Evaluation of Four Probability Distribution Models for Speckle in Clinical Cardiac Ultrasound Images

Zhong Tao, Hemant D. Tagare*, *Member, IEEE*, and James D. Beaty, *Member, IEEE*

***Abstract*—Segmenting cardiac ultrasound images requires a model for the statistics of speckle in the images. Although the statistics of speckle are well understood for the raw transducer signal, the statistics of speckle in the image are not. This paper evaluates simple empirical models for first-order statistics for the distribution of gray levels in speckle. The models are created by analyzing over 100 images obtained from commercial ultrasound machines in clinical settings. The data in the images suggests a unimodal scalable family of distributions as a plausible model. Four families of distributions (Gamma, Weibull, Normal, and Log-normal) are compared with the data using goodness-of-fit and misclassification tests. Attention is devoted to the analysis of artifacts in images and to the choice of goodness-of-fit and misclassification tests. The distribution of parameters of one of the models is investigated and priors for the distribution are suggested.**

***Index Terms*—Cardiac image analysis, speckle probability density, ultrasound segmentation.**

I. Introduction

Maximum-likelihood (ML) and maximum *a posteriori* (MAP) approaches to image segmentation require gray level probability densities in different regions of the image. In principle, exact probability densities can be derived using acoustic physics and knowledge of the ultrasound machine signal processing chain.

In practice, however, this is very difficult to carry out. It is especially difficult in a clinical setting where images are often acquired without complete information about the acquisition process. For example, machine parameters might not be recorded by the operator or the signal processing chain in the machine might not be completely disclosed by the manufacturer. Sometimes the machine malfunctions, or an inexperienced operator uses nonoptimal settings. Finally, many clinical images reside in archives which do not store machine settings.

Lacking complete information for theoretical modeling, the image processor is often forced to use empirical—and admittedly inexact—models. These models are not perfect, but the hope is that they are useful for the task at hand.

Manuscript received January 30, 2006; revised April 27, 2006. This work was supported by the National Library of Medicine Grant R01-LM06911. *Asterisk indicates corresponding author*.

Z. Tao is with R2 Technology Inc., Sunnyvale, CA 94087 USA.

*H. D. Tagare is with the Department of Diagnostic Radiology and the Department of Biomedical Engineering, Yale University, New Haven, CT 06520 USA.

J. D. Beaty is with the Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723 USA.

Digital Object Identifier 10.1109/TMI.2006.881376

In this paper, we investigate this frankly pragmatic point of view. We want to know how effective empirical models are in the task of segmenting ultrasound images. We evaluate four probability density models with a data set of over 100 real-life cardiac short-axis B-mode clinical images.

We intend this to be a nuts-and-bolts paper about real data. Real data often contains unpleasant surprises, and indeed we found a few in ultrasound data (the spike artifacts in Section III). We report all of the surprises as well as tricks we used to overcome them—reporting the scruffiness of real data is an integral part of our validation.

A. Short-Axis B-Mode Images

At clinical frequencies of 2.5–5 MHz, there is significant backscatter from myocardial tissue but very little backscatter from the blood pool. Nevertheless, speckle-like graininess can be seen in the blood pool regions of the image. This is most likely due to backscatter from the tissue into the side-lobes of the transducer and is sometimes referred to as "stationary" speckle. Whatever the ultimate origin of the signal in the image blood pool, its probability density is required for a ML or a MAP segmentation algorithm.

We measure the ability of empirical models to fit the observed gray level densities in the myocardial tissue and blood pool regions. We use two criteria for evaluating empirical models: 1) the goodness-of-fit to real image histograms, and 2) the ability of the models to distinguish between blood and myocardium in the image. The first measures accuracy of modeling, while the second measures segmentation ability.

All images in this paper were obtained from three clinical echo-cardiogram machines: Acuson Sequoia C256, ATL Sono-CT models 5000 and 3000. Most of the images came from the Acuson Sequoia C256 and were imaged at 3.5 MHz. The other images were obtained at 2.5, 4.5, and 5 MHz.

No attempt was made to control acquisition parameters—the images were obtained under whatever settings the clinical technician found useful. The settings were not recorded, nor were the images recorded in a standard format such as DICOM. These are real conditions that image processors face.

B. Organization of the Paper

The paper is organized as follows. Section II contains a brief review of the literature. Section III contains preliminary analysis that motivates the rest of the paper. Section IV contains the four models for the distributions of tissue and blood. Section V discusses goodness-of-fit. Section VI contains the details of the misclassification rate test. Section VII contains the experimental results and their analysis. Section VIII concludes the paper.

II. Background and Literature Review

B-mode ultrasound images are formed by recording the envelope of the backscattered ultrasound acoustic wave as it propagates through an acoustic medium. B-mode images appear "grainy," and the grain is commonly referred to as *speckle*. The statistics of speckle depends on size and organization of scatterers in the acoustic medium.

A. Theoretical Models

A comprehensive overview of theoretical speckle modeling is offered by Insana, Myers, and Grossman [10]. Goodman [7] describes the process of scattering for speckle in laser images. He showed that fully developed speckle can be described by a Rayleigh probability density. Burckhardt [5] showed that this also holds for narrow band ultrasound speckle. Wagner *et al.* [16] considered quasi periodic scatterers and derived a Rician model for speckle. Wagner *et al.* [18] also analyzed the envelope-detected signal and presented a straightforward procedure for separating the specular from the speckle based on second-order statistics at the transducers. Jakeman and Tough [11] suggested the K distribution as a generalized statistical model for weak scattering. Weng *et al.* [20] also proposed the K distribution. Dutt and Greenleaf [22] developed a statistical model for log-compressed envelope signal. Shankar [23] proposed the Nakagami distribution. Abyratne and Petropulu [1] proposed a complex model containing the combined effect of unresolvable diffuse scatterers, unresolvable periodicity from structural scatterers, correlated nonperiodic component of diffuse and structural scatterers, and periodic scatterers. These models are widely used in speckle reduction [24]–[28] and tissue characterization [29]–[32].

However, these models only give the speckle probability density at the transducer. The density has to be transformed into speckle density in the image. This is a complicated task for two reasons. First, most clinical ultrasound machines process the transducer signal through stages such as envelope detection, logarithmic amplification, time gain correction, interpolation, etc., before the signal is presented as an image. Propagating densities through this complex signal processing chain is not an easy task for many of the afore mentioned densities. Second, as mentioned in Section I, acquisition parameters are not always available to the image processor.

B. Empirical Models

Although empirical models are routinely used in segmenting ultrasound images, very little has been reported regarding their validity. In [21], Zimmer *et al.* evaluated a log-normal model for speckle in liver images. Xiao *et al.* [35] used a log-normal model for modeling speckle in breast images. Speckle is also observed in laser imaging [33] as well in synthetic aperture radar (SAR) imagery [34]. The Gamma distribution is often used as a good model for speckle in SAR (e.g., [34]).

