Using Human and Model Performance to Compare MRI Reconstructions

M. Dylan Tisdall* and M. Stella Atkins

***Abstract*—Magnetic resonance imaging (MRI) reconstruction techniques are often validated with signal-to-noise ratio (SNR), contrast-to-noise ratio, and mean-to-standard-deviation ratio measured on example images. We present human and model observers as a novel approach to evaluating reconstructions for low-SNR magnetic resonance (MR) images. We measured human and channelized Hotelling observers in a two-alternative forced-choice signal-known-exactly detection task on synthetic MR images. We compared three reconstructions: magnitude, wavelet-based denoising, and phase-corrected real. Human observers performed approximately equally using all three reconstructions. The model observer showed very close agreement with the humans over the range of images. These results contradict previous predictions in the literature based on SNR. Thus, we propose that human observer studies are important for validating MRI reconstructions. The model's performance indicates that it may provide an alternative to human studies.**

***Index Terms*—Denoising, magnetic resonance imaging (MRI), observers, signal detection, signal processing.**

I. Introduction

IMAGES acquired using magnetic resonance imaging (MRI) are initially complex-valued and corrupted with complex additive white Gaussian noise (AWGN), mostly due to thermal noise in the patient [1]. In order to display these images, each complex pixel is first reduced by some reconstruction operation to a real value that can be displayed as a greyscale intensity. Since the signal-to-noise ratio (SNR) of the images is often quite low, many reconstruction techniques have been proposed that include filters for noise reduction [2]–[4]. However, in the literature presenting and comparing these algorithms the quality of the output is usually represented in terms of summary statistics such as SNR, contrast-to-noise ratio (CNR), or mean-to-standard-deviation ratio (MSR) calculated over a set of example images [2]–[8]. Noting that the vast majority of magnetic resonance (MR) images will be viewed by radiologists for diagnosis, we propose that observer performance is a more relevant quality metric for validating reconstruction techniques.

There is a large body of literature using human observers and/or mathematical models of human response on simple detection tasks using X-ray [9], [10] and nuclear-medicine images [11]. It has been suggested that these models can assist in the selection of imaging parameters and comparison of different equipment. Additionally, there has been work using similar studies to determine the correct parameters for lossy image compression algorithms [10]. We take a similar approach, using both human and model observers to compare three MRI reconstruction algorithms. In particular, we used high-SNR MR images as backgrounds combined with complex AWGN to produce our simulated low-SNR raw MRI data and then reconstructed it using three techniques. Our human observers were volunteers without radiological experience and we compared their performance against a channelized Hotelling observer (CHO) model [12], [13] in a two-alternative forced-choice (2AFC) signal-known-exactly (SKE) detection task.

In Section II, we present a simple model of the complex MRI signal and describe the three reconstruction techniques used. In Section III, we present the process used to generate the synthetic data. Section IV describes the experiment performed by the human observers and describes the CHO model used for comparison. The results of our experiments are presented and discussed in Section V. Section VI contains our conclusion and suggestions for future work.

II. MRI Reconstruction

The image acquired from an MRI scanner is initially complex-valued and, for our purposes, can be described by the equation

$$\mathbf{Y}[\boldsymbol{x}] = \mathbf{S}[\boldsymbol{x}] \exp\left(\mathrm{i}\boldsymbol{\Theta}[\boldsymbol{x}]\right) + \mathbf{N_r}[\boldsymbol{x}] + \mathrm{i}\mathbf{N_i}[\boldsymbol{x}] \tag{1}$$

where $\boldsymbol{x}$ is the two-dimensional (2-D) index of a pixel, $\mathbf{Y}$ is the matrix of complex-valued image pixels, $\mathbf{S}$ is the matrix of real-valued signals, $\boldsymbol{\Theta}$ is the matrix of signal phases, and $\mathbf{N_r}$ and $\mathbf{N_i}$ are matrices of samples from a normal distribution $\mathcal{N}(0, \sigma)$ representing thermal noise in the real and imaginary components, respectively. An example of a single pixel in this model is presented in Fig. 1. It is important to note that $\boldsymbol{\Theta}$ is not the phase of the recorded pixel, but the phase of the signal component in the recorded pixel. The complex AWGN represented with $\mathbf{N_r}$ and $\mathbf{N_i}$ will alter the phase of the recorded pixel. For a more detailed model of MRI signal and noise, see Macovski [14].

In order to display the image in greyscale, a real-valued matrix of pixels $\mathbf{Y}'$ must be produced. Ideally, we would display the real-valued $\mathbf{S}$, but since we only know $\mathbf{Y}$, we attempt to calculate and display a real-valued $\mathbf{Y}'$ that is as close as possible to the unknown $\mathbf{S}$. We will represent an MRI reconstruction for an image with n pixels as a mapping $f : \mathbb{C}^n \rightarrow \mathbb{R}^n$ that gives $\mathbf{Y}' = f(\mathbf{Y})$.

Manuscript received February 1, 2006; revised May 1, 2006. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). *Asterisk indicates corresponding author.*

*M. D. Tisdall was with the School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada (email: mtisdall@cs.sfu.ca).

M. S. Atkins was with the School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada.

Digital Object Identifier 10.1109/TMI.2006.881374

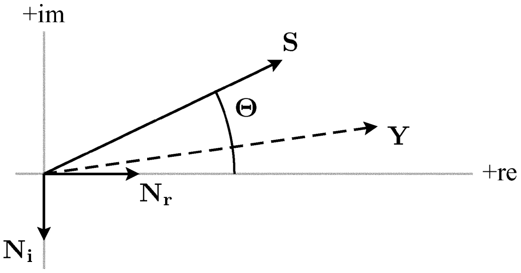

Fig. 1. Example of an MRI pixel showing the relationship of the various components in (1). Solid black lines represent the three components of our signal model and the indicated angle is the signal phase. Dashed line shows the sum of the components. This dashed line is the recorded value of the pixel while all the other parts of the diagram are unknown.

