A GENERALIZED STEIN'S ESTIMATION APPROACH FOR SPEECH ENHANCEMENT BASED ON PERCEPTUAL CRITERIA

Sunder Ram Krishnan and Chandra Sekhar Seelamantula

Indian Institute of Science, Bangalore
Department of Electrical Engineering
Bangalore-560012, India

sunder@ee.iisc.ernet.in, chandra.sekhar@ieee.org

ABSTRACT

We address the problem of speech enhancement using a risk-estimation approach. In particular, we propose the use the Stein's unbiased risk estimator (SURE) for solving the problem. The need for a suitable finite-sample risk estimator arises because the actual risks invariably depend on the unknown ground truth. We consider the popular mean-squared error (MSE) criterion first, and then compare it against the perceptually-motivated Itakura-Saito (IS) distortion, by deriving unbiased estimators of the corresponding risks. We use a generalized SURE (GSURE) development, recently proposed by Eldar for MSE. We consider dependent observation models from the exponential family with an additive noise model, and derive an unbiased estimator for the risk corresponding to the IS distortion, which is non-quadratic. This serves to address the speech enhancement problem in a more general setting. Experimental results illustrate that the IS metric is efficient in suppressing musical noise, which affects the MSE-enhanced speech. However, in terms of global signal-to-noise ratio (SNR), the minimum MSE solution gives better results.

Index Terms— Stein's unbiased risk estimator (SURE), perceptual distortion metrics, generalized SURE (GSURE), speech enhancement.

1. INTRODUCTION

Charles M. Stein, while addressing the problem of estimating the mean of a multivariate normal distribution using an independent and identically distributed (i.i.d.) assumption [1], derived an unbiased estimator of the MSE based on a lemma, the proof of which relies on an identity satisfied by the Gaussian density and integration by parts. It is a remarkable result in statistics, since he also proved that the resulting shrinkage-type estimator of the mean would dominate the classical least-squares (LS) estimator, provided the number of data samples is greater than equal to three. In the realm of statistics, cost functions such as the MSE are referred to as *risks*, and therefore, such an estimator of the risk is referred to as Stein's unbiased risk estimator (SURE). SURE of the MSE (which is quadratic, and therefore mathematically tractable) has been used in image and speech processing for optimally determining the parameters of the denoising algorithm involved in the application [2–9].

Stein's original formalism is based on an i.i.d. Gaussian model [1], which was later extended to other, more general cases. Hudson [10] considered certain density models falling within the exponential family, together with the i.i.d. assumption. Again, developments for improving upon inadmissible estimators considering random variables within continuous and discrete exponential families, were reported in [11] and [12], respectively. Raphan and Simoncelli provided a generalization of SURE for the Gaussian case, showing the derivation of a SURE-optimal parametric LS estimator [13]. Eldar [7] has extended the SURE principle to any density model within the exponential family of densities, without assuming independence. This provides a favorable situation, as we can potentially apply Stein's principle to a variety of denoising problems. However, the risk considered is the classical MSE.

Using a point-wise linear denoising function referred to as the modified James Stein (MJS) estimator, Muraka and Seelamantula developed an unbiased estimator for the IS risk in the i.i.d. Gaussian case [8]. The derivation of unbiased estimators for highly non-linear and non-quadratic perceptual distortion functions is not straightforward. However, they showed that with a high SNR assumption, a Taylor series development of the distortion metric suitably truncated, followed by a recursive form of Stein's lemma, enables one in deriving an unbiased estimator for the non-quadratic IS distortion. This proved to be a significant development in the risk-estimation approach to speech enhancement, as perceptual distortion measures are more appealing for the speech enhancement problem as compared with the popular squared error distortion. In this paper, after briefing the reader on SURE theory for the classical MSE, using an additive noise model, we show a detailed theoretical development of an unbiased estimator for the IS measure under a general setting. That is, without assuming any particular functional form for the denoising function, we focus on general, not necessarily i.i.d. observation models from within the exponential family, and derive the estimator of the risk. We satisfy ourselves that the derived risk estimator reduces to that derived in [8], for the i.i.d. Gaussian, MJS estimator case. Following this, we validate our theoretical findings with some experiments with correlated Gaussian noise, which confirm the superior performance of the IS-based enhancement algorithm with respect to musical noise suppression. In terms of global SNR, the MSE-based denoising algorithm is observed to provide better results. Denoising is performed in the discrete cosine transform (DCT) domain as in [8].