III. Preliminary Investigation

We begin with a preliminary analysis of two cardiac ultrasound images. The aim of the preliminary investigation is to suggest plausible models for gray level distributions of tissue and blood. Below, we illustrate the procedure with two representative images.

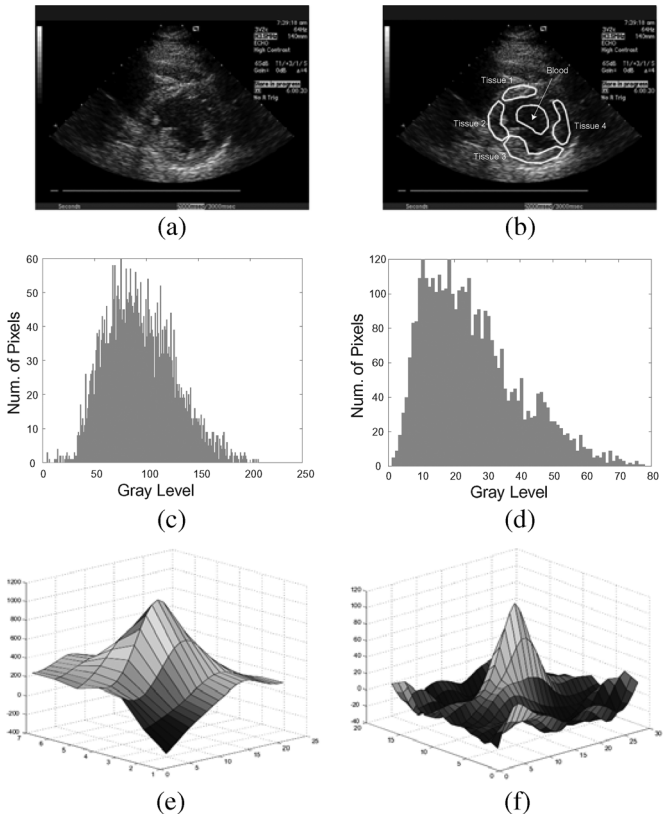


Fig. 1. Analysis of example image 1. (a) Original image. (b) Regions of interest. (c) Histogram of gray levels in tissue region 3. (d) Histogram of gray levels in blood. (e) Correlation in tissue 3. (f) Correlation in blood.

A. First Image

A B-mode short axis cardiac ultrasound image is shown in Fig. 1(a). It has a dark region in the center which is the blood pool and a bright annulus around the center which is the myocardium. From the image, it is clear that the brightness of the myocardium is variable. The myocardium is brighter in the 12 and 6 o'clock positions than at 3 and 9 o'clock. To capture this variation, we defined four regions of interest in the myocardium at roughly 12, 3, 6, and 9 o'clock, as shown in Fig. 1(b) and labeled tissue 1, tissue 2, tissue 3, and tissue 4, respectively. A region of interest in the blood pool was also defined.

The histograms of gray levels in tissue 3 and blood are shown in Fig. 1(c) and (d). While these histograms indicate that the distributions are unimodal, they do not necessarily represent the first-order distributions in these regions. This is because the gray levels in tissue and blood regions are correlated. Fig. 1(e) and (f) shows the correlation functions for tissue 3 and blood. The correlation is high for small shifts, but drops to substantially low values at 7 pixels. This was observed in all other regions and in other images as well.[1]

We subsampled the region of interest by a factor of seven along each dimension and then obtained the histogram of gray

[1]The length at which correlation drops low varies between the four tissue regions. However, this length is always less than seven pixels in our images, so that at seven pixels, the correlation is quite small.

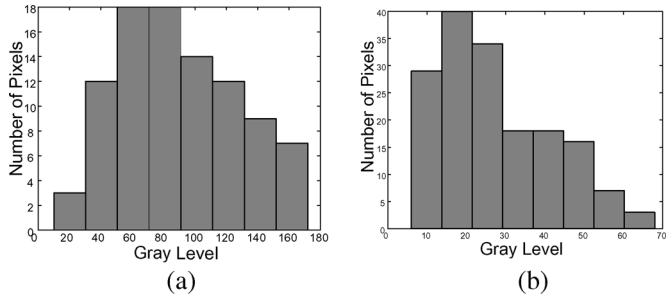


Fig. 2. Histogram of subsampled gray levels. (a) Histogram of subsampled gray levels in tissue 3. (b) Histogram of subsampled gray levels in blood.

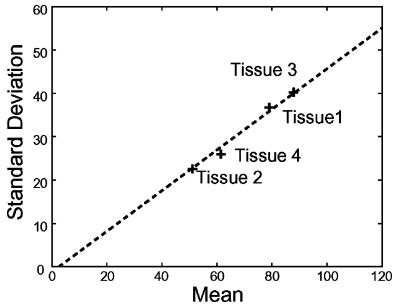


Fig. 3. Mean versus standard deviation of tissue gray levels.

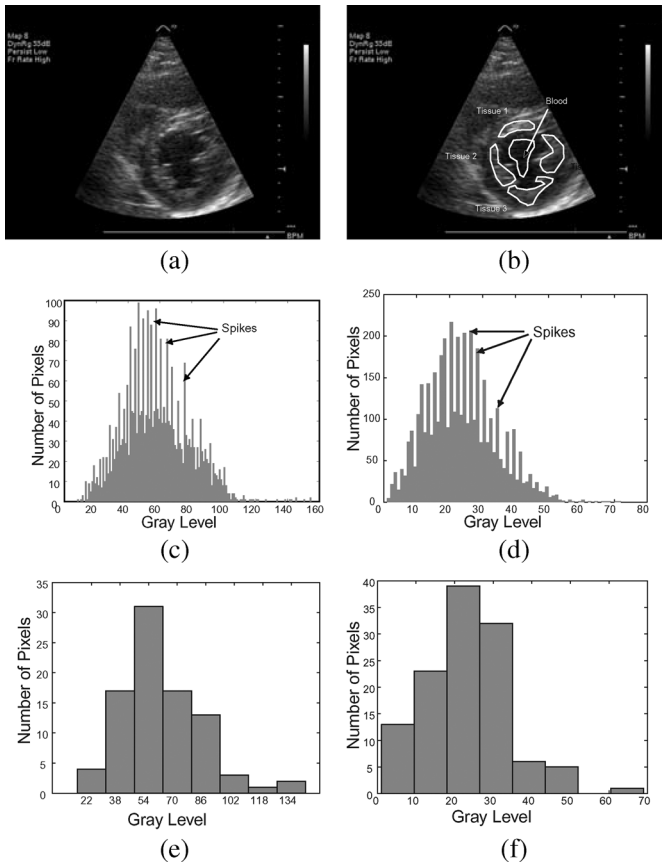


Fig. 4. Analysis of example image 2. (a) Original image. (b) Regions of interest. (c) Histogram of gray levels in tissue 3. (d) Histogram of gray levels in blood. (e) Coarse histogram for tissue 3. (f) Coarse histogram for blood.

levels of these pixels. This eliminates the effect of correlation on the histogram. The subsampled regions do not have sufficient number of pixels to give continuous histograms. So we binned the gray levels into eight bins covering the range from the minimum to the maximum gray level. The histograms after binning are shown in Fig. 2(a) and (b) for tissue region 3 and the blood pool. The figures suggest that a unimodal smooth distribution is a good model for both tissue and blood. Varying the number of bins does not alter the essential shape of these histograms. The histograms for tissue regions 1, 2, and 3 are similar to that of tissue region 4 and we do not display them here to conserve space.

