects, handling subject votes, randomizing tune play order, and ensuring that each search terminates. The MNRU and T-Reference algorithms described in Section 3.1 execute rapidly; so it was possible to generate the required audio signals just before they were played. Likewise, the image processing described in Section 3.2 executes very quickly and the required pairs of images were created on demand.

For our second experiment, "tune" objects were renamed to be "pics," but they and the outer algorithm were otherwise largely unchanged. Fixes for two unforeseen corner cases were integrated, methods to store and retrieve metadata were added, and 3D graph support was added to the plotting code. A terminating condition was added that prevented the algorithm from initiating a sixth-direction finding stage, used the resting point of the fifth line search for the overall resting place of the object, and marked the object (i.e., GAST task) as complete. Finally, the ability to randomly reverse parameter output order and compensate the subjective scores for this reversal (thus randomizing stimulus presentation order) was added to the objects, thus relieving the outer algorithm of that responsibility.

GAST software is available at http://www.its.bldrdoc.gov/audio/ for those who wish to experiment with the GAST technique.

3. GAST Experiments

We have applied GAST in three different applications. Our initial experiment was a proof-of-concept experiment using audio reference conditions to create a simple, controlled quality surface over a two-dimensional parameter space. The experiment and the results are described in Section 3.1. We later used GAST to find the optimizing values of two quantization parameters in a wavelet-based image compression scheme and full details are given in Section 3.2.

In an additional experiment, we created a modified version of the GAST algorithm to locate quality matches, rather than quality maxima. The application was a one-dimensional experiment, and the goal was to identify bit-error rates (BER) that resulted in specific reference speech quality levels. In one-dimensional problems there is only one line to search—no direction finding is required. Each paired comparison involved a reference recording and a recording from the speech coder under test at the BER under test. The result of the comparison would cause the BER to be increased or decreased accordingly (a line search) until the point of equivalence was found.

Each of the three experiments has affirmed the utility and efficacy of the GAST algorithm.

3.1. Audio Quality GAST. As an initial test of the GAST concept, we devised an audio experiment using two reference conditions that simulate audio coding. The use of two reference conditions (instead of two actual coding or transmission system parameters) allowed us to create a two-dimensional parameter space with a known region of maximal audio quality.

3.1.1. Audio Quality Parameter Space. Audio signals were passed through the two reference conditions in sequence to generate a controlled, known quality surface over a two-dimensional parameter space. The first reference condition was the modulated noise reference unit (MNRU) [17]. This condition adds signal-correlated Gaussian noise to the audio signal at the specified SNR of Q dB:

$$y_k = x_k + x_k \cdot n_k \cdot 10^{-Q/20} = x_k \cdot \left(1 + n_k \cdot 10^{-Q/20}\right), \quad (12)$$

where x_k, y_k, and n_k are input, output, and unit-variance zero-mean Gaussian noise samples, respectively. The noise added by the MNRU sounds like that produced by some waveform coders.

The second reference condition was modeled after the T-Reference described in [18, 19]. The T-Reference imparts a controlled level of audio distortion through short-term time warping. This distortion can be described as "warbling" or "burbling" and is similar to that produced by some parametric coders.

The T-Reference operates on frames of 256 audio samples (5.8 milliseconds). In each group of three sequential frames, the first is temporally compressed, the second is untouched, and the third is temporally stretched.

More specifically, with frames labeled 1 through N, the T-Reference applies temporal compression to frames numbered $1 + 3 \cdot k$, it does not change frames numbered $2 + 3 \cdot k$, and it applies temporal expansion to frames numbered $3 + 3 \cdot k, k = 0, 1, 2, \ldots$. Temporal compression is accomplished by deleting every Tth sample, and the complementary temporal expansion is accomplished by interpolating a sample between every Tth and $T+1$st sample. Since $\lfloor 256/T \rfloor$ samples are deleted from the first frame in the group and the same number of samples are interpolated into the third frame in the group, the total number of samples in each group of three frames is preserved at $3 \cdot 256$.

The unit-less parameter T can be set to any integer in the range from 2 to 256. Larger values of T correspond to less distortion.

We developed GAST software to work in a normalized $[0, 1]$ parameter space. Thus, we mapped this range to Q and T values according to

$$\begin{aligned} Q &= -85 \cdot p_1^2 + 100 \cdot p_1, \\ T &= 1 + \left\lfloor 2^{(-15 \cdot p_2^2 + 13 \cdot p_2 + 2)} \right\rfloor, \end{aligned} \quad (13)$$

where $\lfloor \cdot \rfloor$ denotes rounding to the nearest integer. These relationships are displayed in Figure 2. They were selected to smoothly traverse a wide range of Q and T values and have different shapes, asymmetric slopes, and a single interior maximum for both Q and T.

From Figure 2 we can conclude that in the two-dimensional space (p_1, p_2), there is a line segment of

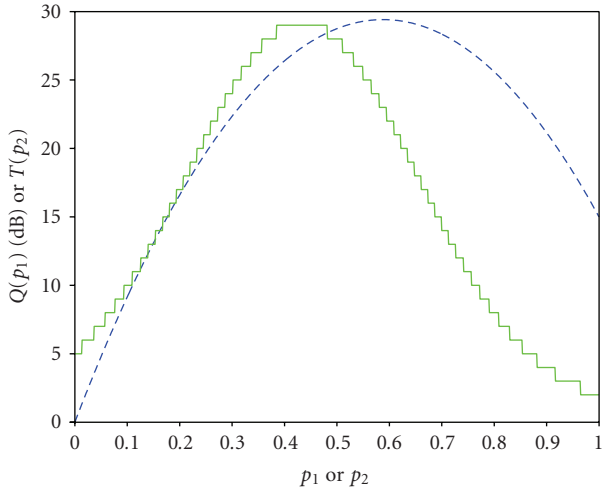


FIGURE 2: Q as a function of p_1 (dashed), and T as a function of p_2 (solid).

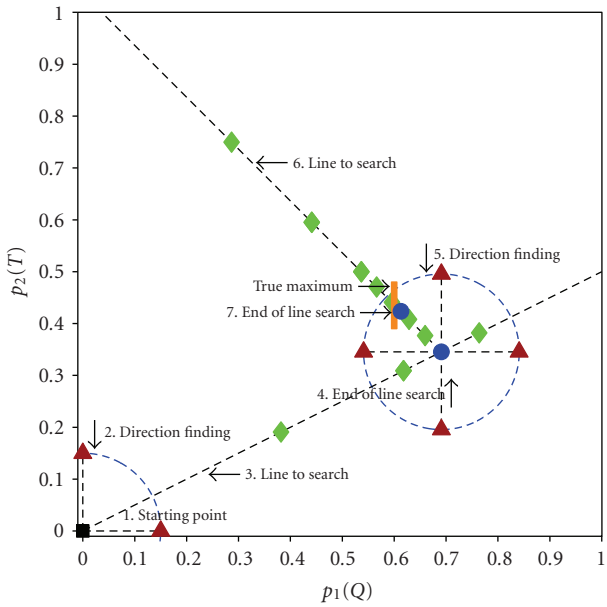


FIGURE 3: Example trajectory of an audio experiment GAST trial; details are in text.

