interspeech 2008 incorporating SST 08

Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis

Chanwoo Kim, Richard M. Stern

Department of Electrical and Computer Engineering and Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA 15213
{chanwook, rms}@cs.cmu.edu

Abstract

In this paper, we introduce a new algorithm for estimating the signal-to-noise ratio (SNR) of speech signals, called WADA-SNR (Waveform Amplitude Distribution Analysis). In this algorithm we assume that the amplitude distribution of clean speech can be approximated by the Gamma distribution with a shaping parameter of 0.4, and that an additive noise signal is Gaussian. Based on this assumption, we can estimate the SNR by examining the amplitude distribution of the noise-corrupted speech. We evaluate the performance of the WADA-SNR algorithm on databases corrupted by white noise, background music, and interfering speech. The WADA-SNR algorithm shows significantly less bias and less variability with respect to the type of noise compared to the standard NIST STNR algorithm. In addition, the algorithm is quite computationally efficient.

Index Terms: SNR estimation, Gamma distribution, Gaussian distribution

1. Introduction

The estimation of signal-to-noise ratios (SNRs) has been extensively investigated for decades and it is still an active field of research (*e.g.* [1-7]). Reliable SNR estimation can improve algorithms for speech enhancement [1][2], speech detection, and speech recognition [3], since knowledge of SNR makes it easier to compensate for the effects of noise.

Techniques for estimating SNR can be classified into several categories. One of the approaches is based on distinguishing the spectra of noise and speech. Noise spectrum estimation (*e.g.* [3]) or spectral subtraction techniques usually belong to this category. Another approach is based on measurement of the energy. The widely used NIST STNR (Signal-To-Noise-Ratio) algorithm is based on this technique. In this approach, a histogram of short-time energy is constructed, from which the signal and noise energy distributions are estimated. In another approach, Martin [4] used the low-energy envelope in frequency bands to estimate the SNR level. Still other approaches are based on statistics that are obtained from waveform samples rather than from energy or spectral coefficients. For example, Nemer [5] used kurtosis values to estimate the SNR in each frequency band. In this approach, short-time voiced signals in a given frequency band are assumed to be sinusoidal with a fixed phase, and short-time unvoiced signals in this band is assumed to be a sinusoidal signal with random phase.

Our approach is based on the fact that the amplitude distribution of a waveform usually can be characterized by a gamma distribution with a shaping parameter value between 0.4 and 0.5. This fact has been observed by several research groups and has been described in numerous books and papers (*e.g.* [8][9]). The only assumptions we make are that (1) the speech and background noise are independent, (2) clean speech follows a gamma distribution with a fixed shaping parameter, and (3) the background noise has a Gaussian distribution. Based on these assumptions, it can be seen that if we model noise-corrupted speech at an unknown SNR using the Gamma distribution, the value of the shaping parameter obtained using maximum likelihood (ML) estimation depends uniquely on the SNR.

While we assume that the background noise can be assumed to be Gaussian, we will demonstrate that this algorithm still provides better results than the NIST STNR algorithm, even in the presence of other types of maskers such as background music or interfering speech, where the corrupting signal is clearly not Gaussian.

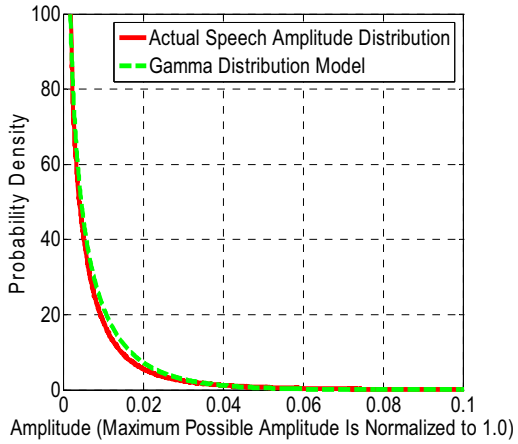
The organization of this paper is as follows: in Sec. 2 we discuss the assumptions about clean speech and additive noise. In Sec. 3 we describe how the SNR measurement can be obtained from the amplitude distribution of the input signal. Section 4 contains experimental results that compare the accuracy of WADA-SNR to the standard NIST STNR algorithm.

