FOUNDATIONS OF SPECTRAL-GAIN FORMULAE FOR SPEECH NOISE REDUCTION

Eric J. Diethorn

ejd@avaya.com
Avaya, Inc., 233 Mt. Airy Road, Basking Ridge, NJ, 07920

ABSTRACT

Over the past years, methods of short-time spectral modification have proven especially effective at reducing noise for the enhancement of speech. Key to these methods are noise-reduction gain formulae, a class of frequency-domain data windows, that are applied to successive blocks of the short-time Fourier transform of the noisy speech. This paper provides a unified treatment of several noise-reduction gain formulae from the context of statistical estimation theory, and in doing so attempts to illuminate relationships shared by the most common gain functions found in the literature.

1. INTRODUCTION

The noise reduction problem is one of recovering a speech signal of interest $s(n)$ from the noisy observation

$$y(n) = s(n) + v(n) \ , \qquad (1)$$

where $v(n)$ represents unwanted additive noise. Signal and noise are treated as sample functions associated with underlying stationary, zero-mean, uncorrelated random processes. In this case, the power spectral density of $y(n)$, $P_y(\omega) = \sum_m R_y(m) e^{-j\omega m}$, is given by

$$P_y(\omega) = P_s(\omega) + P_v(\omega) \ , \qquad (2)$$

where $R_y(m) = E\{y(n)y^*(n+m)\}$ is the covariance associated with $y(n)$, $E\{\cdot\}$ denotes expectation, $P_s(\omega)$ and $P_v(\omega)$ are power spectral densities, and ω denotes Fourier frequency. Because $P_y(\omega)$ can be estimated from the observation, if $P_v(\omega)$ is known or can be sufficiently estimated, the power spectral density of $s(n)$ can be estimated by rearranging (2) to yield

$$P_s(\omega) = P_y(\omega) - P_v(\omega) \ . \qquad (3)$$

To estimate $s(n)$ we may, equivalently, estimate its discrete Fourier transform $S(\omega) = \sum_n s(n) e^{-j\omega n}$. Suppose we wish to do so by applying a linear filter $H(\omega)$ to the noisy observation, that is, by forming an estimate $\hat{S}(\omega)$ given by

$$\hat{S}(\omega) = H(\omega) Y(\omega) \ . \qquad (4)$$

And, consider the following improvised filter:

$$H(\omega) = \frac{P_y(\omega) - P_v(\omega)}{P_y(\omega)} \ . \qquad (5)$$

This is simply the right-hand side of the recovery relation in (3) normalized to result in a unit-less frequency-dependent gain function. What does this gain function do? First, because $P_s(\omega)$ and $P_v(\omega)$ are related through (2), $H(\omega)$ satisfies $0 \le H(\omega) \le 1$. In the absence of noise $P_v(\omega) = 0$, and so $H(\omega) = 1$ as is desired. When noise is present and $P_v(\omega) >> P_s(\omega)$, $H(\omega) << 1$ and tends to zero as $P_v(\omega) / P_s(\omega)$ increases. Consequently, at frequencies ω where the signal dominates the noise, $\hat{S}(\omega)$ is nearly $S(\omega)$; and at frequencies where the noise dominates, $\hat{S}(\omega)$ is close to zero. The improvised filter in (5) therefore provides near recovery of the signal at least at frequencies where the noise is absent or nearly so. (Note that (5) provides perfect recovery of $S(\omega)$ when $P_s(\omega)$ and $P_v(\omega)$ share no frequencies of common support.) Some will recognize (5) as the frequency domain form of the Wiener filter, a type of optimum estimator. In this work we are interested in the general form in (4), in which the (noise-reduced) signal estimate is produced by applying a filter, or frequency domain data window, to the noisy spectrum.