1.1. Organization of the paper

In Section 2, we present the problem statement. In Section 3, we detail the GSURE development for the classical MSE. The detailed derivation of the unbiased estimator for the IS-risk assuming a general observation model is shown in Section 4 in which, we also compare our theoretical results with those in [8]. We draw the concluding remarks in Section 5.

1.2. Notations

We use bold-face lower case letters to denote vector quantities, bold-face upper case letters to denote matrices, and v_n to denote n^{th} component of a vector $\mathbf{v}$.

2. PROBLEM STATEMENT

We assume that the observation vector $\mathbf{x}$ has a pdf parameterized by $\boldsymbol{\theta}$ as follows:

$$p(\mathbf{x}\,;\boldsymbol{\theta}) = a(\mathbf{x})\exp\left(\boldsymbol{\theta}^T\psi(\mathbf{x}) - b(\boldsymbol{\theta})\right), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^N$ denotes the noisy speech, and $\boldsymbol{\theta} \in \mathbb{R}^N$ is the clean speech. We assume an additive noise model as follows:

$$x_n = \theta_n + w_n\,;\; n = 1, 2, 3, \cdots, N, \tag{2}$$

with $\mathbf{w}$ being a random vector, whose pdf belongs to the exponential family. Our problem is to estimate $\boldsymbol{\theta}$ from the noisy observation vector $\mathbf{x}$, by minimizing a chosen distortion metric. Our approach is one of risk-minimization, which is outlined below.

2.1. Our approach

The basic idea of our approach is schematically presented in Figure 1. We would like to emphasize that the index n in (2) need not necessarily refer to time. Indeed, our risk-estimation based denoising algorithm acts in the DCT domain. The fundamental idea is that, since the actual distortion functions are dependent on the ground truth, we could minimize unbiased, finite-sample estimators of the risks corresponding to the distortion measures. Before detailing our approach, as an example, consider the case when $\mathbf{w}$ in (2) is a zero mean Gaussian random vector with covariance matrix $\mathbf{C}$, which is positive definite. Writing out the pdf of $\mathbf{x}$ in this case, and comparing with (1), we get:

$$a(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{C})}}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}\right), \tag{3}$$

$$\psi(\mathbf{x}) = \mathbf{C}^{-1}\mathbf{x}, \tag{4}$$

$$b(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^T\mathbf{C}^{-1}\boldsymbol{\theta}. \tag{5}$$

In (1), from the Neyman-Fischer factorization theorem [14], we note that $\mathbf{t} \triangleq \psi(\mathbf{x})$ is sufficient for estimating $\boldsymbol{\theta}$. Therefore, the denoising function chosen should be a function of $\mathbf{t}$ alone, as otherwise we can always condition it on the sufficient statistic to arrive at a dominating estimator, as suggested by the Rao-Blackwell theorem [14]. We denote the estimate of θ_n by $f_n(\mathbf{t})$. Now, our objective is to optimize the estimator form f_n by minimizing the risk corresponding to a distortion d:

$$\mathcal{R}_d = \mathcal{E}\left\{d(\theta_n, f_n(\mathbf{t}))\right\}. \tag{6}$$

In this work, we consider two forms for d as follows:

$$d_{\text{SE}}(\theta_n, f_n(\mathbf{t})) = (\theta_n - f_n(\mathbf{t}))^2, \quad \text{and} \tag{7}$$

$$d_{\text{IS}}(\theta_n, f_n(\mathbf{t})) = \frac{f_n(\mathbf{t})}{\theta_n} - \log\left(\frac{f_n(\mathbf{t})}{\theta_n}\right) - 1. \tag{8}$$

