HHS Public Access

Author manuscript

Proc SPIE Int Soc Opt Eng. Author manuscript; available in PMC 2010 September 21.

Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 1996 January 1; 2712: 47–58. doi:10.1117/12.236860.

Observer signal-to-noise ratios for the ML-EM algorithm

Craig K. Abbey[1,2], **Harrison H. Barrett**[1,2,3], and **Donald W. Wilson**[2]

[1] Program in Applied Mathematics, University of Arizona, Tucson, AZ 85724

[2] Dept. of Radiology, University of Arizona, Tucson, AZ 85724

[3] Optical Sciences Center, University of Arizona, Tucson, AZ 85724

Abstract

We have used an approximate method developed by Barrett, Wilson, and Tsui for finding the ensemble statistics of the Maximum Likelihood-Expectation Maximization algorithm to compute task-dependent figures of merit as a function of stopping point. For comparison, human-observer performance was assessed through conventional psychophysics.

The results of our studies show the dependence of the optimal stopping point of the algorithm on the detection task. Comparisons of human and various model observers show that a channelized Hotelling observer with overlapping channels is the best predictor of human performance.

Keywords

model observers; observer performance; ML-EM algorithm; image quality

1 Introduction

Studies of model-observer performance can generally be placed into one of two categories: studies in which statistical properties of the imaging system are well specified and hence model-observer performance is reported using ensemble statistics (mathematical formulas), and studies in which the statistical properties are not well specified and hence model-observer performance is reported using sample statistics usually through Monte-Carlo studies. In studies where the samples (images) may be difficult to obtain either because of computational intensity or large-sample requirements, the ensemble approach is preferable if the ensemble statistics can be computed. The goal of this work is to use ensemble statistical approaches to analyze model-observer performance on images produced by the Maximum Likelihood-Expectation Maximization (ML-EM) algorithm. For comparison, we also report human performance measured by conventional psychophysics.

The ML-EM algorithm has received considerable attention as a method of image reconstruction and restoration ([1] – [4]). It has the theoretically attractive properties of converging to a maximum-likelihood solution, implicitly applying a positivity constraint, and enforcing better agreement with the data in subsequent iterations. However it is not clear that these mathematical considerations translate into improved performance in diagnostic tasks. In the past, it has been difficult to study the statistical properties of this algorithm because of its nonlinear nature. However, a recent article by Barrett, Wilson, and Tsui [5]

derives an approximate method for finding the necessary noise properties of this algorithm. We use these ensemble noise properties to compute model observer signal-to-noise ratios for task-based assessment of image quality. The present paper extends the work of Barrett, Wilson, and Tsui by considering stochastically defined image backgrounds in which the distribution of intensity within the object is presumed to follow a prior probability law.

Since humans are the end user of most medical images it is important to understand the effects of different choices made during the image reconstruction process on the resulting diagnostic performance. Due to the proliferation of reconstruction algorithms and the possibly large number of free parameters within a given algorithm, conventional psychophysical studies become infeasible for a thorough investigation of optimal processing. To this end, model observers and, specifically, linear model observers have been proposed as a more efficient way to optimize medical imaging systems for diagnostic performance.

In this work we consider the effect of stopping point on detection performance. Terminating the iterative scheme well before convergence is a simple way to reduce the high level of noise usually found in the unconstrained maximum-likelihood estimate of the reconstructed image ([6] – [8]). As with most methods of regularization, the strength with which the regularizer is applied (in this case the number of iterations the algorithm is allowed to run) is left to the user.

Our conclusions are twofold. Our observer studies indicate that the optimal stopping point is highly dependent on the task being considered. Similar results have been reported for linear iterative algorithms [9] using the ensemble statistical approach and for other forms of regularization in nonlinear algorithms [10] using Monte-Carlo methods. These results reinforce the argument that measures of image quality must take the diagnostic task into account. We also report on the ability of a number of proposed model observers to predict the outcome of human psychophysical studies.

2 Theory

In signal detection theory, one adopts the view that an image is a multidimensional random variable coming from one of two possible probability distributions. The distribution describing the image depends on whether the signal is actually present or absent in the image. The task of the observer is to decide from which distribution a given image comes. In theory, the observer makes its decision by forming a scalar response to the input image and subjecting this response to threshold. It is the statistics of this response variable that determine the observer performance. Detection performance measures are generally thought of as measures of separation between the distribution of responses to signal-present images and signal-absent images.