2. FREQUENCY DOMAIN FORMULATION

The majority of work in noise reduction has focused on methods of short-time spectral modification as a means of reducing noise in speech time series. This general technique, which uses the short-time Fourier transform (STFT), is both computationally simple and, more importantly, relatively effective in enhancing speech. Specifically, any time series $x(n)$, stationary or otherwise, can be represented by its short-time Fourier transform (STFT)

$$X(k,m) = \sum_{n=0}^{N-1} w(n) x(m-n) e^{-j2\pi kn/K} \ , \qquad (6)$$

where $w(n)$ is the analysis window, N dictates the duration or window of time over which $x(n)$ is considered stationary, K is the number of frequencies at which the STFT is computed, and $-\infty < m < \infty$ and $k = 0,...,K-1$ are indices of time and frequency, respectively. Noise reduction is achieved through a short-time spectral gain function applied to successive blocks of the STFT. Ideally, if the noise can be eliminated from the STFT of $y(n)$, the signal estimate $\hat{s}(n)$ can be recovered through an appropriate synthesis procedure.

For the analysis presented in the remainder of this paper, we consider the frequency domain dual of (1). Let,

$$Y = S + V \qquad (7)$$

denote the noisy spectral sample where, for given k and m, $Y = Y(k,m)$, $S = S(k,m)$ and $V = V(k,m)$. We assume S

and V are zero-mean, independent, complex Gaussian random processes. Though these conditions are not strictly true for the short-time Fourier transform spectral representations used in the application of speech noise reduction, they are neither unusual nor problematic assumptions. A real-valued Gaussian time series passed through a bandpass-filter-demodulator structure results in a complex-valued Gaussian with independent real and imaginary components [1]. This same structure is used in any noise reduction scheme employing block frequency-domain processing. It is also assumed that, for any $k \neq l$, $S(k,m)$ and $S(l,m)$ are independent and $V(k,m)$ and $V(l,m)$ are independent. The importance of this assumption in the derivation of noise reduction gain formulas is that the problem of optimally estimating the entire signal waveform can be decomposed into multiple independent problems of estimating the spectral components of the signal.

2.1. Minimum mean-squared-error spectral (MMSES) estimate

We consider first the MMSE estimate of the complex spectral sample S in (7). The following derivation is based on Van Trees' presentation of Bayesian estimation theory [2].

The MMSE estimate of S is that $\hat{S}$ which minimizes the mean-squared-error Bayes "cost"

$$C(S,\hat{S}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(S-\hat{S})(S-\hat{S})^{*}\,p(S_r,S_i\,|\,Y)\,dS_r dS_i \qquad (8)$$

where $p(S_r,S_i\,|\,Y)$, or simply $p(S\,|\,Y)$, is the conditional, or posteriori, probability of S given the observation Y. In (8), S has been decomposed as $S = S_r + jS_i$ to emphasize that $p(S\,|\,Y)$ is a joint probability defined over the real and imaginary parts of S. Because $p(S\,|\,Y)$ is nonnegative, (8) is a convex function of $\hat{S}$. Computing the complex gradient of (8) with respect to $\hat{S}$ and setting the result equal to zero to arrive at the minimum gives

$$\hat{S}_{\mathrm{MMSES}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S\,p(S_r,S_i\,|\,Y)\,dS_r dS_i \qquad (9)$$

as the optimum estimate. The MMSE estimate is recognized as the conditional mean of S, a result that is fundamental to signal estimation theory.

To compute (9), the conditional density $p(S\,|\,Y)$ is needed. Assume signal and noise are (complex) Gaussian with probability density functions [1, 3]

$$p(S) = \frac{1}{\pi\sigma_S^2}\exp\left[-\frac{|S|^2}{\sigma_S^2}\right],\; p(V) = \frac{1}{\pi\sigma_V^2}\exp\left[-\frac{|V|^2}{\sigma_V^2}\right], \quad (10)$$