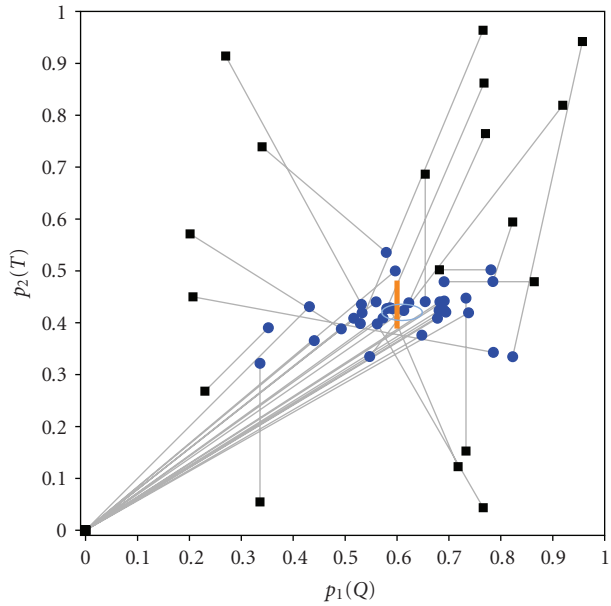


FIGURE 4: Start and end points for 35 audio experiment GAST trials shown with black squares and blue circles, respectively. The light blue ellipse shows the mean and 95-percent confidence interval for all end points. The bold orange vertical line represents region of numerically maximal audio quality.

numerically maximal audio quality extending from the point (0.60,0.39) to the point (0.60,0.48). This segment is shown as a solid vertical line in Figures 3 and 4. The reference condition parameter values associated with this region of maximal audio quality are $Q = 29.4$ dB and $T = 29$.

3.1.2. Audio Quality Protocol. This audio GAST experiment used eight five-second musical segments covering a range of instruments and musical styles. These were excerpted from compact discs and the native sample rate of 44,100 samples per second was maintained through the experiment.

A PC testing protocol was used. Two audio signals were presented sequentially and five possible subjective responses were allowed: "The audio quality of the second recording is much better than, better than, the same as, worse than, or much worse than, the first recording." The associated subjective scores are 2, 1, 0, −1, and −2, respectively. After the presentation of each pair of signals, a subject could submit a vote or request to hear the pair played again.

Subjects were seated in a sound isolated room with background noise measured below 20 dBA SPL. Audio signals were presented through studio-quality headphones at the individually preferred listening level. A PDA was used to present the prompts and collect the votes.

Six subjects participated in the experiment. Each ran the GAST algorithm on four of the eight musical selections, using two different starting places per selection. One starting place was the origin of the parameter space; the other was randomly chosen for each musical selection and each subject. Thus, each subject started eight different GAST tasks, and in each trial the subject made one step of progress on one task randomly selected from the eight. We used the direction-finding step size $\Delta_d = 0.15$ and the terminating condition $\Delta_t = 0.20$.

3.1.3. Audio Quality Results. In this initial GAST experiment, some tasks ended prematurely due to implementation issues, subject time limitations, and lack of a quality gradient near the corners of the parameter space. Excluding these special cases, the GAST algorithm consistently located a point of maximal perceived quality and then terminated as expected.

Figure 3 shows an example GAST task trajectory. The region of numerically maximal audio quality is shown with a bold orange vertical line. The square at the origin indicates the starting location. The triangles connected

to that square indicate the two points used in the first direction-finding step. The audio signal parameterized by the triangle at $(0.15, 0)$ was voted "much better" than the signal associated with the origin; so $S((0,0)^T, (0.15,0)^T) = 2$, where $(\cdot)^T$ indicates the transpose operator. Similarly, $S((0,0)^T, (0,0.15)^T) = 1$.

These two scores yielded the normalized direction vector $\hat{\boldsymbol{\delta}}(\mathbf{x}) = (1/\sqrt{5}) \cdot (2,1)^T$ and this led to a search of the line that runs up and to the right. Points played on this line are shown with diamonds, and the result of the line search is shown with a circle. The four points connected to that circle were played as part of the second direction-finding step. This led to a search of the line that runs toward the upper left corner of the figure. Again, points played are shown with diamonds, and the final result is shown with a circle. This result is very close to the location of numerically maximum audio quality. This task required 13 votes.

Different musical selections can reveal or mask distortions in different ways, and these distortions may be perceived differently by individual subjects. Thus, perceived quality is a function of signals and subjects as well as the device under test. Averaging results over a representative sample of relevant signals and subjects gives the most meaningful perceived quality results.

Figure 4 shows the GAST algorithm start (black squares) and end (blue circles) points for the 35 audio experiment GAST tasks that ran to completion. An average of 15.6 votes was required per task. The end points cluster around the line segment of numerically maximal audio quality (the bold orange vertical line), as expected. The mean and 95-percent confidence intervals for the p_1 and p_2 dimensions are shown with a light blue ellipse. For the 35 combinations of subjects and musical selections, we are 95 percent confident that the mean location of maximal perceived audio quality is between 0.571 and 0.649 in p_1 dimension ($29.1 \le Q \le 29.4$ dB), and between 0.404 and 0.436 in the p_2 dimension ($T = 29$). This result is consistent with the known location of numerically maximal audio quality and required $15.6 \times 35 = 546$ PC presentations (not including any replays) and 546 votes.

To locate this point with the same resolution using ES ACR testing, one would need about 13 samples $((0.649 - 0.571)^{-1} = 12.8)$ in the p_1 dimension and 32 samples $((0.436 - 0.404)^{-1} = 31.3)$ in the P_2 dimension, resulting in a 416-sample grid on the parameter space. Evaluating each point with all 35 combinations of musical selections and subjects would require $416 \times 35 = 14,560$ ACR presentations (not including any replays) and votes. This is a lower bound. If 35 trials per point in the parameter space do not result in statistically significant differences between adjacent parameter space samples in the neighborhood of the quality maximum, then additional trials would be required to locate the maximum with a resolution that matches GAST. Thus, we find that the number of votes required is reduced by at least a factor of $14,560/546 = 26.7$.

Figure 5 shows the average convergence of the 35 GAST trials. Seventeen trials started at the origin and eighteen started at random locations. The resulting average Euclidean distance between starting places and the nearest point in the region of maximal audio quality is 0.54. With each iteration of the GAST algorithm this average distance decreases and an asymptotic value of 0.1 is approached after two iterations.