An analysis of the means and standard deviations of the gray levels in the tissue regions is also revealing, as shown in Fig. 3. To a good approximation, the standard deviation varies linearly with the mean. This suggests that a unimodal family of scalable distributions is a good model for tissue.

B. Second Image

The second image and its regions of interest are shown in Fig. 4(a) and (b). The gray level correlation functions are similar to the one in the previous image and are not shown. As before, the correlation is substantially low at seven pixels. The histograms of gray levels in tissue 3 and blood (after subsampling by seven pixels) are shown in Fig. 4(c) and (d). In contrast to the histograms of the previous image, these histograms have significant spikes. The spikes occur in the blood and all of the tissue regions. They seem to occur roughly periodically in the histogram, and each spike is followed by an empty, or low count bin. The extent of the spike above a smooth histogram base line is approximately the missing count in the next bin.

Spike artifacts were present in almost all the images that we obtained from one of the machines. Possible explanations for the spikes are: 1) they are the result of internal image compression and decompression used to reduce the bandwidth to the display buffer in the machine; 2) they are the result of poor a/d conversion; and 3) they are the result of a speckle reduction algorithm in the machine.[2]

Any goodness-of-fit test that attempts to fit a smooth distribution to the histograms of Fig. 4(c) and (d) is bound to fail since the test will attempt to measure the fit between the distribution and the spike-empty-bin pattern. To avoid this, we choose to bin the histograms. Some experimentation with bin sizes showed that the choice of eight bins covering the histogram suppressed the artifact while retaining information about the overall shape of the distribution. Fig. 4(e) and (f) shows the coarse binned histograms for the data in Fig. 4(c) and (d).

Fig. 5 shows the means and standard deviations of the four tissue regions. In spite of the artifacts, the standard deviation again appears to be linearly related to the mean. This supports the idea mentioned above that scalable unimodal family of distributions could be used to model tissue.

Based on these preliminary results, we turn to a more rigorous analysis of the data. To conduct the analysis, we first select the

[2]The last two explanations were suggested by the anonymous reviewers.

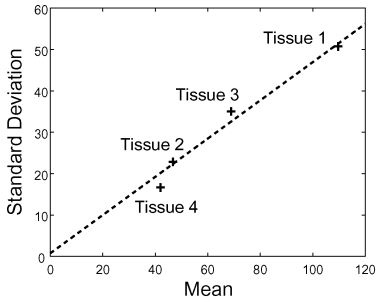


Fig. 5. Mean versus standard deviation for tissue gray levels in Fig. 4.

probability models and then evaluate the goodness-of-fit and the ability of the models to classify tissue and blood.

IV. Models

We chose four parameteric distributions as potential models for tissue and blood gray levels. The distributions are all unimodal and scalable. They are as follows.

1) *The Gamma Distribution*. This distribution has a shape parameter α and a scale parameter β with $\alpha, \beta > 0$. The distribution is given by

$$p_\gamma(x|\alpha,\beta) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, & \text{if } x \geq 0, \\ 0, & \text{otherwise} \end{cases}. \quad (1)$$

We included the Gamma distribution because it is used to model speckle in SAR [34]. Also the Rayleigh distribution is a member of this family of distributions (for $\alpha = 2$).

2) *The Weibull Distribution*. The Weibull distribution also has a shape parameter α and a scale parameter β, with $\alpha, \beta > 0$. This distribution is

$$p_w(x|\alpha,\beta) = \begin{cases} \frac{\alpha}{\beta} x^{\alpha-1} e^{-x^\alpha/\beta}, & \text{if } x \geq 0, \\ 0, & \text{otherwise} \end{cases}. \quad (2)$$

The Weibull distribution is often used to model the maximum of independent and identically distributed (i.i.d.) random variables. Since ultrasound images are formed from the envelope of the transducer signal, and envelopes are local maxima, this is a reasonable distribution to consider.

3) *The Normal Distribution*. This has a location parameter μ and a scale parameter $\sigma > 0$

$$p_n(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}. \quad (3)$$

The appeal of the normal distribution is that it gives simple algorithms for ML and MAP algorithms. We include it because it is often the first choice of a model in many image processing problems.

4) *The Log-normal Distribution*. This too has a location parameter μ and a scale parameter $\sigma > 0$

$$p_{ln}(x|\mu,\sigma) = \begin{cases} \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}. \quad (4)$$

This distribution has been used to model ultrasound liver images [21]. The log-normal distribution has one advantage in the ultrasound context. It is very easy to create a model for ultrasound dropout with it. This has been exploited in breast ultrasound images [35].

V. Goodness of Fit Test

A goodness-of-fit test determines whether a data set comes from a given theoretical probability distribution. Our exposition of goodness-of-fit tests closely follows the one in [4].

Suppose we want to test whether data $x_1, \ldots, x_n$ are i.i.d. samples from the probability distribution $p_x(x)$. We start by partitioning the range of the data into M bins $B_i, i = 1, \ldots, M$ and counting the number of data N_i that fall into each bin. If we let $\int_{B_i} f(x)dx$ denote the integral of the function f over all x in the bin B_i, then

$$p_i = \int_{B_i} p_x(x)dx \quad (5)$$

is the probability that the random variable x, distributed according to $p_x(x)$, will take values in the bin B_i. It is easy to show that the quadratic form in N_i defined by

$$X^2 = \sum_1^M \frac{(N_i - np_i)^2}{np_i} \quad (6)$$

has approximately $\chi^2(M-1)$ distribution. The significance level, also called the P-value [4], of accepting the hypothesis that the observations are distributed according to p_x can be obtained from standard χ^2 significance tables using (6). This is the chi-square goodness-of-fit test [4].

The standard significance table for χ^2 gives the probability of a Type I error: that is, the probability that the hypothesis being tested (the so-called null hypothesis) is rejected by the test when it actually holds. On the other hand, we are interested in accepting the null hypothesis (that the data come from the given distribution). Hence, we interpret the significance value in the standard table in the opposite way: a larger significance value indicates a better fit.