where σ_S^2 and σ_V^2 are the respective variances, or powers. Using Bayes' Rule, any valid conditional probability density satisfies $p(S\,|\,Y)p(Y) = p(Y\,|\,S)p(S)$. Consequently, $p(S\,|\,Y) = p(Y\,|\,S)p(S)/p(Y)$. Now, $p(Y\,|\,S)$ is the probability of Y occurring given that S has occurred. From (7), with S constant, Y is seen to have the density of V but with mean S, that is, $p(Y\,|\,S) = p(V)\,|_{V=Y-S}$. Combining $p(Y\,|\,S)$ with $p(S)$ in (10) through Bayes' Rule gives

$$p(S\,|\,Y) = \frac{1}{p(Y)\pi^2\sigma_V^2\sigma_S^2}\exp\left[-\frac{1}{\sigma_Z^2}\left|S-\frac{\sigma_S^2}{\sigma_S^2+\sigma_V^2}Y\right|^2\right] \quad (11)$$

where $\sigma_Z^2 = \sigma_S^2\sigma_V^2/(\sigma_S^2+\sigma_V^2)$. Because $\int_{-\infty}^{\infty} p(S\,|\,Y)dS$ is, for any valid conditional density, equal to unity, $p(Y)$ occurring in the denominator of (11) serves only as a normalization factor. Furthermore, since (11) itself is recognized as a Gaussian probability density function, the mean of $p(S\,|\,Y)$ can be read directly from (11) and is, from the discussion above, equal to the MMSE estimate of S. Thus,

$$\hat{S}_{\mathrm{MMSES}} = \frac{\sigma_S^2}{\sigma_S^2+\sigma_V^2}\,Y \quad . \qquad (12)$$

Since S and V are uncorrelated, (7) implies $\sigma_Y^2 = \sigma_S^2 + \sigma_V^2$. So, (12) can be rewritten

$$\hat{S}_{\mathrm{MMSES}} = H_{\mathrm{MMSES}}Y \;,\; H_{\mathrm{MMSES}} = \frac{\sigma_Y^2-\sigma_V^2}{\sigma_Y^2} \quad . \quad (13)$$

The form in (13) is identical to the Wiener linear filter in (5). This is because, for Gaussian quantities, the MMSE waveform estimate is always a linear function of the noisy observation and so the two optimum MMSE solutions coincide [2].

2.2. Maximum a-posteriori spectral estimate

Also fundamental to estimation theory is the maximum a posteriori (MAP) estimate. The MAP estimate of the signal in (7) is that S which is most likely to occur given that the noisy observation has already occurred. In other words,

$$\hat{S}_{\mathrm{MAP}} = \arg\max_{S} p(S\,|\,Y) \quad . \qquad (14)$$

The maximum occurs at the peak of the conditional probability density. From (11), $p(S\,|\,Y)$ achieves its peak when the exponential's argument equals zero, and this occurs at the conditional mean, that is, at (12). Thus, the MAP estimate of the signal component coincides with the MMSES estimate. This association always occurs, as long as density $p(S\,|\,Y)$ is unimodal (single maximum) and symmetric about its mode [2].

2.3. Maximum likelihood spectral-power (MLSP) estimate

By assuming Gaussian processes for the signal and noise, Y in (7) is also Gaussian and has density

$$p(Y) = \frac{1}{\pi(\sigma_S^2+\sigma_V^2)}\exp\left[-\frac{|Y|^2}{(\sigma_S^2+\sigma_V^2)}\right] \quad . \qquad (15)$$

Assuming σ_V^2 is known (in practice it can be estimated well), the variance of the signal is all that is necessary to fully describe the density of Y. Consider the problem of estimating σ_S^2 from observation Y and the known noise variance σ_V^2. If the signal variance in (15) is viewed as a conditioning parameter, that is, by writing $p(Y) = p(Y\,|\,\sigma_S^2)$, the maximum-likelihood estimate [2] of σ_S^2 is that which maximizes (15) given observation Y. Eq. (15) is a convex function of σ_S^2;

taking its derivative with respect to σ_S^2 and setting the result equal to zero gives