A. Magnitude Reconstruction

The simplest and most common reconstruction algorithm is the magnitude transform f_m [15]. In this mapping, each pixel's real value is set to the magnitude of the complex pixel. If $\mathbf{Y}$ is the complex vector of recorded data, $\mathbf{Y}'_m$ is the real image matrix resulting from the magnitude transform, and $\overline{\mathbf{Y}[\boldsymbol{x}]}$ is used to denote the complex conjugate of $\mathbf{Y}[\boldsymbol{x}]$, the magnitude reconstruction of pixels is written

$$\mathbf{Y}'_m[\boldsymbol{x}] = \sqrt{\mathbf{Y}[\boldsymbol{x}]\overline{\mathbf{Y}[\boldsymbol{x}]}}. \tag{2}$$

This approach benefits from discarding the effects of the signal phase $\boldsymbol{\Theta}$. However, the principle difficulty with the magnitude transform is that the complex AWGN in $\mathbf{Y}$ is transformed to a Rician noise in $\mathbf{Y}'_m$ with the attendant problem of bias in low-signal areas reducing image contrast [16].

B. Wavelet Reconstruction

In order to enhance contrast and edges, a wavelet transform can be applied and particular wavelet coefficients thresholded. MRI reconstruction algorithms involving wavelets can be applied to the real and imaginary components of $\mathbf{Y}$ before applying a magnitude transform [4] or applied to the real-valued image after the magnitude transform [2]. Algorithms using the wavelet-then-magnitude format will smooth Gaussian noise in each component, but may be affected by the phase of the signal. The alternative, magnitude-then-wavelet approach has the advantage of working on data from which the phase has been discarded, but must be tailored to Rician noise.

For the purposes of our experiment, we will use the magnitude-then-wavelet algorithm and denote it f_w. This reconstruction relies on the magnitude transform f_m and a wavelet filter, $g : \mathbb{R}^n \to \mathbb{R}^n$, that are combined as

$$f_w = g \circ f_m. \tag{3}$$

We used a wavelet filter developed specifically for Rician noise by Nowak [2] as our implementation of g. In particular, we implemented Nowak's algorithm using the full-scale Harr wavelet transform without any shift-invariant approximations. We chose this filter because it is often cited in the MRI noise reduction literature as a point of comparison for new techniques.

C. Phase-Corrected Real Reconstruction

The phase-corrected real reconstruction f_p is an alternative to using the magnitude transform either alone or with wavelets [5]–[8]. This reconstruction first estimates $\boldsymbol{\Theta}$ with $\tilde{\boldsymbol{\Theta}}$ and then performs a point-by-point multiplication of $\mathbf{Y}$ and $\exp(-\mathrm{i}\tilde{\boldsymbol{\Theta}})$. Performing this multiplication on (1) gives

$$\mathbf{Y}[\boldsymbol{x}]\exp\left(-\mathrm{i}\tilde{\boldsymbol{\Theta}}[\boldsymbol{x}]\right) = \mathbf{S}[\boldsymbol{x}]\exp\left(\mathrm{i}\left(\boldsymbol{\Theta}[\boldsymbol{x}] - \tilde{\boldsymbol{\Theta}}[\boldsymbol{x}]\right)\right) + (\mathbf{N_r}[\boldsymbol{x}] + \mathrm{i}\mathbf{N_i}[\boldsymbol{x}])\exp\left(-\mathrm{i}\tilde{\boldsymbol{\Theta}}[\boldsymbol{x}]\right). \tag{4}$$

Assuming that $\tilde{\boldsymbol{\Theta}} \simeq \boldsymbol{\Theta}$, the signal phases cancel giving $\mathbf{S}[\boldsymbol{x}]\exp(\mathrm{i}(\boldsymbol{\Theta}[\boldsymbol{x}] - \tilde{\boldsymbol{\Theta}}[\boldsymbol{x}]) \simeq \mathbf{S}[\boldsymbol{x}]$. Additionally, rotating the complex AWGN composed of $\mathbf{N_r}$ and $\mathbf{N_i}$ has no effect on the distribution. Although the noise samples in any recorded pixel will be rotated by the multiplication in (4), the underlying distribution is not affected by the rotation. Noting this, we can simplify our pixel model to

$$\mathbf{Y}[\boldsymbol{x}]\exp\left(-\mathrm{i}\tilde{\boldsymbol{\Theta}}[\boldsymbol{x}]\right) \simeq \mathbf{S}[\boldsymbol{x}] + \mathbf{N_r}[\boldsymbol{x}] + \mathrm{i}\mathbf{N_i}[\boldsymbol{x}]. \tag{5}$$

Taking just the real component of $\mathbf{Y}[\boldsymbol{x}]\exp(-\mathrm{i}\tilde{\boldsymbol{\Theta}}[\boldsymbol{x}])$ gives our final definition for $\mathbf{Y}'_p[\boldsymbol{x}]$, the phase-corrected real reconstructed image

$$\mathbf{Y}'_p[\boldsymbol{x}] = \mathrm{Re}\left(\mathbf{Y}[\boldsymbol{x}]\exp\left(-\mathrm{i}\tilde{\boldsymbol{\Theta}}[\boldsymbol{x}]\right)\right) \simeq \mathbf{S}[\boldsymbol{x}] + \mathbf{N_r}[\boldsymbol{x}]. \tag{6}$$

This shows that, assuming we have estimated $\boldsymbol{\Theta}$ closely, the result of a phase-corrected real reconstruction is an image containing the signal and a real AWGN.

It is important to note that while $\mathbf{Y}'_p$ is an unbiased estimator of $\mathbf{S}$, when $\mathbf{Y}'_p$ is displayed on a monitor the result is that dark regions will be biased positively simply due to the fact that a monitor cannot display negative intensities. However, the displayed bias in $\mathbf{Y}'_p$ is less than that in $\mathbf{Y}'_m$ and so $\mathbf{Y}'_p$ will be closer to $\mathbf{S}$ than $\mathbf{Y}'_m$.

III. Synthetic Images

When attempting to locate a target feature in an MR image, there are two major sources of distraction: nontarget patient anatomy and thermal noise. In terms of our signal model (1) the anatomy and target feature combine to form the image signal $\mathbf{S}$ while the thermal noise is the complex AWGN process $\mathbf{N_r} + \mathrm{i}\mathbf{N_i}$. By producing synthetic images with both of these components, we propose that target feature detection in our synthetic images will have approximately the same results as similar tasks in clinical MR images when using the reconstructions described in Section II.