2. Characterization of clean speech and additive noise

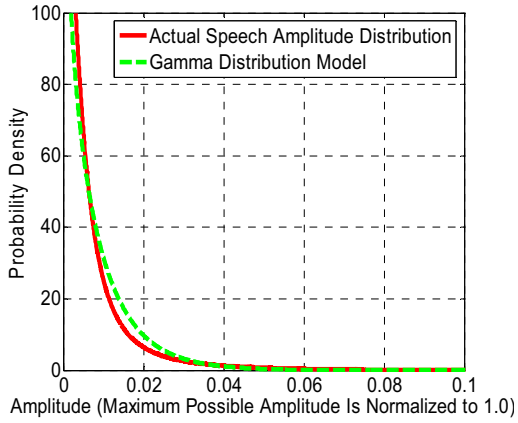
It is widely known that the symmetric gamma distribution is a good approximation to the amplitude distribution of a large speech corpus (*e.g.* [8][9]). Specifically, the probability density function, $f_{\mathbf{x}}(x)$ of clean speech can be represented by the following equation [8-11]:

$$f_{\mathbf{x}}(x\,|\,\beta_{\mathbf{x}}) = \frac{\beta_{\mathbf{x}}}{2\Gamma(\alpha_{\mathbf{x}})}\left(\beta_{\mathbf{x}}\,|\,x\,|\right)^{\alpha_{\mathbf{x}}-1}\exp(-\beta_{\mathbf{x}}\,|\,x\,|) \qquad (1)$$

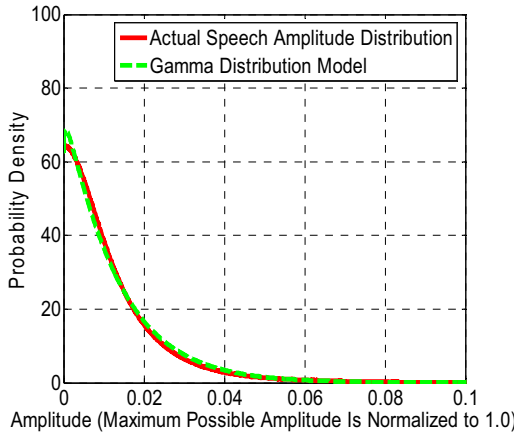
where x is the amplitude of the speech, and $\alpha_{\mathbf{x}}$ and $\beta_{\mathbf{x}}$ are the shaping and rate parameters of the gamma distribution, respectively [10][11]. Fig. 1 (a) illustrates this property. Many research results show that values of 0.4 or 0.5 for $\alpha_{\mathbf{x}}$ provide the best fit for clean speech (*e.g.* [8][9]). We will assume for now that a clean speech signal $x[n]$ exhibits a gamma distribution with a fixed shaping parameter $\alpha_{\mathbf{x}}$ of 0.4 and an arbitrary value of β_x. (The parameter β_x serves to normalize the density function and has no impact on the SNR estimation.) As will be shown later (*cf.* Fig. 2), the SNR value estimated by our algorithm is relatively independent of $\alpha_{\mathbf{x}}$ if the true value of the SNR is less than 20 dB. Throughout this paper, $x[n]$, $v[n]$, and $z[n]$ will denote sample functions for clean speech, noise, and corrupt speech respectively. The variables x, v, and z will denote sample values without regard to time, and $\mathbf{x}$, $\boldsymbol{v}$, and $\mathbf{z}$ will represent the random variables that describe them.



(a) Clean speech



(b) 10-dB additive white Gaussian noise



(c) 0-dB additive white Gaussian noise

Figure 1: *Comparison between the actual amplitude distribution of speech and the gamma distribution model. A subset of 1,600 utterances of the DARPA RM test set was used.*

If a clean speech signal is corrupted by additive Gaussian noise $v[n]$, its probability density function can be expressed as:

$$f_{\mathbf{v}}(v) = \frac{1}{\sqrt{2\pi}\sigma_{\mathbf{v}}} \exp\left(-\frac{v^2}{2\sigma_{\mathbf{v}}^2}\right) \tag{2}$$

where $\sigma_{\mathbf{v}}$ is the standard deviation of the noise.