(7) is the classical squared-error distortion, and (8) expresses the perceptually-motivated IS distortion. Since these measures depend upon the unknown θ_n, we propose to derive unbiased estimators of the risks corresponding to (7) and (8), and minimize the risk estimators instead of the actual risks. We show our theoretical developments in a general framework, though the experimental results are shown for the correlated Gaussian case. We note that our results hold true for any additive noise model, with the noise vector pdf belonging to the class of exponential densities. In particular, for speech enhancement, one could specialize our results to a double-exponential observation model. The development for d_{SE} shown in the following section, is based on [7].

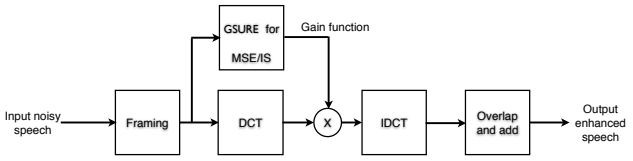


Fig. 1. Block diagram representation of the GSURE-approach to speech enhancement.

3. GSURE FOR d_{SE}

In this case, we would like to obtain the optimum form for $f_n(\mathbf{t})$ by minimizing the risk given by:

$$\begin{aligned}\mathcal{R}_{\text{SE}} &= \mathcal{E}\left\{(f_n(\mathbf{t}) - \theta_n)^2\right\} \\ &= \mathcal{E}\left\{f_n^2(\mathbf{t}) - 2f_n(\mathbf{t})\theta_n + \theta_n^2\right\}.\end{aligned} \tag{9}$$

Note that the expectation is with respect to the density of $\mathbf{t}$. As discussed in [14], one can show that the sufficient statistic $\mathbf{t}$ also has a pdf from within the exponential family given as follows:

$$p(\mathbf{t}\,;\boldsymbol{\theta}) = c(\mathbf{t})\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right). \tag{10}$$

The term in (9) that renders direct optimization infeasible is the second term, because of direct dependence on the unknown θ_n. The term $\mathcal{E}\left\{\theta_n^2\right\}$ does not affect the minimization, as we are optimizing over the form of $f_n(\mathbf{t})$. Using SURE, we can eliminate the dependence of true risks on unknown parameters and arrive at unbiased estimators of these risks, which are functions of the noisy data samples alone. For this, we make use of certain identities satisfied by the density model under consideration. In particular, together with some mild assumptions on the estimator form, the identity that makes possible the whole development for the case of densities in the exponential family is the following:

$$\theta_n\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right) = \frac{\partial}{\partial t_n}\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right). \tag{11}$$

Consider the term $\mathcal{E}\{f_n(\mathbf{t})\theta_n\}$. The following series of equalities hold:

$$\begin{aligned}\mathcal{E}\{f_n(\mathbf{t})\theta_n\} &= \int f_n(\mathbf{t})c(\mathbf{t})\theta_n\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right)\mathrm{d}t_n \\ &= \int f_n(\mathbf{t})c(\mathbf{t})\frac{\partial}{\partial t_n}\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right)\mathrm{d}t_n \\ &= -\int \frac{\partial f_n(\mathbf{t})c(\mathbf{t})}{\partial t_n}\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right)\mathrm{d}t_n.\end{aligned} \tag{12}$$