A. Rao–Rabson Statistic

When the probability distribution is parametric, the parameters have to be estimated from the data before the goodness of the fit can be calculated. This violates the assumption in the usual goodness-of-fit test that the distribution is completely known.

One solution to this is to seek other quadratic forms in N_i which have the property that they tend to the chi-square distribution when estimates of the parameters are obtained from the data. The Rao–Rabson statistic is one such quadratic form [4]. If $p_x(x|\theta)$ is the parametric form of the distribution, and the $\hat{\theta}$ is the maximum likelihood estimate of θ, then the Rao–Rabson statistic is given by

$$RR = V^T(\hat{\theta}) Q(\hat{\theta}) V(\hat{\theta}) \quad (7)$$

where, $\hat{\theta}$ is the maximum-likelihood estimate of the parameter vector θ, and

$$V(\theta) = \frac{N_i - np_i}{(np_i)^{1/2}}$$

$$Q(\theta) = I_M + D(\theta)\left[J(\theta) - D^T(\theta)D(\theta)\right]^{-1} D^T(\theta)$$

where $J(\theta)$ is the Fisher information matrix at θ and the matrix D is given by

$$D_{ij}(\theta) = p_i(\theta)\frac{\partial p_i(\theta)}{\partial \theta_j}$$

where, as before, $p_i(\theta) = \int_{B_i} p_x(x|\theta)dx$ is the integral of $p_x(x|\theta)$ in the bin B_i, and θ_j is the jth component of θ. The Rao–Rabson statistic RR is chi-square distributed with $M - p - 1$ degrees-of-freedom. As above, the significance level of accepting the hypothesis that the observations are distributed according to $p_x(x|\theta)$ can be obtained from standard χ^2 significance tables.

Besides allowing for parameterized densities, the Rao–Rabson statistic has the advantage that it let us bin the data at a given level of coarseness. This reduces the effect of the spike artifact and allows more approximate fits.

VI. Classification Test

A goodness-of-fit test does not directly evaluate the ability to segment an image. To measure this, we use the misclassification rate of the generalized likelihood ratio test (GLRT). The GLRT classifies a region R as blood or tissue depending on whether the generalized likelihood ratio

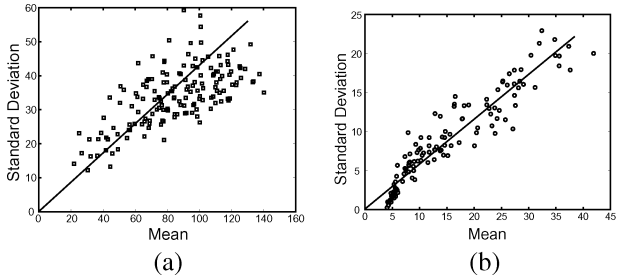
$$\Lambda = \frac{\sum_{i\in R} \log p_b(x_i|\hat{\theta}_b)}{\sum_{i\in R} \log p_t(x_i|\hat{\theta}_t)} \tag{8}$$

is greater than or less than 1. Here, $p_t(x_i|\hat{\theta}_t)$ and $p_b(x_i|\hat{\theta}_b)$ are the probability distributions of gray levels in tissue and blood, $\hat{\theta}_t$ and $\hat{\theta}_b$ are the maximum-likelihood estimates of the parameters of the tissue and blood distributions, and x_i is the gray level of the pixel $i \in R$.

In the experiments, we manually outline myocardial tissue and blood regions. The distribution parameters are estimated from the gray levels in the regions. Then we randomly choose a square regions of $N \times N$ pixels in tissue and blood regions and use the GLRT to classify the entire region as tissue or blood. All misclassifications are reported as a percentage.

VII. Evaluations

A total of 195 short axis cardiac ultrasound images were collected from the three clinical machines described in Section I. Of the 195 images, some images where found to have inappropriate dynamic ranges—the blood pool gray level values were cut off at 0, and/or the tissue gray levels were saturated at 255. These images were discarded. There were 119 images in which the blood was within the dynamic range of the machine and these images were used to analyze the blood gray levels. There were 135 images in which the tissue gray levels were within the dynamic range and these were used for analyzing tissue.



(a) (b)

Fig. 6. The relation between mean and standard deviation. (a) Mean versus standard deviation for tissue for all images. (b) Mean versus standard deviation for blood for all images.

As in the preliminary investigation, the four tissue regions (tissue 1, tissue 2, tissue 3, and tissue 4) and the blood region of interest were outlined manually. The regions were subsampled by a factor of 7 to avoid correlated pixels. Also, as for the preliminary investigation, the tissue 3 region gray levels were used in the experiments.

The aim of experimental analysis was to evaluate the suggestions of preliminary analysis. The first question was whether the linearity between the standard deviation and mean suggested in Figs. 3 and 5 holds for the entire data set. This is a key point since this linearity is the basis for using a scalable family as a model.

A. Linear Relation Between Mean and Standard Deviation

To check whether linearity between the standard deviation and the mean holds over the entire data set, we plotted the mean and standard deviation for all tissue and blood regions in all images. To be consistent, within each region we chose a fixed number of pixels (60 pixels after subsampling by 7) from which to calculate the mean and standard deviation.

Fig. 6(a) shows the mean versus standard deviation plot for tissue and Fig. 6(b) shows the mean versus standard deviation plot for blood for all images. The two plots clearly suggest a linear relation between the mean and standard deviation.

B. Goodness-of-Fit Test

The aim of the next analysis was to evaluate how well the four models of Section IV fit the data. The Rao–Rabson statistic of the goodness-of-fit of the four probability models was calculated for each tissue and blood region. The Rao–Rabson statistic requires a choice of bins. We choose a fixed number of bins M for every histogram. Of the M bins, $M - 2$ bins are uniformly spaced between the minimum and maximum data value. Two additional bins, one from the maximum value to $+\infty$ and one from $-\infty$ to the minimum value are added. The point of these infinite bins was to evaluate the "excess" distribution in the tails of the probability models. After some experimentation, we found that a value of M around 10 was a good compromise between compensating for the effect of the spike artifact and retaining the overall shape of the histogram. Hence, we evaluated the goodness-of-fit at $M = 7$, 9, 11, and 13.