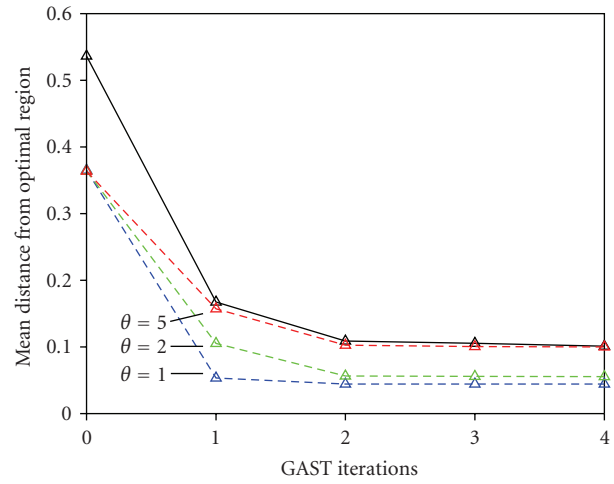


Figure 5: Average convergence performance for human subjects and Monte Carlo simulations for a parametrized family of "perfect subjects."

Figure 5 also shows the results of three Monte Carlo simulations. In these simulations, software emulated a family of "perfect subjects." These hypothetical subjects could decompose the audio signals and independently measure the levels of impairment due to MNRU and T-Reference relative to the best audio quality in the experiment ($Q_{\max} = 29.4$ and $T_{\max} = 29$):

$$\zeta_i = \sqrt{(Q_i - Q_{\max})^2 + \left(\frac{1}{2}(T_i - T_{\max})\right)^2}. \tag{14}$$

The index $i = 1, 2$ indicates internal measurements for the first and second audio recordings heard, respectively. Changes in T are harder to detect than changes in Q and the factor of 1/2 in (14) provides a very rough match between the two scales.

The "perfect subjects" then voted with perfect consistency but finite sensitivity (θ) according to

$$\begin{aligned}
(\zeta_1 - \zeta_2) &\le -2\theta \Longrightarrow S = -2 \text{ (much worse)},\\
-2\theta < (\zeta_1 - \zeta_2) &\le -\theta \Longrightarrow S = -1 \text{ (worse)},\\
-\theta < (\zeta_1 - \zeta_2) &< \theta \Longrightarrow S = 0 \text{ (same)},\\
\theta \le (\zeta_1 - \zeta_2) &< 2\theta \Longrightarrow S = 1 \text{ (better)},\\
2\theta \le (\zeta_1 - \zeta_2) &\Longrightarrow S = 2 \text{ (much better)}.
\end{aligned} \tag{15}$$

For each simulation 16,000 tasks with random starting places were used. This produced an average initial distance of 0.37.

As expected, smaller values of θ result in quicker convergence to lower asymptotic distance values. The setting $\theta = 5$ gives an average convergence curve similar to that of our human subjects, excepting the fact that the average starting distances are different. This corresponds to a baseline

MNRU sensitivity of $Q = 5$ dB and a baseline T-Reference sensitivity of 10 T units.

3.2. Image Quality GAST. We were invited to contribute our work on the GAST algorithm to this special issue of this journal. This motivated us to apply the GAST algorithm to image quality assessment to demonstrate its applicability in that domain.

A typical problem in image coding is rate minimization: minimize the number of bits used to encode an image while holding the image quality at or above some target level (e.g., transparent coding). The dual to this problem is the quality maximization problem: maximize image quality while holding the bit-rate at some constant value. This problem fits well with GAST and is the subject of the experiment.

3.2.1. Image Quality Parameter Space. There are many image coding frameworks that one could invoke for this experiment and we elected to use the JPEG 2000 framework [20–22]. JPEG 2000 is generally considered an advance over the original DCT-based JPEG standard [23] in terms of rate-distortion performance, and this advance comes with additional cost in terms of computational complexity. JPEG 2000 offers lossy-to-lossless progressive coding, scalable resolution, region of interest features, and random access. JPEG 2000 is used in digital cinema, fingerprint databases, remote sensing applications, and medical imaging [22]. We recognize JPEG 2000 as a mature, successful, and highly optimized coding technique. As such, it also provides a natural basis for further investigations in image coding.

Lossy JPEG 2000 compression transforms level-shifted YUV pixel values with the Daubechies 9/7 discrete wavelet transform (DWT). The key to minimizing rate or maximizing quality in JPEG 2000 lies in the quantization and encoding of the resulting DWT coefficients. In typical operation, the quantization step-size is made much smaller than would be ultimately necessary—"overquantization" is performed. This is followed by a multipass bit-plane significance coding algorithm with lossless entropy coding that uses an adaptive arithmetic coding strategy. The quantization and coding stages are tied together through a sophisticated rate-control algorithm that seeks to reduce mean-squared error (MSE) or visually weighted MSE as much as possible as it assigns the available bits.

Quantization of DWT coefficients in the context of JPEG 2000 has been studied extensively. The basis functions of the DWT decomposition from different levels and orientations have differing visual importances. Quantization noise imposed on the associated coefficients produces visual distortions that are localized in spatial frequency and orientation and can also be correlated to the image. Thus, quantization noise on different DWT coefficients will have differing levels of visibility.

The pioneering experiments in [24] found visibility thresholds for each of the various levels and orientations of the wavelet basis functions. These thresholds translate to step-sizes for uniform quantizers—following these step sizes would keep DWT quantization noise for each individual DWT basis function below the visible threshold.

Numerous additional empirical studies and theoretical derivations have treated the topics of contrast sensitivity functions, visual summation of quantization errors, self-masking, neighborhood masking, and others. (These often jointly address the intrinsically linked issues of quantization and rate control.) Individual examples can be found in [25–28] and more comprehensive overviews can be found in [22, 29]. Much of this work has been incorporated (perhaps implicitly) into JPEG 2000, Part 1, and (more explicitly) into Part 2.

Our GAST experiment also treats the quantization of DWT coefficients. Instead of overquantizing and then seeking rate reduction in a coding stage, we use GAST to drive the design of rate-constrained, nonuniform quantizers with arbitrary dead-zones that maximize image quality. Clearly, this is not a proposal for a practical image coding implementation. Instead, it is an experimental investigation of nonuniform quantization and arbitrary dead-zones in the context of DWT coefficients. This investigation is driven by true human visual perception (not MSE, SNR, or a visually based computed distortion metric). To our knowledge, both the optimization problem and the optimization technique that we describe below are unique.