We will further assume that both $x[n]$ and $v[n]$ have zero means and that they are statistically independent. The

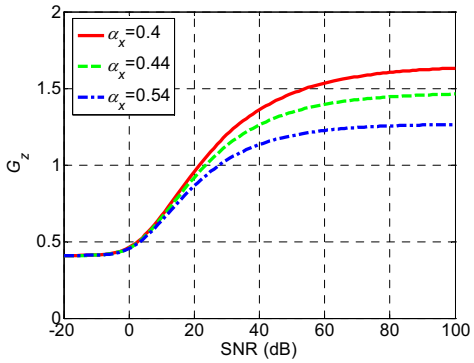


Figure 2: *Calculated dependence of the parameter Gz on the SNR in dB.*

corrupt speech signal $z[n]$ is represented by the following equation:

$$z[n] = x[n] + v[n] \tag{3}$$

From (1) and (2), the power of the speech and noise parts can be obtained from the above distributions using some arithmetic:

$$P_{\mathbf{x}} = \frac{\alpha_{\mathbf{x}}(\alpha_{\mathbf{x}}+1)}{\beta_{\mathbf{x}}^2} \tag{4}$$

$$P_{\mathbf{v}} = \sigma_{\mathbf{v}}^2 \tag{5}$$

where $P_{\mathbf{x}}$ and $P_{\mathbf{v}}$ are the signal and noise power, respectively. Hence, the SNR of this signal $z[n]$ is given by

$$\eta_{\mathbf{z}} = \frac{P_{\mathbf{x}}}{P_{\mathbf{v}}} = \frac{\alpha_{\mathbf{x}}(\alpha_{\mathbf{x}}+1)}{(\sigma_{\mathbf{v}}\beta_{\mathbf{x}})^2} \tag{6}$$

3. SNR measurement based on the gamma distribution

In Figs. 1(a) to 1(c), we observe the amplitude distribution of clean and corrupt utterances. It can be seen that even for noisy speech utterances, the gamma distribution model is still quite close to the actual amplitude distribution. More importantly, we can easily see that the shapes of Figs. 1(b) and 1(c) are very different from that of Fig. 1(a). Hence we observe that the value of the parameter $\alpha_{\mathbf{z}}$ characterizes the amplitude distribution of noisy speech using the Gamma-function model depends on the SNR. From the probability density function of the gamma distribution, we can obtain the following relation for corrupt speech [11]:

$$\ln(\alpha_z) - \psi_0(\alpha_z) = \ln\left(\frac{1}{N}\sum_{n=0}^{N-1}|z[n]|\right) - \frac{1}{N}\sum_{n=0}^{N-1}\ln[|z[n]|] \tag{7}$$

where $\psi_0(\alpha_z)$ is the digamma function.

From the above equation, we can see that the shaping parameter depends on the right hand side of (7). Based on this observation, we define the parameter G_z:

$$G_{\mathbf{z}} = \ln\left(\frac{1}{N}\sum_{n=0}^{N-1}|z[n]|\right) - \frac{1}{N}\sum_{n=0}^{N-1}\ln(|z[n]|) \tag{8}$$

We now show that $G_{\mathbf{z}}$ in (8) can be employed to uniquely determine SNRs. Let us assume that $|z[n]|$ and $\ln(|z[n]|)$ are both ergodic in the mean. For N sufficiently large, we can replace the time averages by their corresponding ensemble averages:

$$\frac{1}{N}\sum_{n=0}^{N-1}|z[n]| = E[|\mathbf{z}|] \tag{9}$$

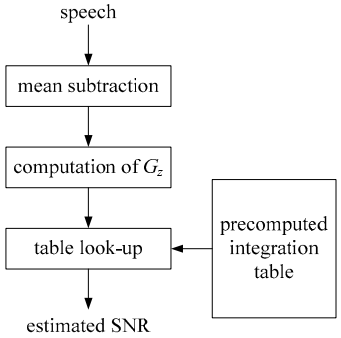


Figure 3: *The structure of the WADA-SNR estimation system*

$$\frac{1}{N}\sum_{n=0}^{N-1}\ln|z[n]| \approx E[\ln|\mathbf{z}|] \tag{10}$$

producing:

$$\begin{aligned} G_{\mathbf{z}} &= \ln(E[|\mathbf{z}|]) - E(\ln(|\mathbf{z}|)) \\ &= \ln(E[|\mathbf{x}+\mathbf{v}|]) - E(\ln(|\mathbf{x}+\mathbf{v}|)) \end{aligned}. \tag{11}$$