We create a complex-valued image without a target feature via

$$\mathbf{Y} = \mathbf{B} + \mathbf{N_r} + \mathrm{i}\mathbf{N_i} \tag{7}$$

where $\mathbf{B}$ is a real-valued background anatomy image, and $\mathbf{N_r}$ and $\mathbf{N_i}$ are images containing real-valued AWGN. Similarly, we produce a complex-valued image with a target feature using

$$\mathbf{Y} = \mathbf{T} + \mathbf{B} + \mathbf{N_r} + \mathrm{i}\mathbf{N_i} \tag{8}$$

where $\mathbf{T}$ is the real-valued image containing only the target.

Equations (7) and (8) do not include signal phases. This is a substantial deviation from the model in (1), but can be be justified by considering the processing that will be applied to these images. The result of the magnitude transform on either the feature-present or feature-absent images defined above will be invariant to signal phase since (2) discards phase. Similarly, because the wavelet transform being considered operates on images after the magnitude transform, its result is also correct without needing a simulated signal phase.

The phase-corrected real reconstruction does rely on an estimator of the signal phase and thus potentially on a simulated signal phase. However, we note that in the best case, $\tilde{\boldsymbol{\Theta}}$ is exactly the same as $\boldsymbol{\Theta}$ and so we can simulate the best case result of phase correction with

$$\mathbf{Y}'_p[\boldsymbol{x}] = \mathrm{Re}\left(\mathbf{Y}[\boldsymbol{x}]\right) = \mathbf{B} + \mathbf{N}_r \tag{9}$$

in the target-feature-absent case and

$$\mathbf{Y}'_p[\boldsymbol{x}] = \mathrm{Re}\left(\mathbf{Y}[\boldsymbol{x}]\right) = \mathbf{T} + \mathbf{B} + \mathbf{N}_r \tag{10}$$

in the target-feature-present case.

We also note that if a phase estimation scheme over-fits the data (i.e., if the phase due to noise is fit, instead of just the signal phase) the result will be an approximation to the magnitude transform. Alternatively, if the phase estimation under-fits the data, there will be spatially varying signal intensity but the noise power will remain the same. This will have the same effect as lowering the target feature intensity relative to the thermal noise. Thus, while we do not simulate the failure conditions directly in our experiment, the effects of both these types of failures will be discernible from the experiment results because we cover both the magnitude transform and a range of target feature intensities and thermal noise powers.

Our target feature $\mathbf{T}$ was an antialiased circular object located in the center of the feature-present image (see Fig. 3). Each pixel in this image matrix was given by

$$\mathbf{T}[\boldsymbol{x}] = \begin{cases} b, & \text{if } |\boldsymbol{x} - \boldsymbol{z}|_2 \leq w \\ b\left(1 - |\boldsymbol{x} - \boldsymbol{z}|_2 + w\right), & \text{if } w < |\boldsymbol{x} - \boldsymbol{z}|_2 < 1 + w \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where b is the amplitude of the feature, $\boldsymbol{z}$ is the index of the image center, $|\cdot|_2$ is the two-norm, and w controls the width of the feature. For our experiments, we set $w = 3$ which was equivalent to an anatomical feature with a diameter of 6 mm. We selected $b \in \{(1/20), (1/12), (5/36)\} \simeq \{0.05, 0.083, 0.139\}$ for each image.

The complex-valued thermal noise was simulated by first selecting $\sigma_{\mathbf{N}} \in \{(9/200), (3/40)\} \simeq \{0.045, 0.075\}$. Based on this choice, we randomly generated two 128 × 128 pixel images for each synthetic MRI. Each pixel in these noise images was sampled from $\mathcal{N}(0, \sigma_{\mathbf{N}})$. One of these images was taken as $\mathbf{N_r}$ and the other as $\mathbf{N_i}$.

To simulate distracting anatomy $\mathbf{B}$ we used regions of slices from high-SNR MR head images of healthy volunteer. These volunteers were scanned using a three-dimensional (3-D) inversion recovery pulse sequence on a Philips Gyroscan Intera 3.0-T MRI scanner. Each volume was reconstructed using the magnitude transform to give real-valued images. These real-valued volumes were then sliced along the axial, coronal, and sagital directions to produce a library of 2-D images. Each slice was then cropped into 16 separate 128 × 128 pixel images. These images were checked to see if they contained enough anatomy by thresholding the central 64 × 64 pixel region of each image and ensuring more than half of the pixels contained anatomy. Images with sufficient anatomy were normalized so that their pixel intensities were on the range (0,1) and retained for use as backgrounds.

As noted in Section II, the background images produced by cropping and normalization will have Rician noise because the magnitude reconstruction was used to produce the greyscale values. Since we used these images as our real-valued data, this Rician noise will have been added to the Gaussian $\mathbf{N_r}$ and included in all of the synthetic image reconstructions. However, because our anatomical images were scanned at 3 T, the magnitude images, after being normalized to the range (0,1), had a standard deviation of approximately 0.002 measured in regions of air. Since the lowest standard deviation, $\sigma_{\mathbf{N}}$ used for our simulated thermal noise is more than 20 times greater than the inherent noise of our anatomical backgrounds, it is unlikely that the spurious noise included in the anatomical images had any effect on our results.

672 images were produced in each of the 36 possible conditions (three reconstructions, three signal powers, two noise powers, target feature present or absent). In Fig. 2 we show one feature-present anatomical background in all 18 possible conditions. To quantify our synthetic images' quality using the same calculations commonly used to describe clinical MRI, we define the PSNR as the ratio of the peak signal to the Rician noise standard deviation in a region of air [16]

$$\mathrm{PSNR} = \frac{\max\limits_{\boldsymbol{x}} \mathbf{S}[\boldsymbol{x}]}{\sigma_{\mathbf{N}}\sqrt{(2 - \pi/2)}}. \tag{12}$$

Due to our normalizing all the anatomical backgrounds to the range (0,1), we know that $\max\limits_{\boldsymbol{x}} \mathbf{S}[\boldsymbol{x}] = 1$. Thus, for images where $\sigma_{\mathbf{N}} = 9/200$, we calculate $\mathrm{PSNR} \simeq 33.92$ and for $\sigma_{\mathbf{N}} = 3/40$ we have $\mathrm{PSNR} \simeq 20.35$ These are a little low, but not implausible PSNR values for clinical MRI.