TABLE I
SIGNIFICANCE VALUES OF THE MODEL FITS FOR BLOOD

Bins	Gamma	Weibull	Normal	Log-normal
7	0.31	0.40	0.28	0.13
9	0.34	0.38	0.26	0.16
11	0.32	0.35	0.25	0.16
13	0.34	0.33	0.23	0.19

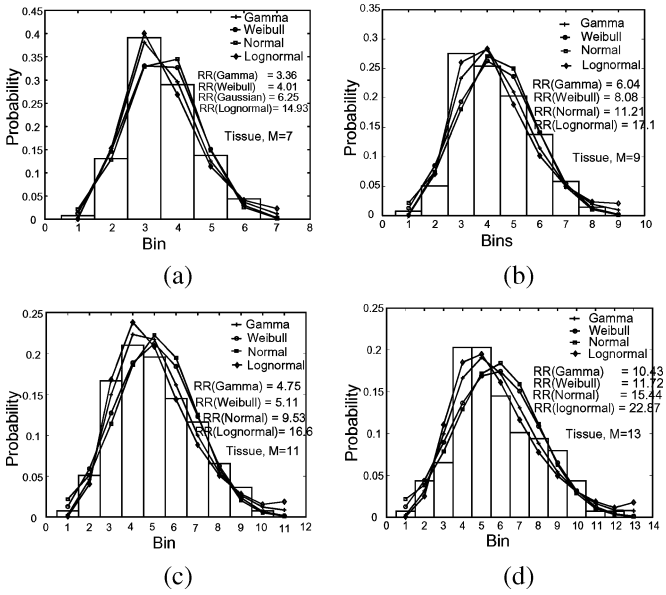


Fig. 7. Typical maximum likelihood fits of the models to tissue. (a) Number of bins $M = 7$. (b) Number of bins $M = 9$. (c) Number of bins $M = 11$. (d) Number of bins $M = 13$.

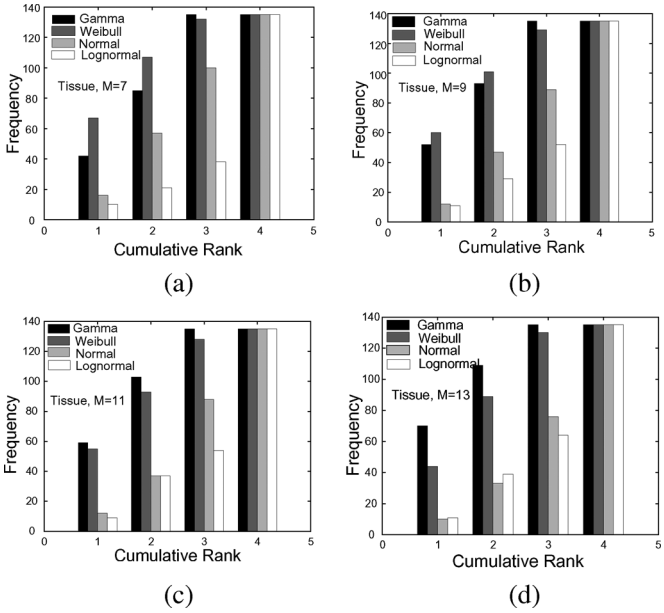


Fig. 8. Ranks of goodness-of-fit of the four models to tissue. (a) Number of bins $M = 7$. (b) Number of bins $M = 9$. (c) Number of bins $M = 11$. (d) Number of bins $M = 13$.

TABLE II
SIGNIFICANCE VALUES OF THE MODEL FITS FOR BLOOD

Bins	Gamma	Weibull	Normal	Log-normal
7	0.23	0.23	0.04	0.06
9	0.23	0.23	0.03	0.08
11	0.21	0.21	0.03	0.08
13	0.20	0.19	0.02	0.08

C. Goodness-of-Fit for Tissue

Table I shows the average significance values over the entire dataset of the goodness-of-fit for tissue as obtained from RR-statistic and a standard chi-square significance table (see the discussion in Section V). The significance values are not high enough to support the null hypothesis at a conventionally acceptable level (which would be 0.9 or more). This point is discussed below in more detail. Nevertheless, as we show in a typical example below, the fits of the models to the data are quite close.

From Table I, it is clear that Gamma and Weibull are better fits than normal or log-normal. To further evaluate the fits, we show in Fig. 7 a typical fit to data from a single image. The figure shows the binned tissue histogram from a typical image for $M = 7, 9, 11, 13$. The first bin of the histogram represents gray levels from $-\infty$ to the minimum gray level in the data and the last bin represents gray levels from the maximum value to $+\infty$. The bins in between are evenly spaced from the minimum to the maximum value. The value of the bar in each bin shows the percentage of gray levels in the tissue region-of-interest that fall into this bin. The figure also shows the probabilities $p_i = \int_{B_i} p_x(s)dx$ (see Section V) due to each model (parameters estimated by maximum-likelihood) as a continuous curve superimposed on the data bars. It quite clear from the figures that the models are close to the data.

To understand the significance values further, for each image we ranked the models according to it. The model with the highest significance value was ranked 1, the next lowest ranked 2, and so on. Fig. 8(a)–(d) shows histograms of goodness-of-fit ranks for the entire tissue data for $M = 7, 9, 11, 13$ (number of bins), respectively. The ranks are shown cumulatively, i.e., for rank $k = 1, \ldots, 4$, the figure shows the number of images for which each model had a goodness-of-fit rank less than or equal to k. Thus, the data in Fig. 8(a) at rank $= 1$ shows the number of images for which each model had a rank less then or equal to 1. At rank $= 2$, it shows the number of images for which each model had a rank less then or equal to 2, etc.

From Fig. 8, it is again clear that the normal and log-normal distributions are consistently worse fits than Gamma and Weibull. The Weibull distribution ranks 1 more often than Gamma for $M = 7$ and 9 (the two are quite close at $M = 9$), with the Gamma distribution ranking better for $M = 11$ and 13. Since increasing M increases the resolution of the goodness-of-fit, it appears that the Gamma distribution is the best of these models for modeling tissue, with Weibull a close second.

D. Goodness-of-Fit for Blood

Table II shows the significance values of the fit of the models to blood gray levels. Again, Gamma and Weibull appear to be better models than Normal and Log-normal.

Fig. 9 shows a typical fit of the distributions to the data. The figure shows the binned blood histogram from a typical image for $M = 7, 9, 11, 13$. As above, the first bin of the histogram represents gray levels from $-\infty$ to the minimum gray level in

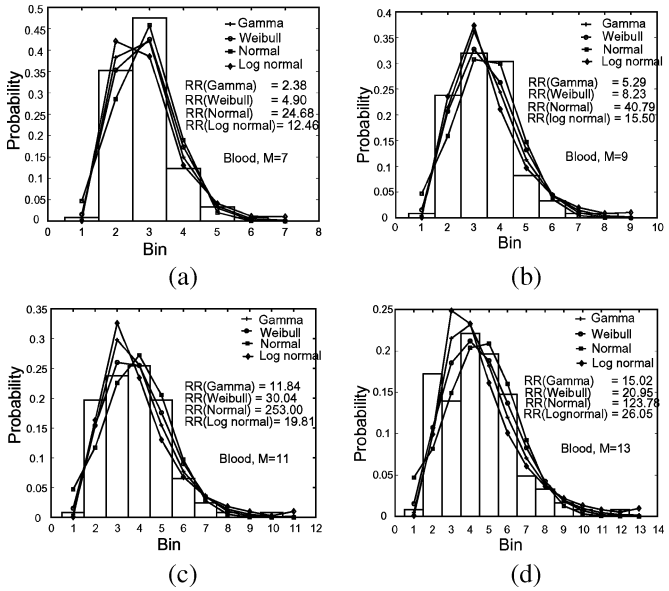


Fig. 9. Typical maximum likelihood fits of the models to blood data. (a) Number of bins $M = 7$. (b) Number of bins $M = 9$. (c) Number of bins $M = 11$. (d) Number of bins $M = 13$.