Let's consider the following normalized random variables $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{v}}$:

$$\tilde{\mathbf{x}} = \beta_{\mathbf{x}}\mathbf{x}, \tag{12}$$

$$\tilde{\mathbf{v}} = \mathbf{v}/\sigma_{\mathbf{v}} \tag{13}$$

From (1), (2), (12), and (13), we see that the probability densities of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{v}}$ are represented by the following distributions:

$$f_{\tilde{\mathbf{x}}}(\tilde{x}) = f_{\mathbf{x}}(\frac{\tilde{x}}{\beta_{\mathbf{x}}})\frac{dx}{d\tilde{x}} = \frac{1}{2\Gamma(\alpha_{\mathbf{x}})}|\tilde{x}|^{\alpha_{\mathbf{x}}-1}\exp(-|\tilde{x}|) \tag{14}$$

$$f_{\tilde{\mathbf{v}}}(\tilde{v}) = f_{\mathbf{v}}(\sigma_{\mathbf{v}}\tilde{v})\frac{dv}{d\tilde{v}} = \frac{1}{\sqrt{2\pi}}\exp(-\tilde{v}^2) \tag{15}$$

Note that Eqs. (14) and (15) have no free parameters at all. Substituting (12) and (13) into (11), we obtain:

$$\begin{aligned} G_{\mathbf{z}} &= \ln(E[|\tilde{\mathbf{x}} + \beta_x\sigma_{\mathbf{v}}\tilde{\mathbf{v}}|]) \\ &\quad - E(\ln(|\tilde{\mathbf{x}} + \beta_x\sigma_{\mathbf{v}}\tilde{\mathbf{v}}|)). \end{aligned} \tag{16}$$

Combining (6), (14), (15), we represent (16) in integral form:

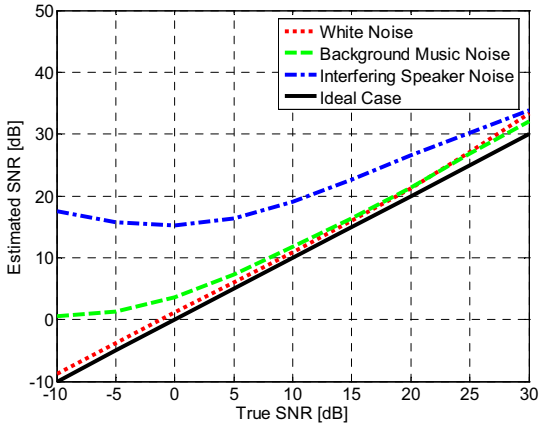
$$G_{\mathbf{z}} = \ln\left(\frac{1}{2\sqrt{2\pi}\Gamma(\alpha_{\mathbf{x}})}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|\tilde{x}+\sqrt{\frac{\alpha_x(\alpha_x+1)}{\eta_{\mathbf{z}}}}\tilde{v}||\tilde{x}|^{\alpha_x-1}\exp(-|\tilde{x}|-\frac{\tilde{v}^2}{2})d\tilde{x}d\tilde{v}\right) \tag{17}$$

$$\frac{1}{2\sqrt{2\pi}\Gamma(\alpha_{\mathbf{x}})}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ln|\tilde{x}+\sqrt{\frac{\alpha_x(\alpha_x+1)}{\eta_{\mathbf{z}}}}\tilde{v}||\tilde{x}|^{\alpha_x-1}\exp(-|\tilde{x}|-\frac{\tilde{v}^2}{2})d\tilde{x}d\tilde{v}$$