IV. Observers

A. Human Observers

Sixteen volunteer observers without any previous radiological training were recruited to participate in our study. The experimental software presented participants with three images

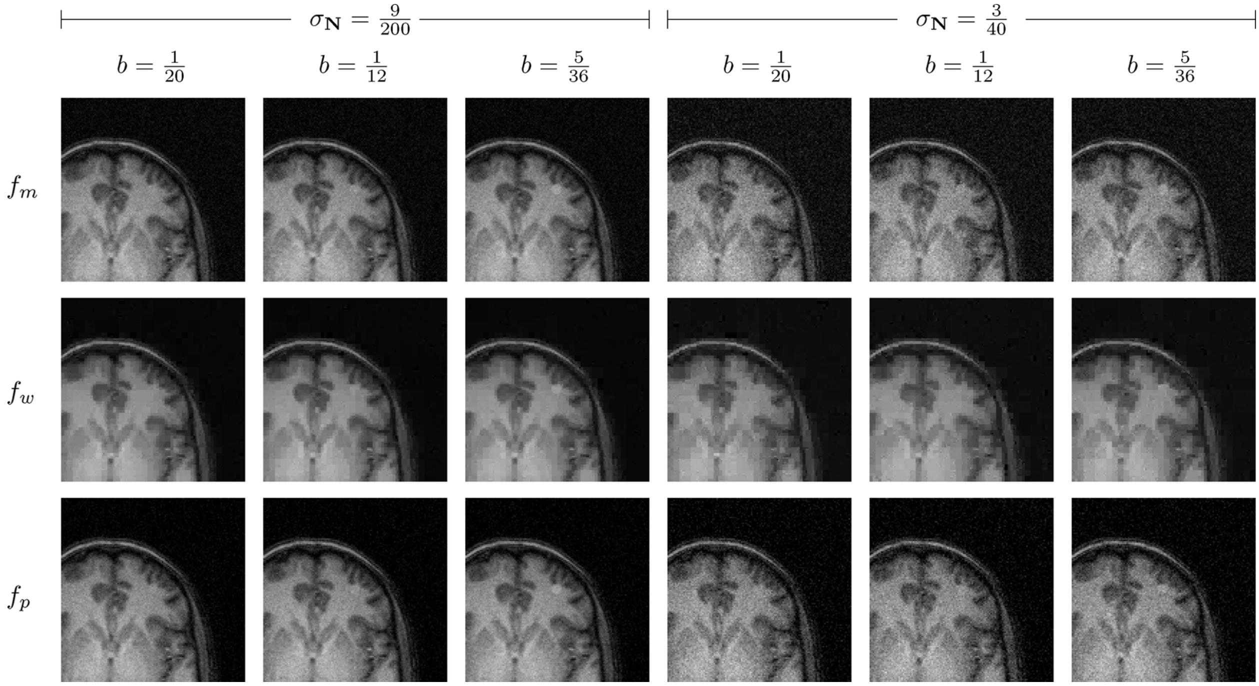

Fig. 2. Examples of feature-present synthetic images for all of the possible 18 experimental conditions. First row was produced with the magnitude reconstruction, the second row with the wavelet reconstruction, and the third row with the phase-corrected real reconstruction. Each column was produced with a different simulated signal intensity, as specified at the top. First three columns were produced using AWGN with $\sigma_{\mathbf{N}} = 9/200$ and the last three with $\sigma_{\mathbf{N}} = 3/40$.

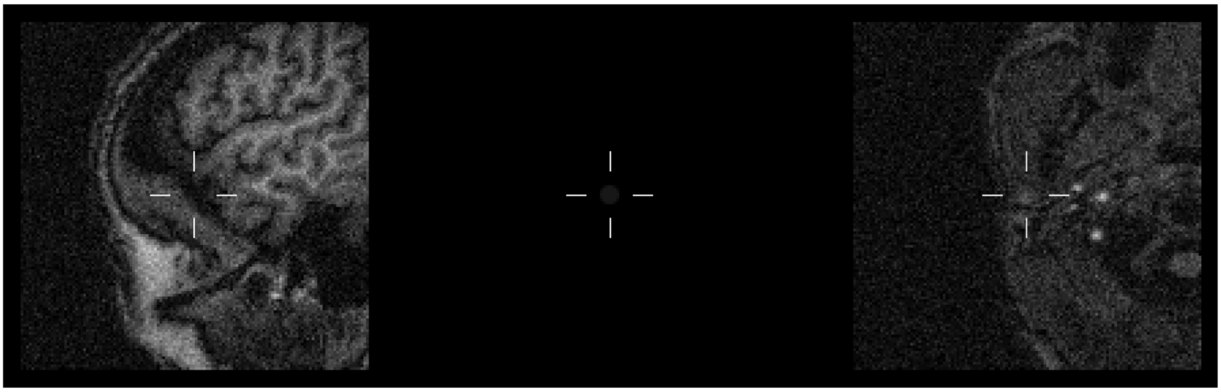

Fig. 3. Example of the user interface used in human observer study. In this case, the target feature is in the right image.

aligned horizontally (see Fig. 3). The center image showed the target feature and the two exterior images represented choices in the 2AFC test. Since this was an SKE task, crosshairs were superimposed over the images in order to reduce the possibility of confusion about the target feature location. The crosshairs could be toggled on and off by the user to reduce visual distraction. Participants were instructed that in every display, one of the exterior images would contain the target feature and that they should use the mouse to click on whichever exterior image they felt most probably contained the target. They were allowed to take as long as they wanted to reach a decision on each image pair. Once a participant clicked on an exterior image, the screen was made completely black for 0.5 s, the mouse pointer was warped to the center of the screen, and then the next set of images was shown and the process repeated.

Each participant was given two training sets, composed equally of all 18 possible combinations of noise power, target feature power, and reconstruction algorithm. If the training took less than 10 min, they were then instructed to wait until ten minutes had elapsed in order to ensure a constant dark adaptation time across all participants. After the training and delay, the participants then proceeded through 16 experiment sets composed equally of the 18 possible combinations for a total of 288 image pairs. In order to minimize order effects, the ordering of the image pairs was selected randomly for each participant from a constant distribution of all possible orderings. As the experiment lasted up to 1 h, fatigue was reduced by displaying a black screen after every 18 image pairs and instructing users to take as long a break as they desired while the display was dark.