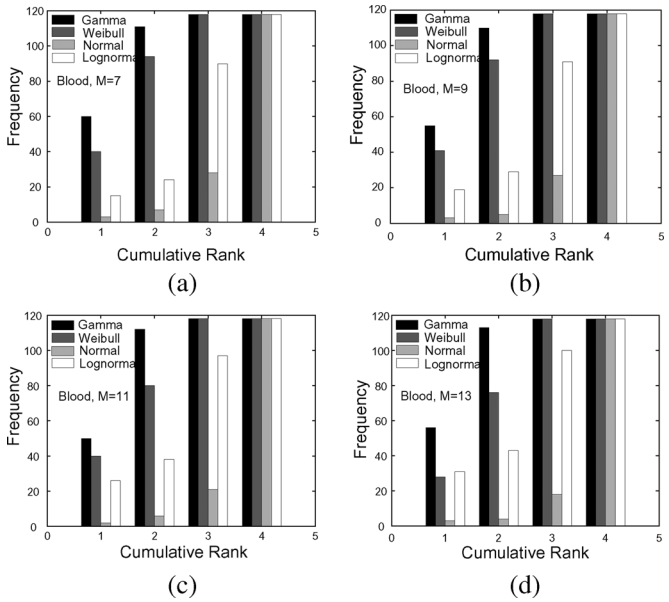


Fig. 10. Ranks of goodness-of-fit of the four models to blood. (a) Number of bins $M = 7$. (b) Number of bins $M = 9$. (c) Number of bins $M = 11$. (d) Number of bins $M = 13$.

the data and the last bin represents gray levels from the maximum value to $+\infty$.

Fig. 10(a)–(d) shows the goodness-of-fit cumulative ranks for blood for $M = 7, 9, 11, 13$, respectively. Here too, it is clear Gamma and Weibull outperform normal and log-normal, with Gamma being the best fit.

E. Classification

Next, we evaluated the ability of the four models to distinguish between tissue and blood by using the misclassification rate of the GLRT. Recall from Section VI, the GLRT is applied to a region of $N \times N$ pixels in order to classify all of it as tissue or blood. We used different sized regions corresponding to $N = 1$, 3, 5, 7, 9, 11, and for each region, we obtained the misclassification rate of the GLRT.

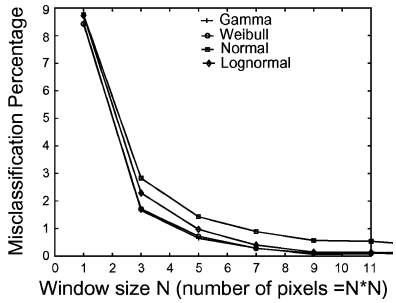


Fig. 11. Misclassification rates for GLRT.

Fig. 11 shows the misclassification rate as a function of N for the four models. The Gamma and Weibull distributions have a lower misclassification rate than normal and log-normal. The Gamma distribution slightly outperforms Weibull. Further, the misclassification rate falls significantly with increasing N, being less than 1 percent at $N = 5$ for Gamma and Weibull.

These experimental results suggest that the Gamma and Weibull distributions are appropriate for modeling blood and tissue, with Gamma slightly better than Weibull.

F. Discussion

One of the fundamental relations we found is the one shown in Fig. 6(a) and (b). These plots show that over the range of subjects, machines, operating frequencies, and other settings, the standard deviation of the gray levels in tissue as well as in blood varies approximately linearly with the mean of the signal. It suggests using a scalable family of probability density functions for modeling the tissue and blood gray levels.

However, some caution is required before blindly using this relation. Recall that we eliminated images that had either blood gray levels below 0 or tissue levels above 255. If such images need to be processed, then the modeling suggested by Fig. 6 is inappropriate. Explicit modeling of the cutoff and saturation may be necessary.

The next matter of interest is the fit of the four probability distributions to measured histograms. Taken at face value, none of the four empirical models fit the data with statistical significance. An informal evaluation of the histograms (such as the ones in Figs. 7 and 9) does not suggest any particular trend in the way the models fit the data. That is, it is not entirely clear that the mismatch is mostly in the central peaked portion of the histogram or the right or left tail of the histogram.

Further, there appears to be very little difference in the maximum-likelihood fits of the four probability densities to the binned histograms, suggesting that perhaps all of these models are flexible enough to capture the main shape characteristics of tissue and blood gray levels.

The models may not fit the histograms perfectly, but are they useful in segmentation? As we mentioned in Section I, this is

the ultimate performance measure of the models. Fig. 11 suggests that in spite of poor goodness-of-fit significance values, the models are able to distinguish between blood and tissue. In fact, using a small 5×5 window gives misclassification rates that are below 1%. In most images, the segmented regions have at least one order of magnitude more pixels than this window size, suggesting that the empirical models are good enough for segmentation.

How is one to resolve the apparent contradiction between goodness-of-fit values of the models and their good classification ability? One possibility is this: recall that the optimal classifier is one that uses the exact probability densities and constructs a decision boundary where the blood and tissue densities are equal. If approximate densities are used, what really matters is how much the decision boundary changes. That is, what matters is how well the empirical distributions approximate the real distribution *near the decision boundary*. Accuracy of approximation far from the boundary is not important. This suggests that although the empirical models may not be very close to the histogram everywhere, they are a good enough fit near the decision boundary to be useful.

A final comment. Although the Gamma distribution seems to have a better overall fit, one may argue that any of the empirical distributions we considered are reasonable for use in segmentation problems. Perhaps that is why many of them are actually used.

G. Model Parameters

Because the Gamma distribution seems to be the best model, we next investigated suitable priors for it. Recall from Section IV that the Gamma distribution has two parameters: shape and scale which are denoted α and β, respectively, in (1) and (2). Let α_t and β_t denote the shape and scale parameters for tissue and α_b and β_b denote the shape and scale parameters for blood. Note that the scale parameters depend on the instrument amplification used to acquire the image. That is, the same image acquired at higher amplification will give scale parameters β_t, β_b multiplied by a scalar factor. On the other hand, the ratio β_t/β_b is independent of the machine gain and the distribution of this ratio is more suitable for use as a prior.