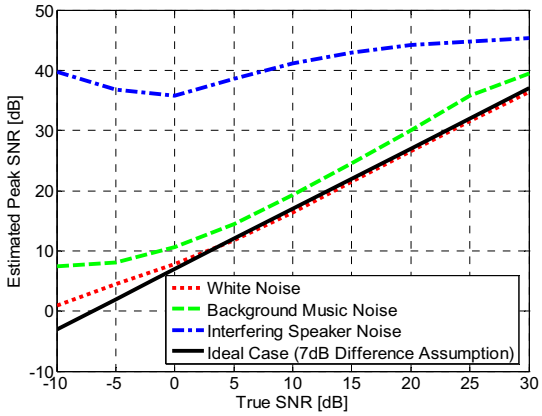
Since $\alpha_{\mathbf{x}}$ is assumed to be the fixed constant 0.4, we see that $G_{\mathbf{z}}$ is uniquely determined for a given SNR $\eta_{\mathbf{z}}$. Hence it can be represented as a function of $\eta_{\mathbf{z}}$ in the following form:

$$G_{\mathbf{z}} = h(\eta_{\mathbf{z}}). \tag{18}$$

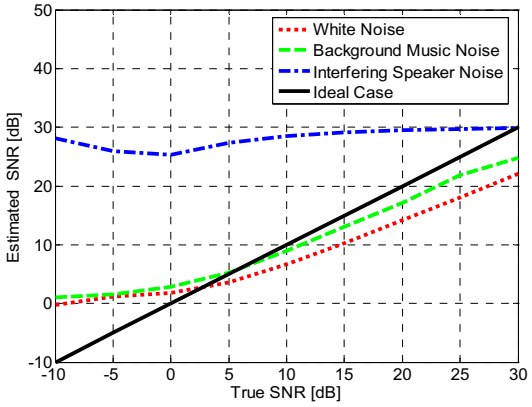
Integration of (17) can be accomplished by numerical Monte-Carlo techniques. Fig. 2 shows the result obtained. In Fig. 2 we observe that the value of G_z is relatively independent of a_x if the true SNR is less than 20 dB. The numerical integration of (17) is computationally intensive, but the calculation can be obtained offline and pre-stored in tabular form. Using the system of Fig. 3 we estimate the SNR based on the relationship between SNR level and G_z implied by (18). Computation of $G_{\mathbf{z}}$ is not very difficult. In some cases, there may be zero values in the utterance, which will cause problems due to the log operation in (8). In practice we either disregard samples with zero values in the computation or we replace them by a small predefined value.



(a) *Results with the WADA-SNR algorithm*

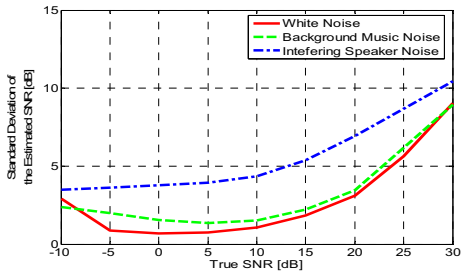


(b) *Results with the NIST STNR algorithm*

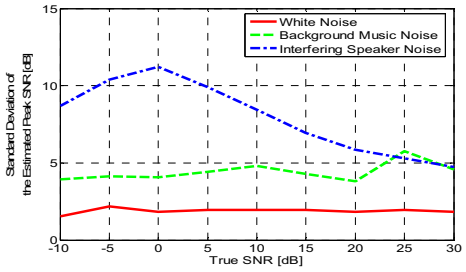


(c) *Result with the modified NIST STNR algorithm*

Figure 4: *Comparison of the average estimated SNR of the NIST STNR algorithm and the WADA-SNR algorithm for the artificially corrupted DARPA Resource Management (RM) database In (c), the mean value of the speech histogram is used instead of the 95 % percentile.*



(a) *Standard deviation of the estimated SNR obtained with the WADA-SNR algorithm*



(b) *Standard deviation of the estimated peak SNR obtained with the NIST STNR algorithm*

Figure 5: *Comparison of the standard deviation of the NIST STNR algorithm and the WADA-SNR algorithm for artificially corrupted DARPA Resource Management (RM) database*

4. Simulation results

The WADA-SNR measurement algorithm described above was evaluated by comparing the estimated SNR with estimates for the same waveforms using the NIST STNR algorithm. The test corpus consists of a subset of 1,600 utterances from the DARPA Resource Management (RM) database. We artificially added noise to the speech in this test set at different SNRs ranging from -10 dB to 30 dB. We used three different types of noise: additive white Gaussian noise, musical segments from the DARPA HUB 4 Broadcast News database, and noise from a single interfering speaker.