We apply the Daubechies 9/7 DWT to each color plane of a 512×512 pixel image with 8 bits/pixel, successively decomposing it to four levels. (Four levels are sufficient to capture most of the available DWT benefit in this context.) At the fourth level the coefficients of each orientation (LL, LH, HL, and HH) form a 32×32 block ($32 = 512 \times 2^{-4}$). Coefficients from the LH and HL orientations follow the same Laplacian distribution:

$$f_c(c) = \frac{1}{\sqrt{2}\sigma} e^{-|c|(\sqrt{2}/\sigma)} \tag{16}$$

so they can share the same quantizer design.

We use GAST to optimize two design parameters for a single quantizer for the fourth-level, Y-plane coefficients from the LH and HL orientations. These are the only coefficients we quantized before application of the inverse DWT to reconstruct the image. The majority of the energy (and thus the majority of the coding problem) lies in the coefficients of the final, fourth level. Additional similar experiments could be designed to further investigate quantization of coefficients from the LL orientation (typically modeled by the Generalized Gaussian distribution or the uniform distribution), the HH orientation (modeled by Laplacian distribution but with lower variance than LH/HL coefficients), or coefficients from lower levels of the decomposition (Laplacian but with lower variance than coefficients from the fourth level).

A histogram (taken across 43 images) confirms that the distribution of the fourth-level, Y-plane, LH/HL DWT coefficients approximately matches that of the zero-mean Laplacian random variable. To allow finite quantization, we limit the coefficient magnitudes to 1200 (limiting occurs for about 0.01% of the coefficients). For ease of presentation here, and without loss of generality, we scale the limited DWT coefficients to the range $[-1, 1]$.

Next we define the quantizer $Q(c, \Delta_{dz}, \alpha, N)$ that operates on the DWT coefficient c:

$$\begin{aligned} |c| \le \Delta_{dz} &\Longrightarrow Q(c, \Delta_{dz}, \alpha, N) = 0, \\ \Delta_{dz} < |c| &\Longrightarrow Q(c, \Delta_{dz}, \alpha, N) \\ &= \text{sign}(c)\left\lceil NF_\alpha\left(\frac{|c| - \Delta_{dz}}{1 - \Delta_{dz}}\right)\right\rceil, \end{aligned} \tag{17}$$

where the compander function $F_\alpha(\cdot)$ is defined:

$$\begin{aligned} \alpha = 0 &\Longrightarrow F_\alpha(x) = x, \\ \alpha \neq 0 &\Longrightarrow F_\alpha(x) = \frac{1 - e^{-\alpha x}}{1 - e^{-\alpha}}. \end{aligned} \tag{18}$$

The quantizer dead-zone is defined by $\Delta_{dz}, 0 < \Delta_{dz} < 1$. The dead-zone extends from $-\Delta_{dz}$ to $+\Delta_{dz}$, so the dead-zone width is $2\Delta_{dz}$, and coefficient values in this range are reconstructed as zero. In addition to this central cell, the quantizer has N cells to cover the remaining negative range and N cells to cover the remaining positive range ($N = 1, 2, 3, \ldots$). Thus the quantizer has $2N + 1$ quantization cells total and it maps real numbers in the interval $[-1, 1]$ to the integers $\{-N, -(N-1), \ldots, N-1, N\}$.

In addition, the quantizer shape (the local quantizer cell width relationship) is controlled by α ($-\infty < \alpha < +\infty$) through the compander function $F_\alpha(\cdot)$. This function maps the range $[-1, 1]$ onto itself. When $\alpha = 0$, $F_\alpha(\cdot)$ is linear and the resulting quantizer has uniform cell widths (with the possible exception of the central, dead-zone cell). If $0 < \alpha$, the resulting quantizer has cell widths that increase as one moves away from the origin. Increasing α strengthens the effect. When $\alpha < 0$, quantizer cell widths decrease as one moves away from the origin and the effect is strengthened by decreasing α. Examples of the quantizer input-output relationship defined by (17) and (18) are shown in Figure 6. Equations (17) and (18) emphasize that nonuniform quantizers can be implemented by a nonlinear function followed by a uniform quantizer.

An approximation, $\tilde{c}$, to the original coefficient value, c, can be recovered by the inverse quantizer:

$$\begin{aligned} Q(c) = 0 &\Longrightarrow \tilde{c} = 0, \\ Q(c) \neq 0 &\Longrightarrow \tilde{c} \\ &= \text{sign}(Q(c))\left((1 - \Delta_{dz})G_\alpha\left(\frac{|Q(c)| - 0.5}{N}\right) + \Delta_{dz}\right), \end{aligned} \tag{19}$$

where the compander function $G_\alpha(\cdot)$ is introduced in order to exactly invert the operation of $F_\alpha(\cdot)$:

$$\begin{aligned} \alpha = 0 &\Longrightarrow G_\alpha(x) = x, \\ \alpha \neq 0 &\Longrightarrow G_\alpha(x) = \frac{-\ln(1 - x(1 - e^{-\alpha}))}{\alpha}. \end{aligned} \tag{20}$$

The resulting mean-squared quantization error is $\epsilon^2 = E((c - \tilde{c})^2)$ and this can be minimized by using a pdf-optimized quantizer design. An approximate design criterion

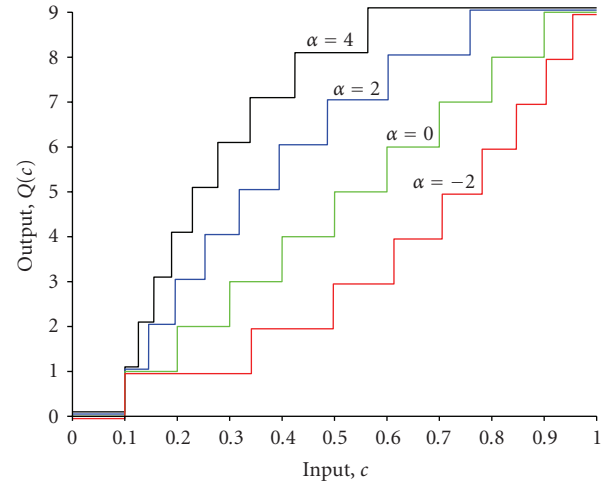


Figure 6: Example quantizer function for positive inputs, $\alpha = -2, 0, 2$, and 4, $\Delta_{dz} = 0.1$, and $N = 9$. (Small vertical offsets have been added for clarity.)