The experiments were conducted in a completely darkened room using a CRT monitor (SGI CMNB024B) as a display. The monitor's spectrum was measured in both the left and right image centers using a telephotometer (Photo Research

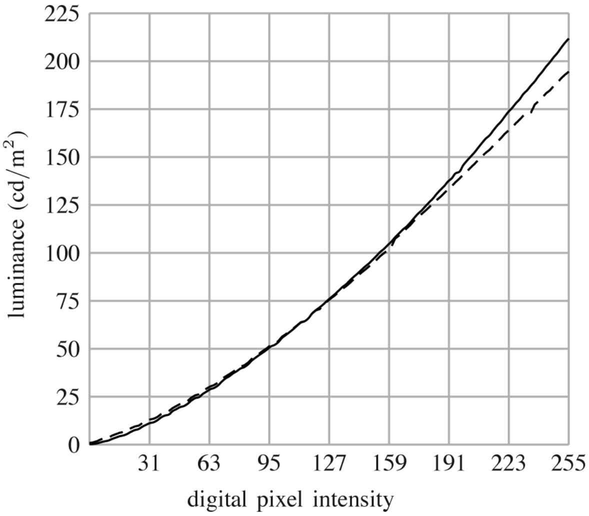

Fig. 4. Luminance in $\mathrm{cd/m^2}$ of the monitor (y-axis) plotted against the digital greyscale intensity sent to the video card (x-axis). Solid line is the luminance measured at the location of the left image's center and the dashed line is the luminance measured at the right image's center.

PR-650) over a range of digital pixel values. The luminance was computed for each digital value using the CIE 1931 observer [17] (see Fig. 4). All participants viewed the monitor from approximately 50 cm away while wearing any corrective lenses they would normally use for computer viewing. When displayed on the screen, the images had a diameter of approximately 8.5 cm and so occupied an angle of approximately 10° from the viewer's eye position. All negative image pixel intensities were truncated at 0 in the reconstructed images since negative intensities are not displayable. The maximum pixel intensity in the set was found (~1.36) and all pixels' intensities were then scaled by the same value (~186.94) to ensure the entire set fell on the range (0,255) for greyscale display. By imposing a consistent scaling on all images, some (e.g., those with a low-intensity target feature and low noise power) did not use the entire range of display intensities. Using the above scaling, we see that target feature intensity for $b = 1/20$ was mapped to 9.35 on the greyscale range, $b = 1/12$ was mapped to 15.59, and $b = 5/36$ was mapped to 25.96. The actual displayed intensity of the target feature was likely higher in almost every image, due to additions from the anatomical background underneath.

B. Channelized Hotelling Observer

We applied the CHO [12], [13] with Gabor channels to our synthetic images to compare the results with human performance. Gabor channels in particular were used because it has been suggested they are a useful approximation for the grating response of the human visual system [9], [10], [18]. The channels are defined by the response equation

$$\mathbf{G}[\boldsymbol{x}] = \exp\left(-4(\ln 2) f^2 \frac{(\boldsymbol{x}-\boldsymbol{x_0})^t(\boldsymbol{x}-\boldsymbol{x_0})}{w_s^2}\right) \times \cos\left(2\pi f \{\cos\phi, \sin\phi\}(\boldsymbol{x}-\boldsymbol{x_0}) + \beta\right) \quad (13)$$

where $\boldsymbol{x_0}$ is the image center, f is the central frequency of the filter in cycles per pixel, w_s is the filter width in octaves of f, ϕ is the angle of the filter, and $\beta \in \{0, (\pi/2)\}$ determines if it is odd or even. We have used a setup with 40 channels, based on the example of Eckstein [9], [10], with $w_s = 0.8825$ and $\phi \in \{0, (2\pi/5), (4\pi/5), (6\pi/5), (8\pi/5)\}$. To compute the central frequencies, note that each pixel subtends 5/64° from the viewer's eye position. We would like our filters to have frequencies of 2, 4, 8, or 16 cycles per degree. Converting this to cycles per pixel gives $f \in \{(5/32), (5/16), (5/8), (5/4)\}$. The 16 384 × 40 channel matrix $\mathbf{C}$ is produced by rearranging each channel $\mathbf{G}$ as a 16 384 × 1 vector and making them each a column of $\mathbf{C}$. We can then compute the 40 × 1 channel response vector, $\boldsymbol{u}$ of image $\mathbf{Y}'$ by rearranging the image into the 16 384 × 1 vector, $\boldsymbol{y}'$ and setting

$$\boldsymbol{u} = \mathbf{C}^t \boldsymbol{y}' \quad (14)$$

where $\mathbf{C}^t$ is the transpose of $\mathbf{C}$.

In order to derive the CHO for each of the 18 experimental conditions, we must compute the specific covariance matrix $\mathbf{K}_{\boldsymbol{u}}^c$ for each condition c. We first note that

$$\mathbf{K}_{\boldsymbol{u}}^c = \frac{1}{2}\left(\mathbf{K}_{\boldsymbol{u},0}^c + \mathbf{K}_{\boldsymbol{u},1}^c\right) + \mathbf{K}_{\boldsymbol{\epsilon}}^c \quad (15)$$

where $\mathbf{K}_{\boldsymbol{u},0}^c$ and $\mathbf{K}_{\boldsymbol{u},1}^c$ are the covariance matrices of the channel responses in condition c's target feature-absent and -present cases, respectively, and $\mathbf{K}_{\boldsymbol{\epsilon}}^c$ is the covariance matrix of the observer's internal noise process in condition c. This process is assumed to add noise independently to each response channel by sampling from a Gaussian with zero mean and variance depending on the channel. Following the example of Eckstein [10], we define $\mathbf{K}_{\boldsymbol{\epsilon}}^c = \alpha \mathrm{Diag}((1/2)(\mathbf{K}_{\boldsymbol{u},0}^c + \mathbf{K}_{\boldsymbol{u},1}^c))$ where $\mathrm{Diag}()$ zeroes all the off-diagonal elements of its argument and α is a proportionality constant that can be varied to reduce the absolute performance of the model observer.

This leaves the problem of determining $\mathbf{K}_{\boldsymbol{u},0}^c$ and $\mathbf{K}_{\boldsymbol{u},1}^c$. Given that we have closed forms for neither the pixel covariance of the anatomical backgrounds, nor the effects of the wavelet filter, we opted to estimate $\mathbf{K}_{\boldsymbol{u},0}^c$ and $\mathbf{K}_{\boldsymbol{u},1}^c$ from the synthetic data. Since we had 672 target feature-present and 672 target feature-absent images in each of the 18 experimental conditions, we used the channel responses from 400 images to estimate $\mathbf{K}_{\boldsymbol{u},0}^c$ and $\mathbf{K}_{\boldsymbol{u},1}^c$ for each condition. This left 272 image pairs for testing each condition.