Fig. 4 shows the average of the estimated SNRs obtained using the NIST STNR algorithm and the WADA-SNR algorithm. It can be seen that the SNR estimates produced by the WADA-SNR algorithm show both less bias and more consistency with respect to noise type than the corresponding estimates produced by the NIST STNR algorithm.

We note that the NIST STNR measures peak SNR rather than ordinary SNR, in that it determines the difference in decibels between the 95th percentile of the signal power and the 50th percentile of the noise power. For additive Gaussian noise, this overestimates the SNR by about 7 dB.

We modified the NIST STNR algorithm to measure the ordinary SNR as well as peak SNR. We conducted experiments using this modified version of the NIST STNR algorithm as well, as shown in Fig. 4(c). However, we found that if STNR is modified in this way, the bias increases compared to the measured SNR. Fig. 5 shows the standard deviation of the estimated values at each SNR level. As can be seen, if the SNR is less than 20 dB, the WADA-SNR results in much smaller standard deviation. If the noise level is greater than this, the NIST STNR algorithm provides smaller standard deviation than the WADA-SNR algorithm.

5. Conclusions

We introduce a novel approach to the estimation of signal-to-noise ratios (SNRs) called the WADA-SNR algorithm, which is based on statistical information obtained from the amplitude distribution of a speech waveform. Our algorithm is based on the two assumptions that clean speech is characterized by a Gamma distribution with a fixed shaping parameter, and that background noise can assumed to be Gaussian. Even though this algorithm is developed under the assumption of Gaussian noise, it was observed empirically to provide good estimates for background music and background speech as well. The algorithm provides estimates of SNR that are more consistent with respect to background noise type than the NIST STNR algorithm. The only major computational cost incurred is in the estimation of the internal parameter G_z, so processing is quite computationally efficient.

6. Acknowledgments

This research was funded in part by the National Science Foundation (Grant IIS-0420866).

7. References

[1] P. Scalart and J. Vieira Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Atlanta, GA, May 1996, vol. 2, pp. 629-632.

[2] C. Plapous and C. Marro, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 6, pp. 2098-2108, Nov. 2006.

[3] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing,* Detroit, MI, May 1995, vol. 1, pp. 153-156.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing,* vol. 9, no. 5, pp. 504-512, July, 2001.

[5] E. Nemer, R. Goubran and S. Mahmoud, "SNR Estimation of speech signals using subbands and fourth-Order statistics," *IEEE Signal Processing Letters,* vol. 6, no. 7, pp. 171-174, July 1999.

[6] R. Martin, "An efficient algorithm to estimate the instantaneous SNR of speech signals," in *Proc.* Eurospeech, Berlin, Germany, 1993, pp. 1093–1096.

[7] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans Speech Audio Processing*, vol. 11, no. 3, pp. 184-192, May 2003.

[8] M. D. Paez and T. H. Glisson, "Minimum mean-squared-error quantization in speech PCM and DPCM," *IEEE Trans. Commun.*, vol. COM-20, no. 2, pp. 225–230, Apr. 1972.

[9] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals,* Englewood-Cliffs, NJ: Prentice-Hall, Inc., 1978.

[10] R. V. Hogg, A. Craig, and J. W. McKean, *Introduction to Mathematical Statistics*, 6th edition, Upper Saddle River, NJ: Prentice-Hall, 2005.

[11] S. C. Choi and R. Wette, "Maximum likelihood estimation of the parameters of the gamma distribution and their bias," *Technometrics*, vol. 11, no. 4, pp. 683-690, Nov. 1969.