In the second equality we have used (11), and integrated by parts to arrive at the final expression. In the simplification, we assume that $f_n(\mathbf{t})$ is weakly differentiable, and that $\mathcal{E}\{|f_n(\mathbf{t})|\}$ is bounded. The

latter assumption guarantees that the first term in the result of integration by parts vanishes to zero. Along with weak differentiability, the other requirement on the form of $f_n(\mathbf{t})$ is that it must be bounded by a fast-increasing function for its expectation to be bounded. For clearly understanding what we require here, consider the i.i.d. Gaussian model assuming Gaussian noise of mean 0 and variance σ^2. In this case, $f_n(\mathbf{x})$ $\left(\text{note that } \mathbf{t} = \dfrac{\mathbf{x}}{\sigma^2}\right)$ must be bounded by a function like $\exp\left(\sum_{n=1}^{N} x_n^2 \Big/ 2\tilde{\sigma}^2\right)$; $\tilde{\sigma} > \sigma$. That is, in any case, the function should not grow exponentially in such a manner that it overrides the exponential decay of the density model under consideration. Using (12), we get the unbiased estimator of the risk in (9) to be:

$$\epsilon_{\text{SE}} = f_n^2(\mathbf{t}) + 2\frac{\partial}{\partial t_n} f_n(\mathbf{t}) + 2f_n(\mathbf{t})\frac{\partial}{\partial t_n}\log c(\mathbf{t}) + \theta_n^2, \quad (13)$$

which is referred to as the GSURE for MSE [7]. Taking the i.i.d. Gaussian model considered by Stein [1], and writing down the pdf of the observation vector $\mathbf{x}$ as in (1), it is straightforwardly seen that $\mathbf{t} = \dfrac{\mathbf{x}}{\sigma^2}$, and so we choose f_n to be a function of $\mathbf{x}$. Here, the components of $\mathbf{t}$ are independent Gaussian random variables with mean $\dfrac{\theta_n}{\sigma^2}$ and variance $\dfrac{1}{\sigma^2}$. In order to fully define the GSURE objective in (13), the additional information required is the form of the function $c(\mathbf{t})$. This is seen to be:

$$c(\mathbf{t}) = A\exp\left(-\frac{\sigma^2}{2}\sum_{n=1}^{N} t_n^2\right), \quad (14)$$

where A is a constant factor. From (14), we derive the GSURE objective for the i.i.d. Gaussian model to be:

$$\epsilon = f_n^2(\mathbf{x}) + 2\sigma^2\frac{\partial}{\partial x_n} f_n(\mathbf{x}) - 2x_n f_n(\mathbf{x}) + \theta_n^2, \quad (15)$$

which is the classical SURE model used extensively. If we define the form of the denoising function as $f_n(\mathbf{x}) = a_n x_n$, which is a pointwise linear estimator (the MJS estimator), the optimal a_ns obtained by minimizing (15) turn out to be:

$$a_n = 1 - \frac{\sigma^2}{x_n^2}. \quad (16)$$

We note that the MJS estimator was used in the DCT domain for speech enhancement in [8]. We are now in a position to derive an unbiased estimator for the risk corresponding to the IS measure in (8).

4. GSURE FOR d_{IS}

The distortion function in (8) is equivalently written as follows:

$$\begin{aligned} d_{\text{IS}} &= \frac{f_n(\mathbf{t})}{x_n}\left(1-\frac{w_n}{x_n}\right)^{-1} - \log f_n(\mathbf{t}) + \log\theta_n - 1 \\ &= \frac{f_n(\mathbf{t})}{x_n}\sum_{k=0}^{\infty}\left(\frac{w_n}{x_n}\right)^k - \log f_n(\mathbf{t}) + \log\theta_n - 1 \\ &= \frac{f_n(\mathbf{t})}{x_n}\sum_{k=0}^{\infty}\left(1-\frac{\theta_n}{x_n}\right)^k - \log f_n(\mathbf{t}) + \log\theta_n - 1, \quad (17) \end{aligned}$$