In Fig. 12(a), we show a scatter plot of $\hat{\alpha}_t$ versus $\log(\hat{\beta}_t/\hat{\beta}_b)$, where $\hat{\alpha}_t$, $\hat{\beta}_t$, $\hat{\beta}_b$ are the maximum likelihood estimates of the tissue parameters for the Gamma distribution obtained from the tissue 3 region of interest over the entire data set. The two random variables in Fig. 12(a) are almost uncorrelated (correlation coefficient $= 0.04$), and as a first approximation they can be modeled as independent. Fig. 12(b) shows the histogram of the shape parameter α_t and a Gamma distribution that can serve as a model for its distribution. The parameters of the Gamma distribution are given in Table III. Fig. 12(c) shows the histogram of $\log(\beta_t/\beta_b)$ and a normal distribution that can serve as a model for its distribution. Table III gives the mean and standard deviation of the normal.

Fig. 13(a) and (b) shows the corresponding results for blood. The distribution for $\log(\beta_t/\beta_b)$ is omitted because it is the same as above. The distribution of α_b can be modeled by a Gamma distribution as before, and its parameters are given in Table III.

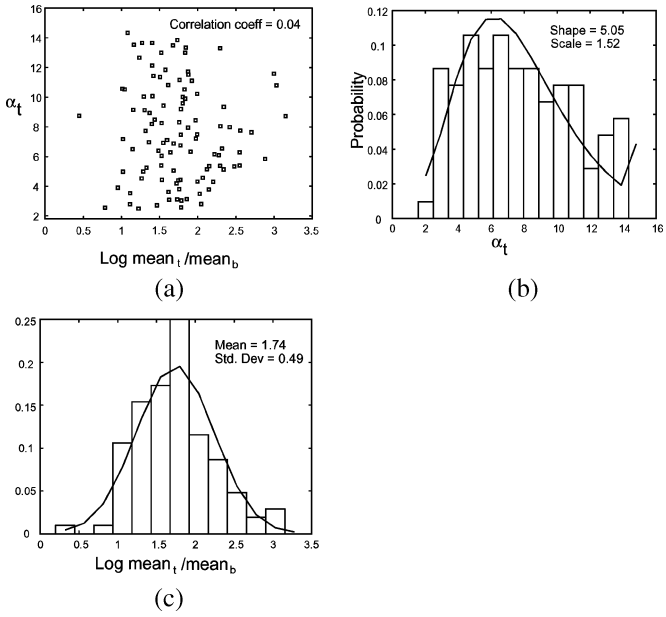


Fig. 12. Distribution of the shape and scale parameters of the Gamma distribution for tissue. (a) Distribution of shape and scale ratio parameters for tissue. (b) Distribution of the shape parameter. (c) Distribution of log scale ratio.

TABLE III
PARAMETER DISTRIBUTIONS

Parameter	Gamma		Normal	
	Shape	Scale	Mean	Std. Dev
α_t	5.68	0.93		
α_b	4.58	0.70		
$\log(\beta_t/\beta_b)$			-0.32	0.99

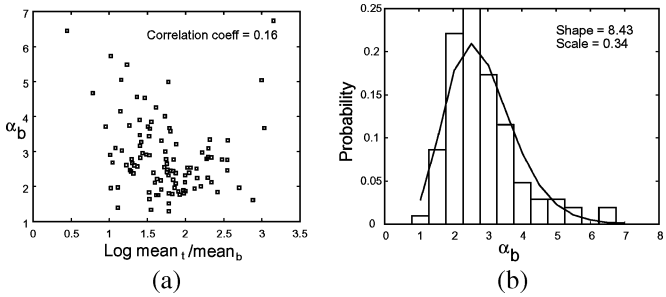


Fig. 13. Distribution of the shape and scale parameters of the Gamma distribution for blood. (a) Estimated shape and scale ratio parameters for blood. (b) Distribution of the shape parameter.

VIII. CONCLUSION

In this paper, we experimentally evaluated four empirical probability models using images obtained from different machines under different operating conditions.

A preliminary investigation revealed a linear relation between the mean and standard deviation of gray levels in all images. This suggested a scalable unimodal probability distribution as a model. We compared the data to four commonly used probability distributions via a modified chi-square goodness-of-fit test using the Rao–Rabson statistic. We also compared the families with respect to their abilities to classify blood an tissue. Although none of the families fit the data at the conventional statistical level of significance of 0.9, the Gamma distribution

had the best fit to the data and also classified blood and tissue at a low misclassification rate. The other three distributions appeared to have comparable fits and acceptable misclassification rates. Finally, an investigation of the parameters of Gamma distribution was also conducted to obtain priors.

ACKNOWLEDGMENT

The authors acknowledge many illuminating discussions with Dr. C. C. Jaffe of the National Cancer Institute, Prof. M. Insana of the University of Illinois at Urbana-Champagne, and Dr. R. Wagner of the Food and Drug Administration.

REFERENCES

[1] U. R. Abeyratne, A. P. Petropulu, and J. M. Reid, "On modeling the tissue response from ultrasonic B-scan images," *IEEE Trans. Med. Imag.*, vol. 15, no. 4, pp. 479–490, Aug. 1996.

[2] J. O. Berger, *Statistical Decision Theory and Baysian Analysis*, ser. Springer Series in Statistics. New York: Springer-Verlag, 1985.

[3] J. S. Bleck, U. Ranft, M. Gebel, H. Hecker, M. Westhoff-Bleck, C. Theismann, S. Wagner, and M. Manns, "Random field models in the textural analysis of ultrasound image of liver," *IEEE Trans. Med. Imag.*, vol. 15, no. 6, pp. 796–801, Dec. 1996.

[4] R. B. D'Agostino and M. A. Stephens, Eds., *Goodness-of-Fit Techniques*. New York: Marcel Dekker, 1986.

[5] C. B. Burckhardt, "Speckle in ultrasound B-mode scans," *IEEE Trans. Son. Ultrason.*, vol. SU-25, no. 1, pp. 1–6, Jan. 1978.

[6] L. Clifford, P. Fitzgerald, and D. James, "Non-Rayleigh first-order statistics of ultrasound backscatter from normal myocardium," *Ultrson. Med. Biol.*, vol. 19, pp. 487–495, 1993.

[7] J. W. Goodman, *Laser Speckle and Related Phenomenon*, J. C. Dainty, Ed. New York: Springer-Verlag, 1975.