$$\hat{\sigma}_{S,\mathrm{MLSP}}^2 = |Y|^2 - \sigma_V^2 \qquad (16)$$

as maximum-likelihood estimate of the signal power. This variance estimate can be used to construct an estimate of the S itself. Consider

$$\hat{S} = \frac{\sigma_S}{|Y|} Y \ . \qquad (17)$$

The mean and variance of (17) are easily shown to match those of S, so (16) is consistent with respect to the second-order statistics of the signal of interest. Using (16) in (17) results in the maximum-likelihood spectral-power estimate of the signal

$$\hat{S}_{\mathrm{MLSP}} = H_{\mathrm{MLSP}} Y \ , \ H_{\mathrm{MLSP}} = \frac{\sqrt{|Y|^2 - \sigma_V^2}}{|Y|} \ . \qquad (18)$$

Some may recognize (18) as the so-called power-subtraction method of noise reduction. The association of this maximum-likelihood problem with the power-subtraction method of noise reduction was presented by McAulay and Malpass [4].

2.4. Maximum likelihood spectral-amplitude (MLSA) estimate

To review, the maximum-likelihood spectral-power estimate is derived from the maximum-likelihood estimate of the signal's variance, or power, given the noisy observation and noise variance, while the MMSE spectral estimate is optimal as an estimate of the (complex) signal itself.

Seeking yet another estimate, McAulay and Malpass [4] also considered a maximum likelihood estimate of the signal's spectral magnitude or amplitude. Representing S in terms of its amplitude and phase components, the noise reduction problem (7) is recast as

$$Y = Ae^{j\phi} + V \ . \qquad (19)$$

Motivation for estimating the envelope of the signal comes from knowledge that human speech perception is less sensitive to phase corruption as it is to corruption of the speech envelope. In this sense, estimation of the phase is secondary to that of the spectral amplitude.

In this approach to estimating the spectral amplitude, A is considered an unknown yet deterministic parameter, while ϕ is treated as a random quantity. The maximum likelihood estimate of the amplitude is that A which maximizes the probability of the observation Y occurring, that is,

$$\hat{A}_{\mathrm{MLSA}} = \arg\max_A p(Y \mid A) \ . \qquad (20)$$

Recognizing that Y is a function of both amplitude and phase, $p(Y \mid A)$ is found from the conditional density of Y given both A and ϕ, $p(Y \mid A, \phi)$, by removing the contribution due to the phase. We have

$$p(Y \mid A) = \int_{-\pi}^{\pi} p(Y \mid A, \phi) p(\phi) d\phi \ . \qquad (21)$$

Now, $p(Y \mid A, \phi)$ is given by the density of V but with a mean determined by the fixed but unknown signal component,

$$p(Y \mid A, \phi) = p(V)\,|_{V = Y - Ae^{j\phi}} = \frac{1}{\pi\sigma_V^2} \exp\left[-\frac{|\, Y - Ae^{j\phi}\,|^2}{\sigma_V^2} \right] \ . \qquad (22)$$

Using (22) in (21), and, lacking additional information, assuming a uniform density for the phase, gives

$$p(Y \mid A) = \frac{1}{\pi\sigma_V^2} e^{-\frac{1}{\sigma_V^2}(|Y|^2 + A^2)} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-2A\,\mathrm{Re}\{Ye^{-j\phi}\}/\sigma_V^2} d\phi \qquad (23)$$

Recognizing the trailing integral in (23) as the zeroth-order modified Bessel function, $I_0(2A|Y|/\sigma_V^2)$, (23) reduces to

$$p(Y \mid A) = \frac{1}{\pi\sigma_V^2} e^{-\frac{1}{\sigma_V^2}(|Y|^2 + A^2)} I_0(2A|Y|/\sigma_V^2) \ . \qquad (24)$$