is that the quantizer cell widths $w(c)$ are proportional to $f_c^{-1/3}(c)$ where $f_c(\cdot)$ is the pdf for the coefficients to be quantized (see e.g., [30] or [31]). Under this design criterion, areas with lower probability densities are assigned wider quantization cells. This design criterion becomes exact (minimizing ϵ^2) in the high-rate (large N) limit. For the Laplace pdf (16), the $f_c^{-1/3}$ rule dictates the cell width relationship:

$$w(c) \sim e^{|c|(\sqrt{2}/3\sigma)}. \tag{21}$$

The local quantizer cell widths defined in (17) and (18) are driven by the reciprocal of the local slope of the compander function $F_\alpha(\cdot)$:

$$\left(\frac{\partial}{\partial c}F_\alpha(c)\right)^{-1} = \frac{(1 - e^{-\alpha})e^{c\alpha}}{\alpha}, \tag{22}$$

resulting in the cell width relationship:

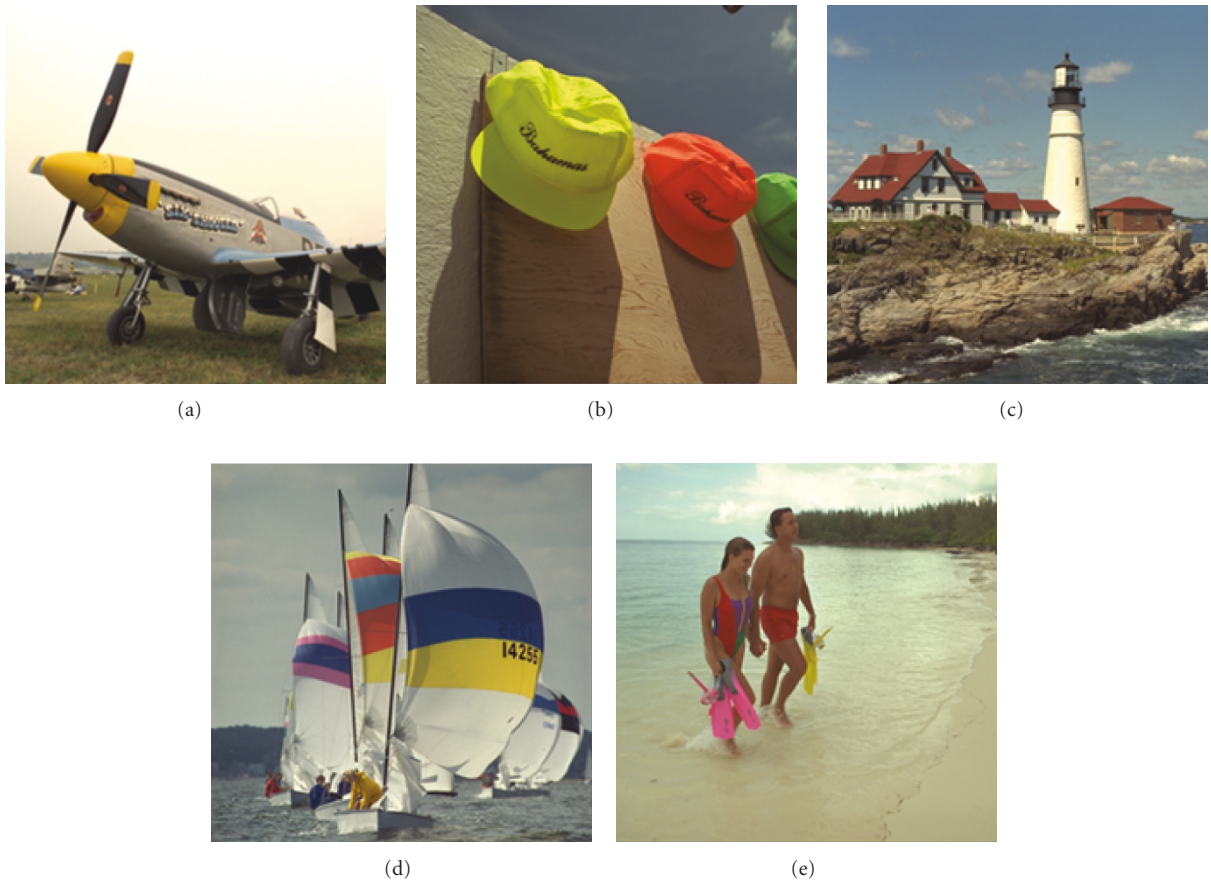
$$w(c) \sim e^{|c|\alpha}. \tag{23}$$

Comparison of (21) with (23) reveals that the choice

$$\alpha = \alpha_0 = \frac{\sqrt{2}}{3\sigma} \tag{24}$$

will give the Laplace pdf-optimized shape to the quantizer defined in (17)-(18). In (24) σ is the standard deviation of the DWT coefficients after scaling to the range $[-1, 1]$.

Thus (17) and (18) define a quantizer parametrized by dead-zone (Δ_{dz}), shape (α), and size (N). Together these three parameters determine the rate and the distortion of the quantizer. Because dead-zone and shape interact in determination of both rate and distortion, they must be optimized jointly. We use the GAST algorithm to find jointly optimal values of Δ_{dz} and α for a fixed quantizer bit rate. And the optimization is with respect to perceived image quality



(a) (b) (c)

(d) (e)

Figure 7: The five images used in the image quality experiment. Original images with dimensions larger than 512 × 512 were cropped as shown.

rather than mean-squared error or some visually weighted variant of mean-squared error.

By convention, GAST parameters range from 0 to 1. Preliminary visual inspection motivated us to apply the mapping

$$p_1 = 12\Delta_{dz} \tag{25}$$

to search Δ_{dz} values from 0 up to 1/12 (DWT coefficients normalized to $[-1, 1]$). Similarly

$$p_2 = 0.5 + 0.5\frac{\alpha}{1.5\alpha_0} \tag{26}$$

allows a search of α values from $-1.5\alpha_0$ to $1.5\alpha_0$. Under this mapping $p_2 = 0.5$ gives the uniform quantizer, and $p_2 = 5/6 \approx 0.83$ gives the pdf-optimized quantizer of (24). For any pair (p_1, p_2) the GAST software calculates and applies the corresponding values of Δ_{dz} and α as given in (25) and (26). This is done for $N = 1, 2, 3 \cdots$ until the entropy of the quantized coefficients approximately matches the target quantizer bit rate.

The target rates are 1.5 or 2.0 bits/coefficient. One of these values was selected for each image in the experiment after preliminary visual inspections. The goal of this manual rate-selection process was to ensure an image quality gradient on the parameter space for each image rather than image quality that is saturated at "very bad" or "very good" due to images that are hard to code or easy to code (or equivalently a target rate that is too low or too high).

Part 1 of JPEG 2000 standard specifies a uniform scalar quantizer ($\alpha = 0$, and quantizer cell width is Δ_q) and a dead-zone that is twice as wide as the other quantizer cells ($\Delta_{dz} = \Delta_q$). Part 2 allows for arbitrary dead-zone widths, but this can interfere with the intrinsic embedding property that follows from the constraint $\Delta_{dz} = \Delta_q$.