With the 18 $\mathbf{K}_{\boldsymbol{u}}^c$ estimates computed, we can determine the optimal channel weights $\boldsymbol{v}^c$ for each condition, c, according to the Hotelling strategy as

$$\boldsymbol{v}^c = \left(\mathbf{K}_{\boldsymbol{u}}^c\right)^{-1}\left(\langle \boldsymbol{u}_1^c \rangle - \langle \boldsymbol{u}_0^c \rangle\right) \quad (16)$$

where $\langle \boldsymbol{u}_1^c \rangle$ and $\langle \boldsymbol{u}_0^c \rangle$ are the sample mean target feature-present and target feature-absent channel response vectors for condition c. From this, we can write the template $\boldsymbol{w}^c$ applied by the observer in condition c as

$$\boldsymbol{w}^c = \mathbf{C}\boldsymbol{v}^c. \quad (17)$$

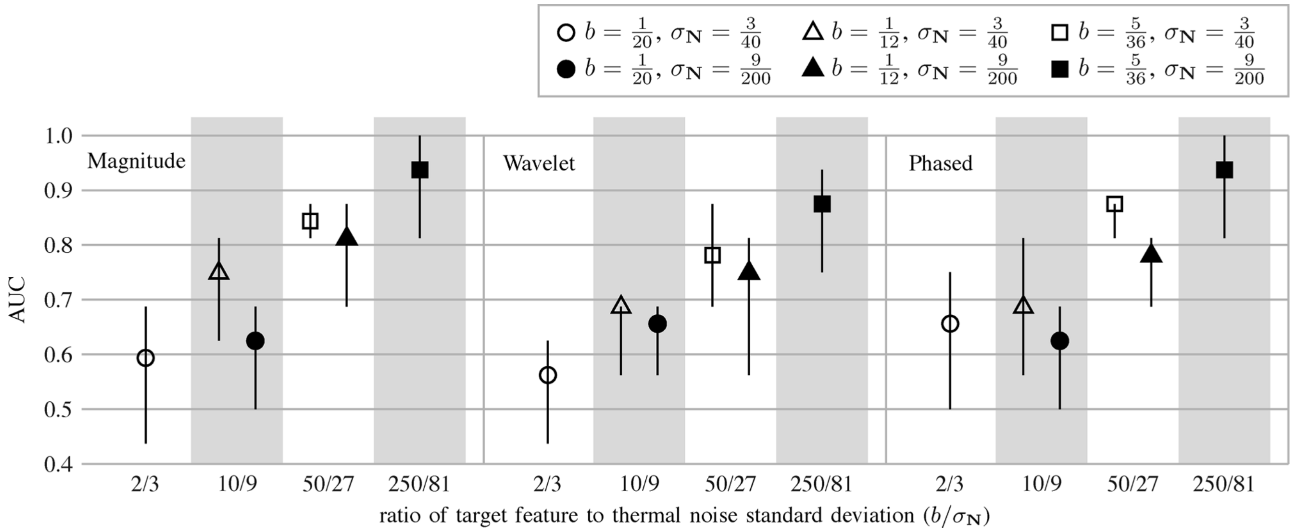

Fig. 5. Plot of the median, first quartile, and third quartile of AUC for the human observers in each of the 18 experimental conditions. The y axis is the AUC score. Each experimental condition is represented by a symbol (one of three shapes, either filled or unfilled) that locates the median AUC and two vertical lines that represent the first and third quartiles of the AUC. The symbol's shape represents a different configuration of target feature and noise power, as explained in the legend at the top-right of the chart. The x axis is divided into thirds, with each third containing results from one of the reconstruction technique as labeled in the top-left corner of each third. Inside of each third, the x axis is further divided by the ratio of target feature to thermal noise standard deviation (complex feature-SNR). In some cases, two experimental conditions have the same feature to noise ratio (e.g., $b = 1/12, \sigma_{\mathbf{N}} = 3/40$ and $b = 1/20, \sigma_{\mathbf{N}} = 9/200$) and so appear in the same band on the diagram.

This template can be used to calculate the response, λ, to a reconstructed image $\mathbf{Y}'$. Reordering the 128 × 128 matrix $\mathbf{Y}'$ to the 16,384 × 1 vector $\boldsymbol{y}'$, we write

$$\lambda = (\boldsymbol{w}^c)^t \boldsymbol{y}' \tag{18}$$

where $\boldsymbol{y}'$ was produced with conditions c.

If there is no internal decision noise (i.e., if $\alpha = 0$) then determining the model observer's choice in a 2AFC experiment requires calculating the λ for each of the two image choices and then selecting the image with the larger score. However, when $\alpha \neq 0$, we must add the internal noise of the observer. Rather than compute a noise for each channel, we note that the effect of the channel decision noises is combined in the final response score. Thus, we can modify the computed response by adding a single sample ϵ from a zero-mean Gaussian distribution with variance

$$\sigma_\epsilon^2 = (\boldsymbol{v}^c)^t \mathbf{K}_{\boldsymbol{\epsilon}} \boldsymbol{v}^c. \tag{19}$$

Adding this noise sample gives our final estimate of the score a human observer would assign to the image

$$\lambda' = \lambda + \epsilon. \tag{20}$$

As before, the image with the greater λ' in each pair is considered the CHO selection in the 2AFC trial.

V. Results and Discussion

A. Human Observer Study

For each participant, we computed the percentage correct, Pc, in each of the 18 experimental conditions. We computed the median Pc over all the participants as well as the first and third quartiles. It can be shown that in a 2AFC task, the Pc is also an estimator for the area under the curve (AUC) of the experiment's ROC [19]. Noting this, we have plotted the first, second, and third quartiles of the AUC in Fig. 5.

While the width of the first and third quartiles indicates substantial inter-subject variability, it seems that there is no clear difference between the three reconstructions across all experimental conditions. The resolution of our AUC quartile measurements is only 1/16 because we opted to cover a variety of experimental cases and thus show each participant each condition only 16 times. Although this structure reduces the effect of observer variability, it also means that we are unable to demonstrate statistically significant differences between the reconstructions given the small effect.