In this work we utilize the approach defined in [11] for objective assessment of image quality. Here the performance metric is the observer signal-to-noise ratio (SNR), the expected difference in the means of the two response distributions divided by their average variance. The model observers used in this work are all linear functions of the image data

and therefore the SNR can be computed directly from the first and second-order statistics of the images. The SNR of a linear observer is defined by the following formula,

$$\mathrm{SNR}^2(\mathbf{w}) = \frac{(\Delta \mathbf{s}^t \mathbf{w})^2}{\mathbf{w}^t \mathbf{K} \mathbf{w}}, \quad (1)$$

where $\Delta\mathbf{s}$ is the vector difference in the expected signal-present and signal-absent images, $\mathbf{K}$ is the average covariance of the two classes of images, and $\mathbf{w}$ is a vector representing the linear observer.

The images analyzed in this work come from a simulated two-dimensional parallel-beam single photon emission computed tomography (SPECT) imaging system. The tomographic imaging system is modeled as a linear system

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \mathbf{n}, \quad (2)$$

where $\mathbf{f}$ is a vector representing the object of interest, $\mathbf{H}$ is the system matrix, $\mathbf{n}$ is a vector representing the noise associated with the system due to the Poisson statistics of gamma-ray emission, and $\mathbf{g}$ is the measured data. The ML-EM algorithm approximates the object of interest by a vector $\hat{\mathbf{f}}^{k+1}$ through the following iterative scheme

$$\hat{f}_n^{k+1} = \frac{\hat{f}_n^k}{\sum_{m'=1}^{M} H_{m'n}} \sum_{m=1}^{M} H_{mn} \frac{g_m}{\sum_{n'=1}^{N} H_{mn'} \hat{f}_{n'}^k}. \quad (3)$$

Approximate formulas for the statistical properties of images produced by ML-EM algorithm have been worked out in [5]. The basic approach is to linearize each step of the iterative algorithm. The statistical properties of the algorithm are computed from the resulting sequence of linear transformations. A related approach is used in [13] to investigate resolution properties of penalized-likelihood algorithms. The expected reconstruction at a given stopping point is approximated by the image produced when the ML-EM algorithm is run on noiseless projection data. This vector is denoted by $\mathbf{a}^{k+1}$. The image covariance is approximated by

$$\mathbf{K}_{\mathbf{f}^{k+1}} \simeq \mathrm{Diag}\left(\mathbf{a}^{k+1}\right) \mathbf{U}^{k+1} \mathbf{K}_{\mathbf{g}} \left[\mathbf{U}^{k+1}\right]^t \mathrm{Diag}\left(\mathbf{a}^{k+1}\right), \quad (4)$$

where $\mathbf{K}_{\mathbf{g}}$ is the data covariance and $\mathbf{U}^{k+1}$ represents the effects linearizing (3). The interested reader is referred to [5] and [14] for expressions regarding the structure of $\mathbf{U}^{k+1}$. Extensive Monte-Carlo studies of this approximate covariance were performed in [12] to test the validity of the first-order approximation required for (4). Generally good agreement was found in the range of 50,000 expected counts per data set corresponding to about 60 expected counts per detector in the center of the image. In this range, the error in approximation was less than 10% at mid-range stopping points (more than 10 iterations and less than 200 iterations). In the present work, a much larger system is used (128 × 128 pixel

images reconstructed at 64 angles as opposed to 32 × 32 pixel images from 32 angles), but with 400,000 expected counts in the data set, there are still about 60 expected counts per detector.

The data covariance takes two forms in this work. For the background-known-exactly (BKE) study, the Poisson statistics of the data collection process are the only source of variation in the collected data. Hence the data covariance is given by

$$\mathbf{K_g} = \mathrm{Diag}(\mathbf{Hf}). \quad (5)$$

In the second study, the object **f** is also presumed to be random, creating a "lumpy" background of the sort used in [15]–[17]. In this case the data covariance is of the form

$$\mathbf{K_g} = \mathrm{Diag}\left(\mathbf{H}\bar{\mathbf{f}}\right) + \mathbf{H}\mathbf{K_f}\mathbf{H}^t, \quad (6)$$

where $\bar{\mathbf{f}}$ and $\mathbf{K_f}$ are the object mean and covariance respectively. The key requirement which allows us to extend the approximate to lumpy background data is relatively small object covariance. The addition of object covariance adds a source of variance to each data point which is 16% of the size of the variance due to Poisson counting statistics.