The work of [22] reports that rate-distortion optimized dead-zone widths follow $(1/2)\Delta_q < \Delta_{dz} < \Delta_q$. The work of [32] suggests the value $\Delta_{dz} \approx (3/4)\Delta_q$. And [33] proposes $\Delta_{dz} \sim 1/C_{95}$ where C_{95} is the 95th percentile point of the coefficient distribution.

These quantizers are special cases of the more general quantizer described by (17) and (18). In Section 3.2.3 we compare three of these with the visually optimal quantizer designs identified by GAST.

3.2.2. Image Quality Protocol. Five 512 × 512 images were used in the test. These were provided by other image processing labs and were in some cases cropped to obtain this size. Thumbnails of the images can be seen in Figure 7.

In each trial two versions of an image (corresponding to quantization based on two points in the parameter space)

were presented side-by-side on an LCD touch-screen. The prompt "Which image has higher quality?" appeared at the top of the screen, and subjects could select either image by touching the button below it, or they could touch a button labeled "No Quality Difference." This produced scores of ± 1 to indicate an image preference, and 0 for no preference.

A 150 cm by 75 cm table was placed in the center of a sound-isolated room. A 54.5 cm color touchscreen monitor was placed 14 cm from and bisecting the long edge of the table nearest the room's entrance. The monitor has a pixel density of approximately 40 pixels per centimeter (1920×1080 pixels, 47.5 cm by 27 cm). A comfortable chair was placed near the monitor.

The lighting level was controlled by a dimmer. In order to comply with the lighting levels specified in [3], the viewing distance must be known. Viewers were given the freedom to choose their viewing distance; so a lookup table was created. The lookup table included ranges from 27.5 to 67.5 cm (or 1100 to 2700 pixels) in increments of 5 cm (200 pixels). A digital lux meter was used to measure the illuminance of the monitor and the wall behind the monitor at a given distance. These readings iteratively served as a guide to correct the position of the dimmer for each viewing distance. These dimmer positions were recorded and linked to viewing distance in the lookup table.

Viewers were instructed to adjust the chair's distance from the monitor; so they could comfortably compare detailed images. After viewers selected a comfortable position, the approximate viewing distance was measured and the lookup table was consulted to find the proper dimmer setting.

Two of the room's four adjustable lights were positioned such that a semiuniform field of light illuminated the gray background wall. The other two lights were pointed towards the side walls. Very little light ended up on the wall behind the subject, thus minimizing reflections on the surface of the monitor.

Color calibration hardware and software was used to optimize the monitor's color profile for accuracy and optimal contrast given the room's lighting environment.

Twenty subjects participated in the experiment. Subjects wore any vision correction that they normally would for screen-based work at their preferred viewing distance. Each subject ran the GAST algorithm on all five images, using two randomly selected starting locations in the parameter space for each image. Thus, each subject performed ten GAST tasks, and in each trial the subject made one step of progress on one task randomly selected from the ten. A total of 199 GAST trials were completed. This falls short of $20 \times 10 = 200$ due to time limitations for one subject. Subjects typically spent around 35 minutes in the test. We used the direction-finding step size $\Delta_d = 0.20$ and the terminating condition $\Delta_t = 0.15$.

3.2.3. Image Quality Results. Figure 8 shows the starting and ending points for the 199 completed GAST trials. The starting points are randomly distributed across the search space, and the ending points are mostly clustered near the center of the search space. Some ending points remain close to or identical to their starting points. This indicates a lack of local quality gradient. Indeed, in the corners of the search space, the image quality is consistently low—there is no local quality gradient. In addition, some random starting places happen to fall near the point of maximum image quality and those trials end quickly.

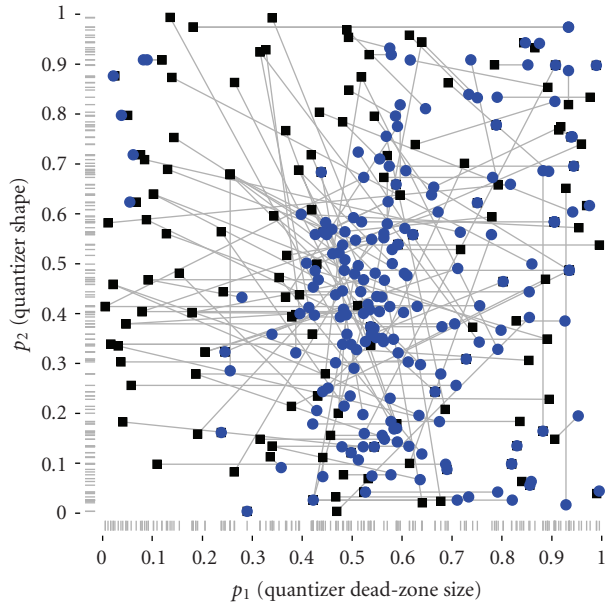


Figure 8: Starting and ending points for all 199 completed image GAST tasks shown by black squares and blue circles, respectively. The gray tick marks in the axes indicate the p_1 or p_2 value of each starting point.

Table 1: Means and 95% Confidence Interval Values of p_1 and p_2 for each Image.

Image	Mean		95% c.i.	
	p_1	p_2	p_1	p_2
a	0.599	0.430	0.060	0.086
b	0.588	0.519	0.073	0.064
c	0.631	0.413	0.046	0.082
d	0.636	0.469	0.048	0.079
e	0.576	0.423	0.073	0.071
all	0.606	0.451	0.027	0.034

Figure 9 shows the ending points for the 199 trials coded by image and Figure 10 shows the mean ending point for each image with a cross. The major and minor axes of the ellipse drawn about each cross indicate the 95% confidence interval for that mean location. While some per-image differences are observable in these results, they are not large, especially in light of the confidence intervals. Table 1 shows the numerical results for each image.

Figure 11 shows the grand mean result and 95% confidence intervals for image GAST experiments taken over the five images. In addition, the mean (across the five images) locations of three different reference quantizers described in Section 3.2.1 are displayed. These quantizers all use uniform

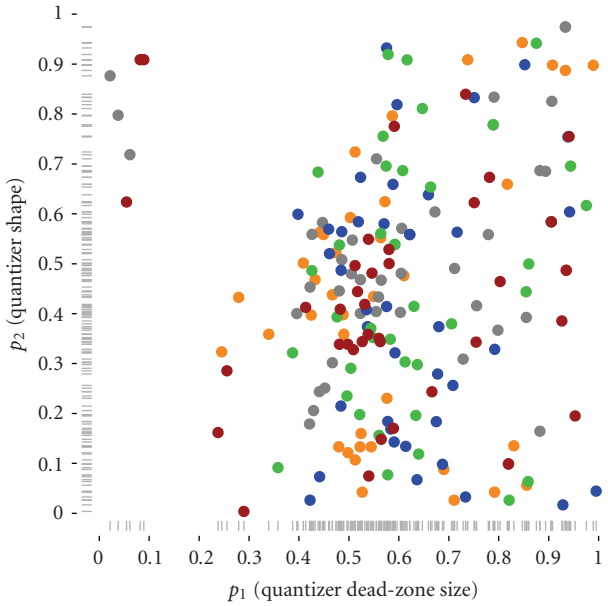


Figure 9: Ending points for all completed image GAST tasks. Red dots correlate with image a, gray with image b, blue with image c, green with image d, and orange with image e. The gray tick marks in the axes indicate the p_1 or p_2 value of each ending point.