Our human observers did not show a measurable increase in performance when using the phase-corrected real reconstruction, despite the oft-noted fact that these images should have higher contrast than the magnitude reconstruction [5], [6], [16]. The lack of effect here may be due the to fact that noise distribution of a magnitude image becomes very similar to that of a real-reconstruction when $(\mathbf{S}[\boldsymbol{x}]/\sigma_{\mathbf{N}}) > 3$. Since many of our target features were added on top of bright anatomy, these two reconstructions should be effectively the same in many of the tested cases. A difference might be found in the more specific task of locating faint features in dark areas of MR images. Alternatively, experiments could be conducted with substantially higher $\sigma_{\mathbf{N}}$, although the realism of increased noise powers given current MRI technology is questionable.

Similarly, we do not show an improvement in signal detection using the wavelet transform, despite the improved image SNR demonstrated by the algorithm [2]. It is not clear if a specific task would be better suited to this filter, since it has been described as useful generally for MRI. Clearly, the wavelet basis underlying

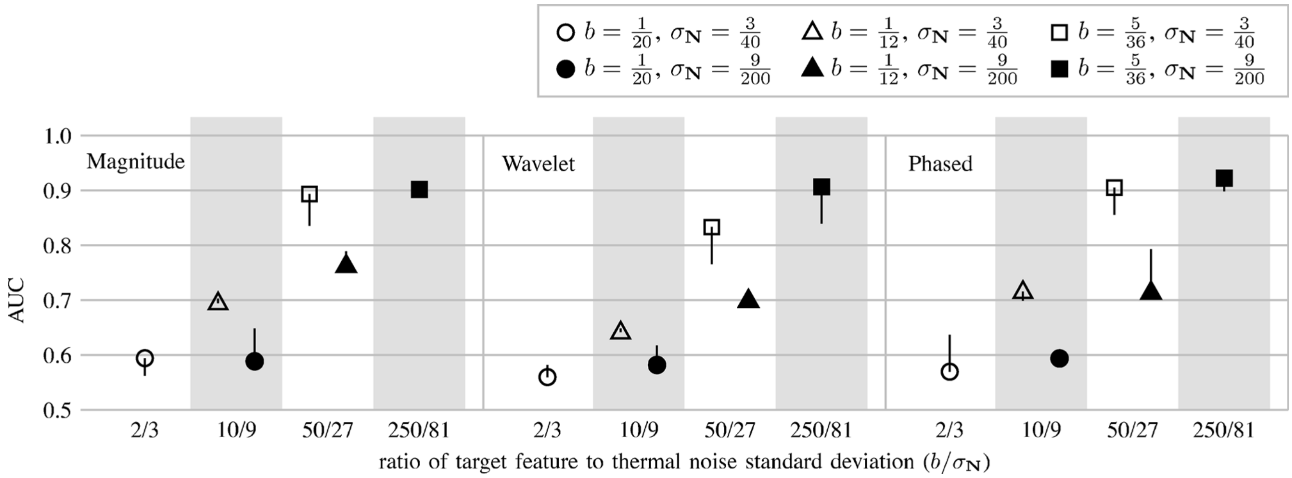

Fig. 6. Plot of the mean AUC of the the model and human observers in each of the 18 experimental conditions. The y axis is the AUC score. Each experimental condition is represented by a symbol (one of three shapes, either filled or unfilled) that locates the mean model observer AUC and one vertical line that represents the AUC of the mean human observer. Longer vertical lines indicate a greater mismatch between the model and human observers. The symbol's shape represents a different configuration of target feature and noise power, as explained in the legend at the top-right of the chart. The x axis is divided into thirds, with each third containing results from one of the reconstruction technique as labeled in the top-left corner of each third. Inside of each third, the x axis is then further divided by the ratio of target feature to thermal noise standard deviation (complex feature-SNR). In some cases, two experimental conditions have the same PSNR (e.g., $b = 1/12, \sigma_{\mathbf{N}} = 3/40$ and $b = 1/20, \sigma_{\mathbf{N}} = 9/200$) and so appear in the same band on the diagram.

the filtering algorithm could be varied, and other wavelet processing algorithms could be implemented as well. However, our results indicate that claiming improved pixel SNR alone does not necessarily imply improved signal detection in MR images

Considering the magnitude and phase-corrected real AUCs, we note that in every case where two conditions share the same ratio of target feature amplitude to thermal noise standard deviation (vertical bands in Fig. 5), the condition with higher target feature intensity outperforms the case with the lower target feature intensity. Since both the feature-to-noise ratio and anatomical background intensity were held constant in these cases, the only changes are the increase in target feature and noise intensity relative to the anatomical background. We hypothesize that this effect is due to the anatomical background obscuring the target feature more often when the feature is less intense. In particular, we note that at $b = 1/20$ (the circles in Fig. 5), there is little difference in the magnitude and phase-corrected real AUCs, despite a change in thermal noise power. This seems to indicate that the anatomical background has become the dominant distractor in the image at low target feature intensity. Combined with anecdotal comments from our subjects, this encourages us that using anatomical backgrounds to provide realistic distractors is important in studying feature detection.

Although the wavelet reconstruction produced similar results, it is unclear if this also demonstrates the effects of the anatomical background. The wavelet filter uses $\sigma_{\mathbf{N}}$ as an input to control the quantity of smoothing performed and so there is clearly a nonlinear relationship, at least in theory, between AUC, the target feature intensity, and the thermal noise power. For example, we note that while the magnitude and phase-corrected reconstructions had approximately the same results in the lowest target feature intensity case (circles in Fig. 5), the wavelet reconstruction was still sensitive to the thermal noise power. We cannot differentiate with our experiment whether the dominant effect in these conditions is the thermal noise or the smoothing artifacts. However, we note that there is still a tendency, although weaker than in the other two reconstructions, towards higher AUC with brighter signal intensity at each feature intensity to thermal noise ratio. Since the brighter-feature/higher-noise images should also have more smoothing artifacts, the fact that they produced marginally higher AUCs than their darker-feature/lower-noise equivalents is an indication that the anatomy still played an important role as a distractor.

B. Model Observer Study

Ideally, we would compute the AUC or some similar metric directly from the description of our model observer and the image statistics. However, due to the fact that we have neither stationary image backgrounds, nor Gaussian-distributed model observer responses λ we calculated the AUC for the CHO by applying the template to each synthetic image, computing λ' using (20), and then calculating the Pc. We performed this operation for 50 separate instances of the model observer, and then computed the result of the mean observer in each of the experimental conditions. As in the human observer case, the Pc values were taken as estimates of the AUC.