Human observer performance was assessed through conventional psychophysics. Two-alternative forced-choice experiments were conducted to estimate the probability of a correct identification (p_c) in image pairs sampled from the signal-present and signal-absent distributions. The p_c estimates were transformed to d' – an estimated human signal-to-noise ratio – by the formula [18]

$$d' = \sqrt{2}\Phi^{-1}(p_C),$$

where Φ^{-1} is the inverse cumulative normal transformation.

3 Experimental results

The task considered was detection of a centered Gaussian bump against a flat or lumpy background. The radius of the signal – measured as the standard deviation of the Gaussian – was 4.0 pixels. Images were reconstructed within a window of radius 64 pixels inscribed in 128 × 128 pixel images. The signal contrast was 19.8% in the BKE images and 29.7% in the lumpy-background images. These contrasts were determined from pilot psychophysical studies to give an acceptable range of d' values for human performance [18]. Parallel-beam projections were collected at 64 angles equally spaced within a semicircle around the object. Because our objective was analysis of the reconstruction algorithm, perfect collimation of the projection data was assumed. Poisson noise in the data was set at 400,000 expected counts per data set.

Performance was assessed for a number of different model observers for comparison to the psychophysical results. One notable omission from the following list of observers is that of the Hotelling or optimal linear observer ([19] and [20]). Computation of the template **w** for this observer involves computing the pseudo-inverse of $\mathbf{K}_{\hat{\mathbf{f}}^{k+1}}$, a 16,384 × 16,384 element matrix. While direct inversion of this matrix can be avoided by an iterative search for **w** [20], the large number of evaluations involving products with $\mathbf{K}_{\hat{\mathbf{f}}^{k+1}}$ make this approach difficult as well. In addition, the Hotelling observer is independent of stopping point for a large class of linear algorithms [9].

The model observers considered here were:

1. A region-of-interest (ROI) observer

This observer has a template w whose elements are 0 for pixels outside the region of interest and 1 for pixels inside the region of interest. For our studies, the region was a disk of radius 5.67 pixels corresponding to the standard deviation of point the signal profile. This observer model is similar to that used by Hanson [21] to analyze the ART algorithm.

2. The Nonprewhitening Matched Filter (NPW)

This observer is given by the expected profile of the signal after reconstruction. From the results of [5], the observer template is well approximated by computing *noiseless* signal-present and signal-absent reconstructions and taking their difference. This expected signal difference, which we shall denote Δs^{k+1} is also used for Δs in Equation (1) for evaluating all model observer SNRs. The NPW observer is known to be optimal in stationary Gaussian white noise [22].

3. Channelized Hotelling Observers

This observer uses a bank of frequency-selective filters to reduce the image to a much smaller number of filter responses. Optimal linear discrimination is then performed on this reduced set of responses. The channel model can be thought of as a bank of image templates represented by the matrix **T**. The template in each column of **T** is one of the channel filters represented in the spatial domain33. The observer template associated with a channelized Hotelling observer is given by [9]

$$\mathbf{w}=\mathbf{T}\left(\mathbf{T}^t\mathbf{K}_{\hat{\mathbf{f}}^{k+1}}\mathbf{T}\right)^{-1}\mathbf{T}^t\Delta\mathbf{s}^{k+1}. \quad (7)$$

Two channelized Hotelling observers are used here, each being defined by the radial frequency profile of its image templates. The first uses has four square non-overlapping channels (SQR). The second uses three channels with overlapping difference-of-Gaussian (DOG) profiles. Plots of the frequency profiles are seen in Figure 1. Channel models have been analyzed in ([23], [24], and [9]) as predictors of human performance.

Performance of the various model observers as a function of stopping point is given in Figure 2. The different observer models show marked differences in performance. Some observers – most noticeably the SQR observer – show improved performance in the lumpy

background experiments. This is solely due to the increased contrast in these experiments. In lumpy backgrounds, all observers exhibited degraded performance relative to fixed backgrounds when the signals were of equal contrast.

Average human-observer performance results are plotted in Figure 3. Six observers participated in the studies, which consisted of 200 image pairs for each stopping point in both the fixed and lumpy backgrounds. All observers had participated in previous studies with similar images and were trained on an independent sample of 50 image pairs before taking each test. The experiments were randomized both in the order in which the image pairs were shown within a given test and in the order of the tests taken by each observer. The images were scaled for constant contrast across stopping point. In the fixed-background studies, the expected signal contrast was held constant at 7 grey levels of 256 in the display. In the lumpy background studies, the signal contrast was 10 grey levels.