It is this density that we wish to insert into (20). Unfortunately, doing so does not result in a closed-form relation since the Bessel function is not reducible. For $|x| >> 1$, however, $I_0(|x|) \approx \exp(|x|)/\sqrt{2\pi|x|}$ [5], and the approximation improves as $|x|$ increases. Applying this approximation to (24) gives

$$p(Y \mid A) \approx \frac{1}{2\pi\sigma_Y\sqrt{\pi A|Y|}} e^{-\frac{1}{\sigma_V^2}\left(|Y|^2 - 2A|Y| + A^2\right)} \qquad (25)$$

for $2A|Y|/\sigma_V^2 > 1$. Eq. (25) is a convex function of A; differentiating with respect to A and setting the result to zero to extract the peak gives

$$\hat{A}_{\mathrm{MLSA}} = \frac{|Y| + \sqrt{|Y|^2 - \sigma_V^2}}{2} \qquad (26)$$

as the (approximate) maximum likelihood estimate of the signal's amplitude. Eq. (26) is seen to be an arithmetic average of the modulus of the observation and an estimate of the modulus of the desired signal. Note that, in the absence of noise, $\sigma_V^2 = 0$ and so $\hat{A}_{\mathrm{MLSA}} = |Y| = |S|$, as desired. An estimate of the desired signal waveform is constructed by appending the phase of Y to (26). This results in the estimate

$$\hat{S}_{\mathrm{MLSA}} = H_{\mathrm{MLSA}} Y \ , \ H_{\mathrm{MLSA}} = \left[\frac{1}{2} + \frac{1}{2}\sqrt{1 - \frac{\sigma_V^2}{|Y|^2}} \right] \qquad (27)$$

2.5. MMSE spectral-amplitude (MMSESA) estimate

The maximum-likelihood spectral-amplitude estimate was derived using a model in which the amplitude is considered deterministic while the phase is treated as random. Alternatively, the spectral amplitude can be modeled as a random quantity, like the phase, and estimated within the MMSE framework.

For the reasons discussed in section 2.1, the MMSE estimate of the (now random) spectral amplitude A is given by the conditional mean

$$\hat{A}_{\mathrm{MMSESA}} = \int_0^\infty A\, p(A \mid Y) dA \ . \qquad (28)$$

Using Bayes' Rule, the conditional probability $p(A \mid Y)$ satisfies $p(A \mid Y) = p^{-1}(Y) \int p(Y \mid A,\phi) p(A,\phi) d\phi$, where $p(A,\phi)$ is the joint density of A and ϕ. Applying this in (28) gives

$$\hat{A}_{\mathrm{MMSESA}} = \frac{\int_0^\infty \int_0^{2\pi} A\, p(Y \mid A,\phi) p(A,\phi) d\phi dA}{\int_0^\infty \int_0^{2\pi} p(Y \mid A,\phi) p(A,\phi) d\phi dA} \qquad (29)$$

where denominator $p(Y)$ has been expanded in terms of the conditional density $p(Y \mid A,\phi)$ to expose like terms. Conditional density $p(Y \mid A,\phi)$ is given in (22). Regarding $p(A,\phi)$, under the aforementioned signal model in which the real and imaginary components of S are independent complex Gaussian random variables, the joint density $p(A,\phi)$ is separable, i.e., $p(A,\phi) = p(A)p(\phi)$, where $p(A)$ is a Rayleigh density and $p(\phi)$ is a uniform density. These densities are given by [2, 3] (for $A > 0$, $-\pi \le \phi \le \pi$).

$$p(A) = 2A \Big/ \sigma_s^2 \exp\left[- A^2 \Big/ \sigma_s^2 \right], \ \ p(\phi) = \frac{1}{2\pi} \qquad (30)$$

Using (30), both numerator and denominator in (29) are reduced by first evaluating the integral with respect phase as was done in (21). The result is

$$\hat{A}_{\mathrm{MMSESA}} = \frac{\int_0^\infty A^2 e^{-\frac{1}{\sigma_V^2}\left[|Y|^2 + \frac{\sigma_S^2+\sigma_V^2}{\sigma_S^2} A^2 \right]} I_0(2A \mid Y \mid / \sigma_V^2) dA}{\int_0^\infty A e^{-\frac{1}{\sigma_V^2}\left[|Y|^2 + \frac{\sigma_S^2+\sigma_V^2}{\sigma_S^2} A^2 \right]} I_0(2A \mid Y \mid / \sigma_V^2) dA} \qquad (31)$$