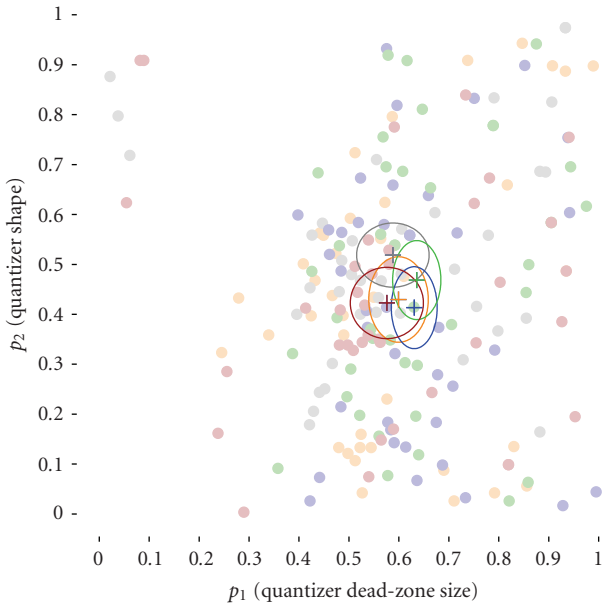


Figure 10: Mean and 95% confidence intervals for all completed image GAST tasks, separated by image. Similarly to Figure 9, the red ellipse correlates with image A, gray with image B, blue with image C, green with image D, and orange with image E.

quantization bins $p_2 = 0.5$ ($\alpha = 0.0$) with the possible exception of the central bin defined by the dead-zone. Thus, they differ only with respect to p_1 which controls Δ_{dz}. These quantizers are (from left to right) the uniform rounding quantizer ($\Delta_{dz} = 1/2\Delta_q$), the quantizer proposed in [32] ($\Delta_{dz} = (3/4)\Delta_q$), and the JPEG 2000 Part 1 quantizer

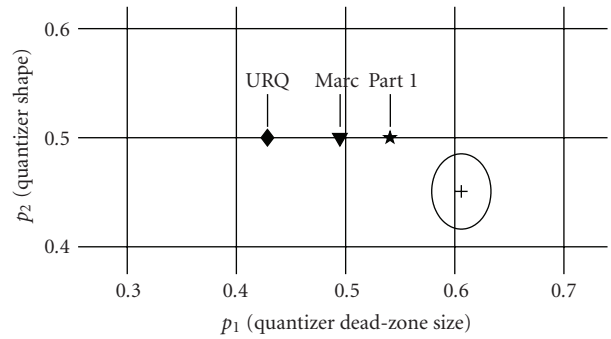


Figure 11: Mean and 95% confidence intervals for all completed image GAST tasks. Compare with three labeled quantizer designs, left to right: Uniform Rounding Quantizer (URQ), Quantizer of [32] (Marc), JPEG 2000, Part 1 Quantizer (Part 1).

($\Delta_{dz} = \Delta_q$). Of these three options, the GAST results are closest to the JPEG 2000 Part 1 quantizer, though in context of this particular experiment, a slightly larger dead-zone ($p_1 = 0.61$, $\Delta_{dz} = 0.051$) may be desirable.

Next, we consider the quantizer shape parameter α which is controlled by P_2. To minimize MSE, one would select a pdf-optimized quantizer, $\alpha = \alpha_0$ and $p_2 = 5/6$. This is a *compressive* function and quantizer bins get larger as one moves away from zero. Consideration of visual self-masking suggests the function, $H(x) = x^{0.7}$ [29]. While not directly comparable with (18), this is also a *compressive* function and thus would correspond to $0 < \alpha$ and $0.5 < p_2$. But the GAST results say that image quality is maximized, on average, by $p_2 = 0.45$, corresponding to a slightly *negative* shape factor ($\alpha = -0.15\alpha_0$) and a slightly *expansive* function, with bins getting slightly smaller as one moves away from zero.

This quantizer shape result is barely statistically significantly different from $p_2 = 0.5$ and $\alpha = 0.0$, which would point to uniform quantization as the optimal strategy, and that may be the safest conclusion to draw. Suffice it to say that this experiment does not suggest the use of a compressive nonlinearity to improve image quality.

The experiment results can be summarized as follows. When the quantizer defined by (17)-(18) is applied to the Y-plane, level 4, LH/HL orientation, Daubechies 9/7 DWT coefficients from the five images shown in Figure 7, the dead-zone size and quantizer shape that maximize mean perceived image quality are very close to the dead-zone and shape used in JPEG 2000, Part 1. From an image coding perspective, we may have simply reinvented the wheel. Or we could argue that we have added additional, and unique, support for the JPEG 2000 Part 1 quantizer design. But from the image quality assessment perspective, we argue that we have demonstrated a new subjective image quality maximization technique that has surveyed a two-dimensional image coding space and efficiently arrived at what is arguably the "right answer."

4. Discussion and Observations

We have presented the motivation for and development of GAST. And we have demonstrated this novel and efficient

subjective testing technique in two different domains: audio quality testing and image quality testing.

In the audio experiment we created a simple controlled, two-dimensional parameter space using reference conditions. Because of the already established monotonic relationships between Q and perceived audio quality, and between T and perceived audio quality, the region of highest audio quality (the "right answer") was known. This is a necessary condition for the evaluation of a new measurement technique. The GAST algorithm identified the right answer accurately and efficiently. Compared with the hypothetical comparable ES ACR subjective test, the number of votes was reduced by at least a factor of 27, and one would expect these savings to increase in higher-dimensional problems.

In the image experiment we optimized the dead-zone size and shape of a quantizer for one class of JPEG 2000 DWT coefficients. Here the "right answer" was not known—this is a natural next step for testing a measurement technique. GAST identified a dead-zone size and quantizer shape that maximize image quality and these are quite close to those defined in JPEG 2000, Part 1. We consider this to be a very plausible "right answer."