The plots show a strong dependence on task in human performance. The fixed-background studies show a fairly consistent decrease in performance with subsequent iterations. The lumpy background studies indicate peaked performance in the range of 16 to 32 iterations.

The DOG observer was the only model observer that provided a reasonably consistent fit to the human-observer data in both backgrounds. Figure 3 also contains plots of the DOG SNR multiplied by a scale factor chosen to best fit the human data. Scaling of model-observer performance has been discussed in [25] as a way to incorporate degradation due to internal noise in the human observer when detecting localized signals. The least-squares scale values required for the DOG SNRs to fit the human data were 0.84 for the fixed background and 0.64 for the lumpy background. The χ^2 goodness-of-fit statistics based on the error bars in the plots were 4.78 ($p = 0.57$) and 6.63 ($p = 0.36$) for the fixed and lumpy backgrounds respectively. These χ^2 values must be interpreted carefully. While these goodness-of-fit statistics do not permit rejection at the usual $p < 0.05$ level, the relatively small sample sizes leave open the possibility that the experiments simply lacked the necessary statistical power to reject the fitted curves. In addition, this statistical analysis has not included the fact that the number, shape, and frequency range of the channels can all be adjusted. Hence the χ^2 values only apply to resolving whether *any* channel model can provide a reasonable fit to our human data.

Visual inspection of the fitted curves is similarly ambivalent. While the model observer performance curves capture the general trends exhibited by the human data, closer inspection reveals a number of possible discrepancies. In both studies the fitted curves seem to undershoot human performance at the first iteration. The fit in the lumpy background experiments also seems to overestimate the high-iteration human performance. These differences are not statistically significant, but nevertheless they remain a source of concern and merit further study.

4 Conclusions

We have used the approximate ensemble statistical properties of the ML-EM algorithm to quantify detection performance of model observers for two different tasks – detection of a Gaussian bump in fixed and lumpy backgrounds. Standard psychophysical methods have

been used to assess human performance. Our primary contribution is to have quantified the effect of the two tasks on optimal stopping point. All observers, both model and human, exhibit optimal performance at different stopping points in the different tasks. The more complex lumpy-background task requires more iterations of the ML-EM algorithm to achieve optimal performance. Similar behavior has been found in previous work on a different reconstruction algorithm [9]. Our results highlight the need for task-based methods to answer questions involving the optimal processing of medical images.

Our second conclusion regards the ability of model observers to reliably predict (and therefore remove the necessity of assessing) human performance. As we have pointed out, the only model that was at all consistent with the human data in both backgrounds was the DOG observer. While this observer does fit the data within standard statistical tolerances, visual discrepancies between the human observer results and the scaled model observer performance leave some doubt of the prospects of generalizing to larger sample sizes and other tasks without further psychophysical evaluations. Furthermore, there is not yet a systematic way to choose parameters of the channel model that guarantee good agreement with human performance. In short, the DOG channelized Hotelling observer is not yet a predictive tool.

Future directions of this work will pursue the conclusions of the present work. Since the task has been shown to have a fundamental effect on observer performance, new tasks need to be explored. Perhaps most important of the future tasks will be those in which the signal is not confined to the center of the reconstruction. Other tasks of interest include varying the lumpy background parameters, size and shape of the signal, and quantifying the effect of penalty functions on Maximum-Likelihood reconstructions.

New model observers need to be considered as well as new tasks. This work underscores the limited capacity of conventional linear observers to predict human performance on tomographic reconstructions. Observers based on nonlinear estimation of signal parameters have been proposed as models for signal detection [25], [26]. The ability of these observers to predict human performance in a wide range of detection tasks remains to be seen.

Acknowledgments

The authors would like to thank Jack Denny, Elizabeth Krupinski, and Ed Soares for many helpful discussions. We also wish to acknowledge the participants in the human-observer studies for their time and patience. This work was supported by NCI grant R01-CA52643.