where in the second equality, we have used the binomial expansion of $\left(1-\dfrac{w_n}{x_n}\right)^{-1}$ assuming $\left|\dfrac{w_n}{x_n}\right| < 1$. With a high SNR assumption, a close enough approximation to the series is got by truncating it to the first five terms in the final expression. The term in the corresponding risk, which needs simplification is: $\mathcal{E}\left\{\dfrac{f_n(\mathbf{t})}{g_n(\mathbf{t})}\sum_{k=0}^{4}\left(1-\dfrac{\theta_n}{g_n(\mathbf{t})}\right)^k\right\}$, where we have assumed that $\psi(\mathbf{x})$ is an invertible function, and denoted its inverse function by $g(\mathbf{t})$. Note that in case of most densities within the exponential family, we have $\psi(\mathbf{x}) \propto \mathbf{x}$ at least in the independent case. Specifically, in the correlated Gaussian noise case as in (2), $\mathbf{x} = g(\mathbf{t}) = \mathbf{Ct}$, so that each x_n is a linear combination of all t_ns. The simplified form of this term is given in (18), where $\xi_\ell(\mathbf{t}) = \dfrac{f_n(\mathbf{t})}{g_n^\ell(\mathbf{t})}$. A recursive form of Stein's lemma was used in [8] for proceeding with the derivation. Presently, we note that a similar situation accrues here, with (11) being used the required number of times to simplify the last four terms in (18). For instance, consider the third term in (18) (see next page). The simplification proceeds as given below:

$$\begin{aligned} \mathcal{E}\{\xi_3(\mathbf{t})\theta_n^2\} &= \int \theta_n\xi_3(\mathbf{t})c(\mathbf{t})\theta_n\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right)\mathrm{d}t_n \\ &= \int \theta_n\xi_3(\mathbf{t})c(\mathbf{t})\frac{\partial\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right)}{\partial t_n}\mathrm{d}t_n \\ &= -\int \frac{\partial\xi_3(\mathbf{t})c(\mathbf{t})}{\partial t_n}\theta_n\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right)\mathrm{d}t_n \\ &= -\int \frac{\partial\xi_3(\mathbf{t})c(\mathbf{t})}{\partial t_n}\frac{\partial\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right)}{\partial t_n}\mathrm{d}t_n \\ &= \int \frac{\partial^2\xi_3(\mathbf{t})c(\mathbf{t})}{\partial t_n^2}\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right)\mathrm{d}t_n. \quad (19) \end{aligned}$$

In arriving at (19), we have assumed that the integrated part $\left|\left[c(\mathbf{t})\dfrac{\partial\xi_3(\mathbf{t})}{\partial t_n} + \xi_3(\mathbf{t})\dfrac{\partial c(\mathbf{t})}{\partial t_n}\right]\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right)\right|$ goes to zero as $|t_n|$ goes to zero. In the cases of many densities like correlated Gaussian, exponential, Rayleigh, etc., along with the independence assumption, when the form of the estimator is again as was assumed for the MSE derivation, one can easily verify that this term vanishes to zero. In fact, as we try to proceed in a similar fashion for the last two terms in (18), we would demand that the first terms arising out of the repeated integration by parts, should decay to zero asymptotically, which is also guaranteed in the cases just mentioned above. That is, for simplifying the fourth term, $|P_{1,\boldsymbol{\theta}}(\mathbf{t})|$ in (20) should vanish to zero, whereas we need $|P_{2,\boldsymbol{\theta}}(\mathbf{t})|$ in (21) to go to zero asymptotically, for simplifying the last term in (18). Now, (19) can equivalently be written as in (22). Thus, the unbiased estimator of the corresponding risk in this case is:

$$\begin{aligned} \epsilon_{\text{IS}} &= 5\xi_1(\mathbf{t}) + 10\frac{\partial\xi_2(\mathbf{t})}{\partial t_n} + 10\frac{\partial^2\xi_3(\mathbf{t})}{\partial t_n^2} + 5\frac{\partial^3\xi_4(\mathbf{t})}{\partial t_n^3} \\ &+\frac{\partial^4\xi_5(\mathbf{t})}{\partial t_n^4} + A_{1,\text{I}}(\mathbf{t})\frac{\partial\log c(\mathbf{t})}{\partial t_n} + A_{2,\text{I}}(\mathbf{t})\left(\frac{\partial\log c(\mathbf{t})}{\partial t_n}\right)^2 \\ &+A_{3,\text{I}}(\mathbf{t})\left(\frac{\partial\log c(\mathbf{t})}{\partial t_n}\right)^3 + A_{4,\text{I}}(\mathbf{t})\left(\frac{\partial\log c(\mathbf{t})}{\partial t_n}\right)^4 \\ &+B_{1,\text{I}}(\mathbf{t})\frac{\partial^2\log c(\mathbf{t})}{\partial t_n^2} + B_{2,\text{I}}(\mathbf{t})\left(\frac{\partial^2\log c(\mathbf{t})}{\partial t_n^2}\right)^2 \\ &+C_{1,\text{I}}(\mathbf{t})\frac{\partial^3\log c(\mathbf{t})}{\partial t_n^3} + D_{1,\text{I}}(\mathbf{t})\frac{\partial^4\log c(\mathbf{t})}{\partial t_n^4} - \log f_n(\mathbf{t}) + \log\theta_n - 1. \quad (23) \end{aligned}$$

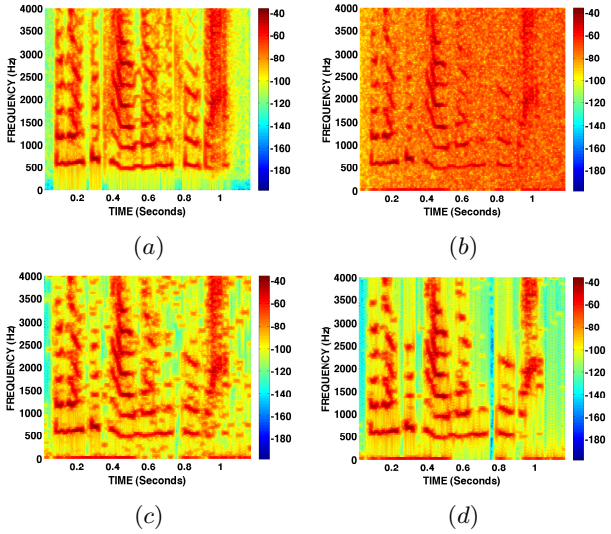


Fig. 2. (Color online) Spectrograms for speech signal 1: (a) Clean speech (b) Noisy speech (SNR=5.50 dB) (c) Enhanced speech using MSE (global SNR=10.86 dB) (d) Enhanced speech using IS (global SNR=9.09 dB).

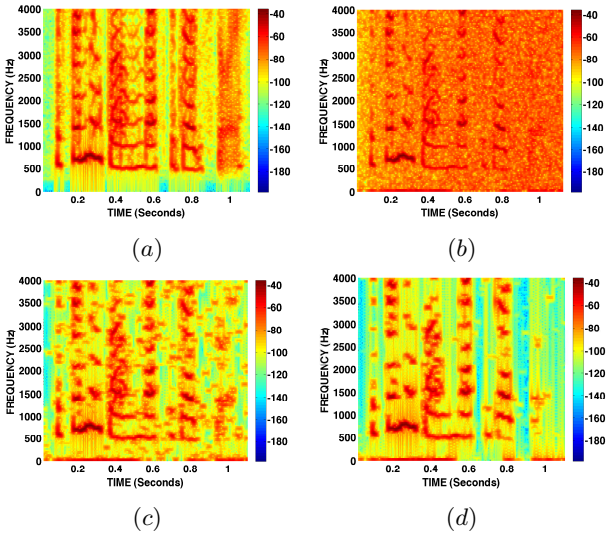


Fig. 3. (Color online) Spectrograms for speech signal 2: (a) Clean speech (b) Noisy speech (SNR=5.40 dB) (c) Enhanced speech using MSE (global SNR=11.04 dB) (d) Enhanced speech using IS (global SNR=9.73 dB).