The mean AUC of the model observers was fit to the mean human observer data by searching for the α that minimized the mean squared error (MSE) of all 18 experimental conditions. We determined the optimal setting to be $\alpha = 3.2$ $(\mathrm{MSE} = 0.00189)$ by doing exhaustive search on an initially coarse range of α values and then gradually refining the range. This computation took approximately 3 h on a 1 Ghz PowerPC G4. The mean AUC of the 50 model observers are shown in comparison to the mean human results in Fig. 6 (note that Fig. 5 displays the median human results while Fig. 6 displays the mean). Overall, the model observer shows a very good match with the human study. There is no appreciable difference between the three reconstructions according to the model observer. We note that there was a small disagreement between the model and mean human observer on the ordering of some of the weaker signals, although the inter-subject variability in these conditions was high as well. Additionally, there is a tendency for the model observer to overestimate mean human performance at the highest target feature intensity.

VI. Conclusion

Among our human observers, we noted no significant difference between the three reconstructions. This contradicts previous predictions based on the improvements in SNR produced by phase correction or wavelet filtering. Studies with finer resolution in their AUC estimates are required in order to better distinguish the effects of these reconstructions. It is also possible that more specific tasks may demonstrate benefits from different reconstruction techniques.

The model matched the human observers very well, also showing no difference between the three reconstructions. Given the computational efficiency of the CHO, and the close match to human results, the CHO may be useful in situations where a human study of MRI reconstruction is not feasible.

The contradiction between our experimental results and the predictions made in the literature from SNR measurements indicate that the use of observer studies seems preferable to SNR and similar metrics for evaluating MRI reconstruction and filtering. Our results also suggest that the choice of background is important due to its distracting effects across the range of feature intensities and noise powers considered. Given the variation in human responses, using a greater number of participants in a study with finer resolution is likely necessary to determine small effects. Additionally, while the difference between the CHO and mean human AUC is very low, further investigation is needed to determine if other model observers would further improve the fit with human data.

Acknowledgment

The authors would like to thank Dr. B. Maedler at Philips Medical Systems for providing the high-field MRI images and Dr. B. Funt at Simon Fraser University for his assistance in measuring the monitor luminance. The authors also thank the reviewers for their helpful comments that substantially improved the quality of our experiment.

References

[1] W. A. Edelstein, G. H. Glover, C. J. Hardy, and R. W. Redington, "The intrinsic signal-to-noise ratio in NMR imaging," *Magn. Reson. Med.*, vol. 3, pp. 604–618, 1986.

[2] R. D. Nowak, "Wavelet-based Rician noise removal for magnetic resonance imaging," *IEEE Trans. Image Process.*, vol. 8, no. 10, pp. 1408–1418, Oct. 1999.

[3] M. E. Alexander, R. Baumgartner, A. R. Summers, C. Windischberger, M. Klarhoefer, E. Moser, and R. L. Somorjai, "A wavelet-based method for improving signal-to-noise ratio and contrast in MR images," *Magn. Reson. Imag.*, vol. 18, pp. 169–180, 2000.

[4] P. Bao and L. Zhang, "Noise reduction for magnetic resonance images via adaptive multiscale products thresholding," *IEEE Trans. Med. Imag.*, vol. 22, no. 9, pp. 1089–1099, Sep. 2003.

[5] M. A. Bernstein, D. M. Thomasson, and W. Perman, "Improved detectability in low signal-to-noise ratio magnetic resonance images by means of a phase-corrected real reconstruction," *Med. Phys.*, vol. 16, pp. 813–817, Sep. 1989.

[6] D. Noll, D. Nishimura, and A. Macovski, "Homodyne detection in magnetic resonance imaging," *IEEE Trans. Med. Imag.*, vol. 10, no. 2, pp. 154–163, Jun. 1991.

[7] M. D. Tisdall and M. S. Atkins, "MRI denoising via phase error estimation," *Proc. SPIE*, vol. 5747, pp. 646–654, 2005.

[8] Z. Chang and Q.-S. Xiang, "Nonlinear phase correction with an extended statistical algorithm," *IEEE Trans. Med. Imag.*, vol. 24, no. 6, pp. 791–798, Jun. 2005.

[9] M. P. Eckstein, C. K. Abbey, and J. S. Whiting, "Human vs model observers in anatomic backgrounds," *Proc. SPIE*, vol. 3340, pp. 16–26, 1998.

[10] M. P. Eckstein, C. K. Abbey, F. O. Bochud, J. L. Bartoff, and J. S. Whiting, "The effect of image compression in model and human performance," *Proc. SPIE*, vol. 3663, pp. 243–252, 1999.

[11] J. P. Rolland and H. H. Barrett, "Effect of random background inhomogeneity on observer detection performance," *J. Opt. Soc. Am. A*, vol. 9, no. 5, pp. 649–658, May 1992.

[12] K. J. Myers and H. H. Barrett, "Addition of a channel mechanism to the ideal-observer model," *J. Opt. Soc. Am. A*, vol. 4, no. 12, pp. 2447–2457, Dec. 1987.

[13] C. K. Abbey and F. O. Bouchud, "Modelling visual detection tasks in correlated image noise with linear model observers," in *Handbook Med. Imag.*. Bellingham, WA: SPIE, 2000, vol. 1, ch. 11, pp. 629–654.

[14] A. Macovski, "Noise in MRI," *Magn. Reson. Med.*, vol. 36, pp. 494–497, 1996.

[15] V. Kuperman, *Magnetic Resonance Imaging—Physical Principles and Applications*. New York: Academic, 2000.

[16] H. Gudbjartsson and S. Patz, "The Rician distribution of noisy MRI data," *Magn. Reson. Med.*, vol. 34, pp. 910–914, 1995.

[17] G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd ed. New York: Wiley, 1982.

[18] S. Park, "Efficiency of the human observer detecting random signals in random backgrounds," *J. Opt. Soc. Am. A*, vol. 22, no. 1, pp. 3–16, Jan. 2005.

[19] H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating function," *J. Opt. Soc. Am. A*, vol. 15, no. 6, pp. 1520–1535, June 1998.