We emphasize again that a successful GAST task identifies a local quality maximum. As with all such search algorithms, there is no guarantee that this local maximum is the global maximum. And as with all such search algorithms, there is a battery of techniques to mitigate this potential problem. Our work here demonstrates one of the simplest of these techniques, the use of multiple random starting points. When the vast majority of searches starting from across the search space end up in the same region, one can have good confidence that the region is preferred in the global and local sense.

Note that GAST can be used with naïve or expert subjects. Expert subjects might benefit from additional information as the test progresses. Since the end point of each line search is the current approximation to the point of maximal quality, experts might advantageously use feedback on search progress to use their time even more efficiently. For example, the message "You have just completed the nth line search for this task" indicates that one has obtained an approximate solution and could end the task despite the fact that a terminating condition has not been met.

Note also that if identifying points of *minimal* quality is of interest (worse case analyses), one can simply multiply all votes by -1 and the GAST algorithm will locate minima instead of maxima.

The work presented here is a fairly straightforward melding of paired-comparison subjective testing and a rather basic search algorithm. There are many potential paths to improve GAST performance, efficiency, and robustness. One might undertake refinement of the terminating conditions, possibly making them adaptive. Line lengths could become adaptive; thus one would search shorter lines as the algorithm progresses, since the start of the line should be getting closer to the sought-after point of maximal quality. The direction finding step size Δ_d might be advantageously adapted as the algorithm progresses (larger early on or when, in flatter regions, smaller later or in steeper regions). Finally, other search algorithms could be employed in a similar fashion.

Acknowledgments

This work was supported by the National Telecommunications and Information Administration's Institute for Telecommunication Sciences. The authors would like to thank Frank Sanders for his managerial support and the many anonymous test subjects who participated in the subjective experiments.

References

[1] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," Geneva, 1996.

[2] ITU-R Recommendation BS.1284, "General methods for the subjective assessment of sound quality," Geneva, 2003.

[3] ITU-R Recommendation BT.500-12, "Methodology for the subjective assessment of the quality of television pictures," Geneva, 2009.

[4] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia applications," Geneva, 1998.

[5] S. Tourancheau, F. Autrusseau, Z. M. P. Sazzad, and Y. Horita, "Impact of subjective dataset on the performance of image quality metrics," in *Proceedings of the 15th IEEE International Conference on Image Processing (ICIP '08)*, pp. 365–368, October 2008.

[6] S. Voran, "Estimation of speech intelligibility and quality," in *Handbook of Signal Processing in Acoustics*, D. Havelock, S. Kuwano, and M. Vorläander, Eds., vol. 2, chapter 28, pp. 483–520, Springer, New York, NY, USA, 2008.

[7] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[8] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[9] S. Voran and A. Catellier, "Gradient ascent paired-comparison subjective quality testing," in *Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEx '09)*, pp. 133–138, San Diego, Calif, USA, July 2009.

[10] I. E. G. Richardson and C. S. Kannangara, "Fast subjective video quality measurement with user feedback," *Electronics Letters*, vol. 40, no. 13, pp. 799–801, 2004.

[11] U. Reiter and J. Korhonen, "Comparing apples and oranges: subjective quality assessment of streamed video with different types of distortion," in *Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEx '09)*, pp. 127–132, San Diego, Calif, USA, July 2009.

[12] A. B. Watson and D. G. Pelli, "QUEST: a Bayesian adaptive psychometric method," *Perception and Psychophysics*, vol. 33, no. 2, pp. 113–120, 1983.

[13] A. Ravindran, K. M. Ragsdell, and G. V. Reklaitis, *Engineering Optimization: Methods and Applications*, Wiley, Hoboken, NJ, USA, 2nd edition, 2006.

[14] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, NY, USA, 2nd edition, 2006.

[15] S. Boyd, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[16] B. Gottfried and J. Weisman, *Introduction to optimization theory*, Prentice Hall, Englewood Cliffs, NJ, USA, 1973.

[17] ITU-T Recommendation P.810, "Modulated noise reference unit (MNRU)," Geneva, 1996.

[18] B. Cotton, "New reference condition for very low bit rate voice coder evaluation," CCITT SGXII Contribution D.108, September 1991.

[19] S. Voran, "Observations on the t-reference condition for speech coder evaluation," CCITT SGXII Contribution SQ-13.92, February 1992, http://www.its.bldrdoc.gov/audio.

[20] ISO/IEC 15444-1, ITU-T T.800, "Information Technology—JPEG 2000 image coding system," Geneva, 2004.

[21] ISO/IEC 15444-2, ITU-T T.801, "Information Technology—JPEG 2000 image coding system: extensions," Geneva, 2004.

[22] P. Schelkens, A. A. Skodras, and T. Ebrahimi, Eds., *The JPEG 2000 Suite*, Wiley, Chichester, UK, 2009.

[23] ISO/IEC IS 10918-1, ITU-T T.81, "Information Technology—Digital compression and coding of continuous-tone still images—part 1: Requirements and guidelines," Geneva, 1993.

[24] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1164–1175, 1997.

[25] M. Long, H. M. Tai, and S. Yang, "Quantisation step selection schemes in JPEG2000," *Electronics Letters*, vol. 38, no. 12, pp. 547–549, 2002.

[26] D. M. Chandler and S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," *Journal of the Optical Society of America A*, vol. 20, no. 7, pp. 1164–1180, 2003.

[27] Z. Liu, L. J. Karam, and A. B. Watson, "JPEG2000 encoding with perceptual distortion control," *IEEE Transactions on Image Processing*, vol. 15, no. 7, pp. 1763–1778, 2006.

[28] H. Oh, A. Bilgin, and M. W. Marcellin, "Visibility thresholds for quantization distortion in JPEG2000," in *Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEx '09)*, pp. 228–232, San Diego, Calif, USA, July 2009.

[29] W. Zeng, S. Daly, and S. Lei, "An overview of the visual optimization tools in JPEG 2000," *Signal Processing: Image Communication*, vol. 17, no. 1, pp. 85–104, 2002.

[30] N. Judell and L. Scharf, "A simple derivation of Lloyd's classical result for the optimum scalar quantizer (corresp.)," *IEEE Transactions on Information Theory*, vol. 32, no. 2, pp. 326–328, 1986.

[31] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Prentice Hall, London, UK, 1984.

[32] M. W. Marcellin, M. A. Lepley, A. Bilgin, T. J. Flohr, T. T. Chinen, and J. H. Kasner, "An overview of quantization in JPEG 2000," *Signal Processing: Image Communication*, vol. 17, no. 1, pp. 73–84, 2002.

[33] A. O. Zaid, C. Olivier, and F. Marmotton, "Wavelet image coding with adaptive dead-zone selection: application to JPEG 2000," in *Proceedings of the International Conference on Image Processing (ICIP '02)*, vol. 3, pp. 253–256, Rochester, NY, USA, September 2002.