[8] J. B. Hampshire, II, J. W. Strohbehn, M. D. McDaniel, J. L. Waugh, and D. H. James, "Probability density of myocardial ultrasound backscatter," in *Proc. 14th Annu. Northeast Bioeng. Conf.*, 1988, pp. 305–308.

[9] M. F. Insana, R. F. Wagner, B. S. Garra, D. G. Brown, and T. H. Shawkar, "Analysis of ultrasound image texture via generlized Rician statistics," *Opt. Eng.*, vol. 25, pp. 743–748, 1986.

[10] M. F. Insana, K. J. Myers, and L. W. Grossman, "Signal modeling for missue mharacterization," in *Handbook Med. Imag.*, M. Sonka and M. J. Fitzpatrick, Eds. Bellingham, WA: SPIE, 2000, vol. 2.

[11] E. Jakeman and R. J. A. Tough, "Generalized K distribution: A statistical model for weak scattering," *J. Opt. Soc. Amer.*, vol. 4, pp. 1764–1772, Sep. 1987.

[12] E. Jakeman and P. N. Pusey, "A model for mon-Raleigh sea echo," *IEEE Trans. Antenn. Propagat.*, vol. AP-24, pp. 806–814, Nov. 1976.

[13] R. C. Molthen, P. M. Shankar, and J. M. Reid, "Characterization of ultrasonic B-scans using non-Rayleigh statistics," *Ultrson. Med. Biol.*, vol. 21, pp. 161–170, 1995.

[14] R. C. Molthen, P. M. Shankar, J. M. Reid, F. Forsberg, E. J. Halpern, C. W. Piccoli, and B. B. Goldberg, "Comparisons of the Rayleigh and K-distribution models using *in vivo* breast and liver tissue," *Ultrson. Med. Biol.*, vol. 24, pp. 93–100, 1998.

[15] R. M. CramBlitt and K. J. Parker, "Generation of non-Rayleigh speckle distribution using marked regularity models," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 46, no. 4, pp. 867–874, Jul. 1999.

[16] R. F. Wagner, S. W. Smith, J. M. Sandrick, and H. Lopez, "Statistics of speckle in ultrasound B-scans," *IEEE Trans. Son. Ultrason.*, vol. SU-30, no. 1, pp. 156–163, Jan. 1983.

[17] R. F. Wagner, M. F. Insana, and D. G. Brown, "Unified approach to the detection and classification of speckle texture in diagnostic ultrasound," *Opt. Eng.*, vol. 25, pp. 738–742, 1986.

[18] R. F. Wagner, M. F. Insana, S. W. Smith, J. M. Sandrik, and H. Lopez, "Statistical properties of radio frequency and envelope-detected signals with applications to medical ultrasound," *J. Opt. Soc. Amer.*, vol. 4, pp. 910–922, 1987.

[19] R. F. Wagner, S. W. Smith, J. M. Sandrik, and H. Lopez, "Statistics of speckle in ultrasound B-scans," *IEEE Trans. Son. Ultrason.*, vol. SU-30, no. 3, pp. 156–163, May 1983.

[20] L. Weng, J. M. Reid, P. M. Shankar, and K. Soetanto, "Ultrasound speckle analysis based on K-distribution," *J. Acoust. Soc. Amer.*, vol. 89, pp. 2992–2995, 1991.

[21] Y. Zimmer, R. Tepper, and S. Akselrod, "A lognormal approximation for the gray level statistics in ultrasound images," in *Proc. World Cong. Med. Phys. Biomed. Eng.*, Chicago, IL, Jul. 2000, pp. 23–28.

[22] V. Dutt and J. F. Greenleaf, "Statistics of the log-compressed echo envelope," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, Jun. 1996.

[23] P. M. Shankar, "A general statistical model for ultrasonic scattering from tissues," *IEEE Trans. Ultrason. Ferroelect. Freq. Control*, vol. 47, no. 3, pp. 727–736, May 2000.

[24] P. M. Shankar and V. L. Newhouse, "Speckle reduction with improved resolution in ultrasound images," *IEEE Trans. Son. Ultrason.*, vol. SU-32, no. 4, pp. 537–543, Jul. 1985.

[25] P. M. Shankar, "Speckle reduction in ultrasound B-scan using weighted averaging in spatial compounding," *IEEE Trans. Ultrason. Ferroelect. Freq. Contr.*, vol. UFFC-33, no. 6, pp. 754–758, Nov. 1986.

[26] J. C. Bamber and C. Daft, "Adaptive filtering for reduction of speckle in ultrasonic pulse-echo images," *Ultrasonics*, pp. 41–44, Jan. 1986.

[27] M. Karaman, M. A. Kutay, and G. Bozdagi, "An adaptive speckle suppression filter for medical ultrasonic imaging," *IEEE Trans. Med. Imag.*, vol. 14, no. 2, pp. 283–291, Jun. 1995.

[28] M. Belohlacek and J. F. Greenleaf, "Detection of cardiac boundaries in echocardiographic images using a customized order statistics filter," *Ultrason. Imag.*, vol. 19, no. 2, pp. 127–137, April 1997.

[29] F. L. Lizzi, E. J. Feleppa, M. Astor, and A. Kalisz, "Statistics of ultrasonic spectral parameters for prostate and liver examinations," *IEEE Trans. Ultrason. Ferroelect. Freq. Control*, vol. 44, no. 4, pp. 935–942, Jul. 1997.

[30] A. Vieli, J. Heiserman, I. Schnittger, and R. L. Popp, "An improved stochastic approach to rf amplitude analysis in ultrasonic cardiac tissue characterization," *Ultrason. Imag.*, vol. 6, pp. 139–151, 1984.

[31] H. J. Huisman and J. M. Thijssen, "An in vivo ultrasonic model of liver parenchyma," *IEEE Trans. Ultrason. Ferroelect. Freq. Contr.*, vol. 45, pp. 739–750, May 1998.

[32] P. M. Shankar, V. A. Dumane, J. M. Reid, V. Genis, F. Forsberg, C. W. Piccoli, and B. B. Goldberg, "Classification of ultrasounic B-mode images of breast masses using Nakagami distribution," *IEEE Trans. Ultrason. Ferrorelect. Freq. Contr.*, vol. 48, no. 2, pp. 569–580, Mar. 2001.

[33] J. C. Dainty, Ed., *Laser Speckle and Related Phenomena*. New York: Springer–Verlag, 1976, vol. 9.

[34] I. B. Ayed, A. Mitchie, and Z. Belhadj, "Multiregion level-set partitioning of synthetic aperture radar images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 793–800, May 2005.

[35] G. Xiao, M. Brady, J. A. Noble, and Y. Zhang, "Segmentation of ultrasound B-mode images with intensity inhomogeneity correction," *IEEE Trans. Med. Imag.*, vol. 21, no. 1, pp. 48–57, Jan. 2002.