The functions $A_{i,\mathrm{I}}, B_{i,\mathrm{I}}, C_{i,\mathrm{I}}, D_{i,\mathrm{I}}$ are expanded in (24)-(31).

$$A_{1,\mathrm{I}}(\mathbf{t}) = 10\xi_2(\mathbf{t}) + 20\frac{\partial\xi_3(\mathbf{t})}{\partial t_n} + 15\frac{\partial^2\xi_4(\mathbf{t})}{\partial t_n^2} + \left(15\xi_4(\mathbf{t}) + 12\frac{\partial\xi_5(\mathbf{t})}{\partial t_n}\right)\frac{\partial^2\log c(\mathbf{t})}{\partial t_n^2} + 4\frac{\partial^3\xi_5(\mathbf{t})}{\partial t_n^3} + 4\xi_5(\mathbf{t})\frac{\partial^3\log c(\mathbf{t})}{\partial t_n^3}. \tag{24}$$

$$A_{2,\mathrm{I}}(\mathbf{t}) = 10\xi_3(\mathbf{t}) + 15\frac{\partial\xi_4(\mathbf{t})}{\partial t_n} + 6\frac{\partial^2\xi_5(\mathbf{t})}{\partial t_n^2} + 6\xi_5(\mathbf{t})\frac{\partial^2\log c(\mathbf{t})}{\partial t_n^2}. \tag{25}$$

$$A_{3,\mathrm{I}}(\mathbf{t}) = 5\xi_4(\mathbf{t}) + 4\frac{\partial\xi_5(\mathbf{t})}{\partial t_n}. \tag{26}$$

$$A_{4,\mathrm{I}}(\mathbf{t}) = \xi_5(\mathbf{t}). \tag{27}$$

$$B_{1,\mathrm{I}}(\mathbf{t}) = 10\xi_3(\mathbf{t}) + 15\frac{\partial\xi_4(\mathbf{t})}{\partial t_n} + 6\frac{\partial^2\xi_5(\mathbf{t})}{\partial t_n^2}. \tag{28}$$

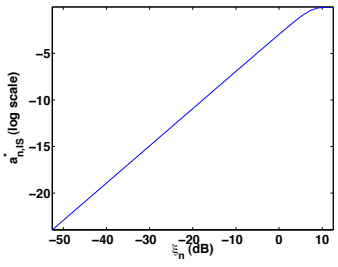


Fig. 4. Variation of $a^*_{n,\mathrm{IS}}$ with a-posteriori SNR.

$$B_{2,\mathrm{I}}(\mathbf{t}) = 3\xi_5(\mathbf{t}). \tag{29}$$

$$C_{1,\mathrm{I}}(\mathbf{t}) = 5\xi_4(\mathbf{t}) + 4\frac{\partial\xi_5(\mathbf{t})}{\partial t_n}. \tag{30}$$

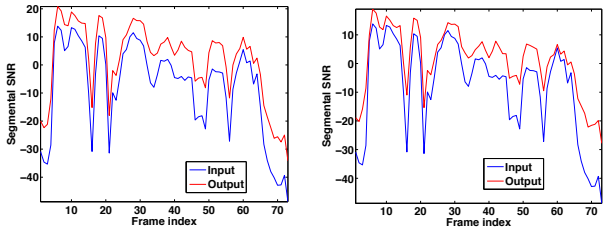
$$D_{1,\mathrm{I}}(\mathbf{t}) = \xi_5(\mathbf{t}). \tag{31}$$

$$\mathcal{E}\left\{\frac{f_n(\mathbf{t})}{\theta_n}\right\} \approx \mathcal{E}\left\{5\frac{f_n(\mathbf{t})}{g_n(\mathbf{t})} - 10\frac{f_n(\mathbf{t})}{g_n^2(\mathbf{t})}\theta_n + 10\frac{f_n(\mathbf{t})}{g_n^3(\mathbf{t})}\theta_n^2 - 5\frac{f_n(\mathbf{t})}{g_n^4(\mathbf{t})}\theta_n^3 + \frac{f_n(\mathbf{t})}{g_n^5(\mathbf{t})}\theta_n^4\right\}$$
$$= \mathcal{E}\left\{5\xi_1(\mathbf{t}) - 10\xi_2(\mathbf{t})\theta_n + 10\xi_3(\mathbf{t})\theta_n^2 - 5\xi_4(\mathbf{t})\theta_n^3 + \xi_5(\mathbf{t})\theta_n^4\right\}. \tag{18}$$