References

1. Dempster AP, Laird NM, Ruben DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B. 1977; 39:1–38.
2. Shepp LA, Vardi Y. Maximum likelihood reconstruction for emission tomography. IEEE trans Med Imaging. 1982; MI-1:113–22. [PubMed: 18238264]
3. Lange K, Carson R. EM reconstruction algorithms for emission and transmission tomography. J Comput Assist Tomogr. 1984; 8:306–16. [PubMed: 6608535]
4. Miller MI, Snyder DL, Miller TR. Maximum-likelihood reconstruction for single-photon emission computed tomography. IEEE Trans Nuc Sci. 1985; NS-32:769–78.

5. Barrett HH, Wilson DW, Tsui BMW. Noise properties of the EM algorithm: I. Theor. Phys Med Biol. 1994; 39:833–46. [PubMed: 15552088]
6. Veklerov V, Llacer J. Stopping rule for the EM algorithm based on statistical hypothesis testing. IEEE Trans Med Imaging. 1988; MI-6:313–19.
7. Liow J-S, Strother SC. Practical tradeoffs between noise, quantitation, and number of iterations for maximum likelihood-based reconstructions. IEEE Trans Med Imaging. 1991; 10:563–71. [PubMed: 18222862]
8. Coakley KJ, Llacer J. The use of cross-validation as a stopping rule in emission tomography reconstruction. SPIE Image Phys. 1991; 1443:226–33.
9. Abbey, CK.; Barrett, HH. In: Bizais, Yves; Barillot, Christian; Di Paola, Robert, editors. Linear iterative reconstruction algorithms: Study of observer performance; Proc. 14th Int. Conf. on Information Processing in Medical Imaging; Dordrecht: Kluwer Academic; 1995. p. 65-76.
10. Wagner, RF.; Myers, KJ.; Hanson, KM.; Brown, DG.; Anderson, MP. Toward optimal observer performance of detection and discrimination tasks on reconstructions from sparse data. In: Hanson, KM.; Silver, RN., editors. Maximum Entropy and Bayesian Methods. Kluwer Academic; Dordrecht: 1996.
11. Barrett HH. Objective assessment of image quality: Effects of quantum noise and object variability. J Opt Soc Am A. 1990; 7:1266–78. [PubMed: 2370589]
12. Wilson DW, Tsui BMW, Barrett HH. Noise properties of the EM algorithm: II. Monte Carlo simulations. Phys Med Biol. 1994; 39:847–71. [PubMed: 15552089]
13. Fessler JA, Rogers WL. Spatial resolution properties of penalized-likelihood image reconstruction: Spatially invariant tomographs. IEEE Trans Image Proc EDICS. :2.3–1995.
14. Wilson, DW. PhD Dissertation. University of North Carolina; 1994. Noise and resolution properties of FB and ML-EM reconstructed SPECT images.
15. Barrett HH, Rolland JP, Wagner RF, Myers KJ. Detection and discrimination of known signals in inhomogeneous, random backgrounds. Proc SPIE. 1989; 1090:176–82.
16. Myers KJ, Rolland JP, Barrett HH, Wagner RF. Aperture optimization for emission imaging: effect of a spatially varying background. J Opt Soc Am A. 1990; 7:1279–93. [PubMed: 2370590]
17. Rolland JP, Barrett HH. Effect of random background inhomogeneity on observer detection performance. J Opt Soc Am A. 1992; 9(5):649–58. [PubMed: 1588452]
18. Burgess AB. Comparison of receiver operating characteristic and forced choice observer performance measurement methods. Med Phys Biol. 1995; 22(5):643–55.
19. Smith WE, Barrett HH. Hotelling trace criterion as a figure of merit for the optimization of imaging systems. J Opt Soc Am A. 1986; 3:717–25.
20. Fiete RD, Barrett HH, Smith WE, Myers KJ. The Hotelling trace criterion and its correlation with human observer performance. J Opt Soc Am A. 1987; 4:945–53. [PubMed: 3598746]
21. Hanson KM. Method of evaluating image recovery algorithms based on task performance. J Opt Soc Am A. 1990; 7:1294–1304.
22. Wagner RF, Brown DG. Unified SNR analysis of medical imaging systems. Phys Med Biol. 1985; 30:489–518.
23. Myers KJ, Barrett HH. Addition of a channel mechanism to the ideal-observer model. J Opt Soc Am A. 1987; 4:2447–57. [PubMed: 3430229]
24. Eckstein MP, Whiting JS. Lesion detection in structured noise. Acad Radiol. 1995; 2:249–53. [PubMed: 9419557]
25. Burgess AB, Ghandeharian H. Visual signal detection. II. Signal-location identification. J Opt Soc Am A. 1984; 1:907–10.
26. Kijewski MF. The Barankin bound: A model of detection with location uncertainty. Proc SPIE. 1992; 1768:153–60.