With manipulation, one can show that the numerator and denominator of this expression are in the form of the second and first moments, respectively, of the Rician density function [6, 7]. Ephraim and Malah [8] have shown that (31) reduces to

$$\hat{A}_{\mathrm{MMSESA}} = \frac{\sqrt{\pi \upsilon}}{2\gamma} e^{-\upsilon/2} \left[(1+\upsilon) I_0(\frac{\upsilon}{2}) + \upsilon I_1(\frac{\upsilon}{2}) \right] \mid Y \mid , \qquad (32)$$

where $I_1(\cdot)$ is the first-order modified Bessel functional, $\upsilon = \gamma\xi/(1+\xi)$, and where $\xi = \sigma_S^2 / \sigma_V^2$ and $\gamma = \mid Y \mid^2 / \sigma_V^2$ are, respectively, the a priori and a posteriori signal-to-noise ratios. To arrive at an estimate of S, the optimum MMSE amplitude is appended to the noisy phase, as in (27), resulting in

$$\hat{S}_{\mathrm{MMSESA}} = H_{\mathrm{MMSESA}} Y \ , \ H_{\mathrm{MMSESA}} = \frac{\hat{A}_{\mathrm{MMSESA}}}{|Y|} \qquad (33)$$

Thorough treatment of this estimator can be found in [8].

2.6. Comparison of gain formulas

A plot of several noise reduction gain formulas is shown in Fig. 1. All gain functions are plotted as a function of a priori signal-to-noise ratio ξ. Where a posteriori SNR γ is required, the substitution $\gamma = \xi + 1$ has been made.

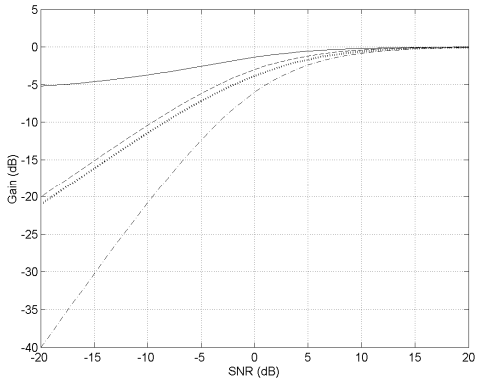

Figure 1. *Gain formulae as a function of a priori input signal-to-noise ratio. Top to bottom are the MLSA, MLSP, MMSESA and MMSES (aka Wiener) estimates.*

2.7. Discussion

For the most part, the noise reduction gain formulas discussed here can be derived using classical detection and estimation principles formalized by Norbert Wiener, with subsequent specific contributions by Rice [6], while at Bell Laboratorties in the 1950s, and Middleton [7]. Much of this theory is presented in seminal texts by Van Trees [2] and Middleton [7].

3. REFERENCES

[1] F. D. Neeser and J. L. Massey, "Proper Complex Random Processes with Applications to Information Theory," *IEEE Trans. Inform. Theory,* vol. 39, No. 4, July 1993.

[2] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I,* New York: John Wiley & Sons. 1968.

[3] R. Arens, "Complex Processes for Envelopes of Normal Noise," *IRE Trans. Infom. Theory*, vol. IT-3, pp. 204-207, Sept. 1957.

[4] R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. ASSP-28, No. 2, April 1980.

[5] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions,* New York: Dover Publications, Inc.

[6] S. O. Rice, "Statistical Properties of a Sinewave Plus Random Noise", *Bell System Tech. J.,* pp. 109-157, Jan. 1948.

[7] D. Middleton, *An Introduction to Statistical Communication Theory,* New York: McGraw-Hill Book Company. 1960.

[8] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. ASSP-32, No. 6, Dec. 1984.