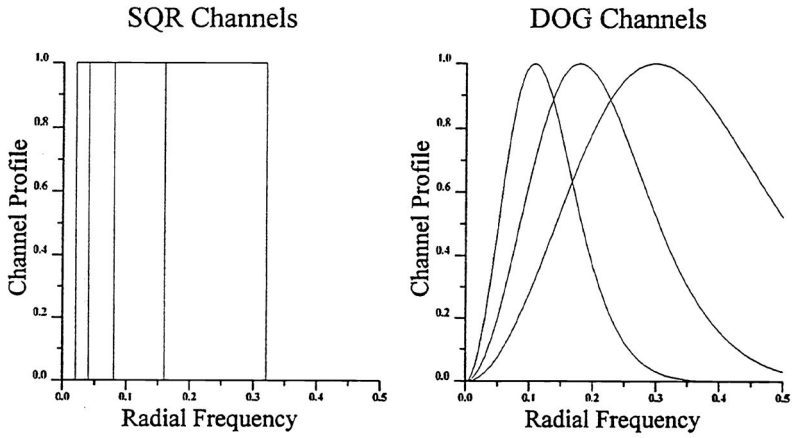

Figure 1.

Plots of the channel profiles for the two channelized Hotelling observers used. The plots range from 0 to the Nyquist frequency of 0.5 pixels^{-1}. The channels for the SQR observer are non-overlapping and confined to a limited radial band of spatial frequencies. The DOG observer has overlapping channel profiles, each spanning a broad range of radial frequencies.

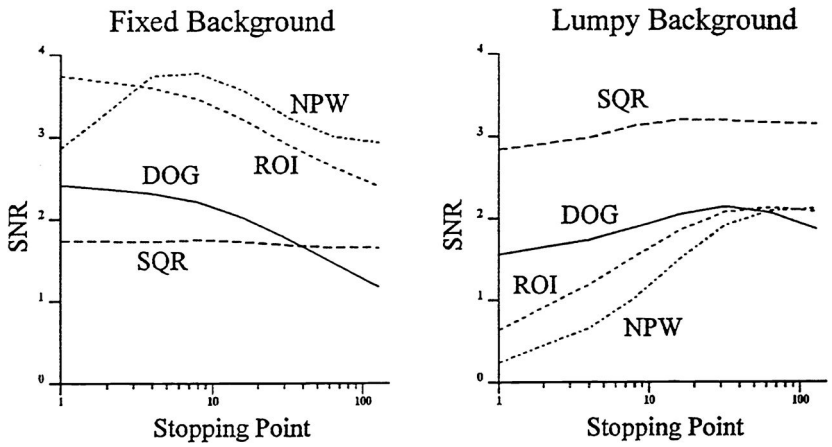

Figure 2.

Model observer performance as a function of stopping point. These plots represent the performance of the model observers for the fixed and lumpy background detection tasks. The observer plots are identified as follows: ROI – Region of Interest, NPW – Non-prewhitening, SQR – Square channel Hotelling, DOG – Difference of gaussian Hotelling.

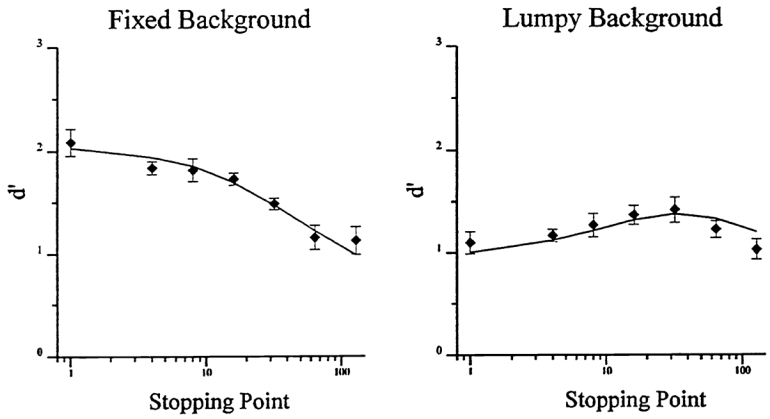

Figure 3.

studies and a constant scale parameter fit of DOG SNRs. Error bars represent one unit of the standard error in average performance.