$$P_{1,\boldsymbol{\theta}}(\mathbf{t}) = \left[c(\mathbf{t})\frac{\partial^2\xi_3(\mathbf{t})}{\partial t_n^2} + 2\frac{\partial\xi_3(\mathbf{t})}{\partial t_n}\frac{\partial c(\mathbf{t})}{\partial t_n} + \phi_3(\mathbf{t})\frac{\partial^2 c(\mathbf{t})}{\partial t_n^2}\right]\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right). \tag{20}$$

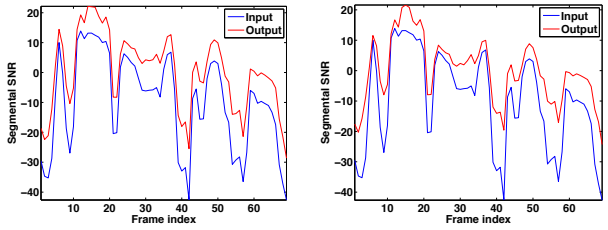
$$P_{2,\boldsymbol{\theta}}(\mathbf{t}) = \left[c(\mathbf{t})\frac{\partial^3\xi_3(\mathbf{t})}{\partial t_n^3} + 3\frac{\partial^2\xi_3(\mathbf{t})}{\partial t_n^2}\frac{\partial c(\mathbf{t})}{\partial t_n} + 3\frac{\partial\xi_3(\mathbf{t})}{\partial t_n}\frac{\partial^2 c(\mathbf{t})}{\partial t_n^2} + \xi_3(\mathbf{t})\frac{\partial^3 c(\mathbf{t})}{\partial t_n^3}\right]\exp\left(\boldsymbol{\theta}^T\mathbf{t} - b(\boldsymbol{\theta})\right). \tag{21}$$

$$\mathcal{E}\{\xi_3(\mathbf{t})\theta_n^2\} = \mathcal{E}\left\{\frac{\partial^2\xi_3(\mathbf{t})}{\partial t_n^2} + 2\frac{\partial\xi_3(\mathbf{t})}{\partial t_n}\frac{\partial\log c(\mathbf{t})}{\partial t_n} + \xi_3(\mathbf{t})\left[\frac{\partial^2\log c(\mathbf{t})}{\partial t_n^2} + \left(\frac{\partial\log c(\mathbf{t})}{\partial t_n}\right)^2\right]\right\}. \tag{22}$$



(a) (b)

Fig. 5. (Color online) Segmental SNRs for speech signal 1: (a) Using MSE (b) Using IS.



(a) (b)

Fig. 6. (Color online) Segmental SNRs for speech signal 2: (a) Using MSE (b) Using IS.

For the special case of the i.i.d. Gaussian model, using $\mathbf{t} = \dfrac{\mathbf{x}}{\sigma^2}$, (14), and $f_n(\mathbf{x}) = a_n x_n$, we get the unbiased estimator of the IS risk to be:

$$\epsilon_{\text{IS}} = a_n \left(1 + 60\frac{\sigma^6}{x_n^6} + 840\frac{\sigma^8}{x_n^8}\right) - \log(a_n x_n) - \log\theta_n - 1, \quad (32)$$

which agrees with the result presented in [8].

5. EXPERIMENTAL RESULTS

For the validation of the theoretical results presented, we selected two utterances from the NOIZEUS database [15]–"The lazy cow lay in the cool grass," and "The friendly gang left the drug store." We use the model in (2), with the noise vector $\mathbf{w}$ assumed to be a Gaussian random vector with zero mean and covariance matrix, $\mathbf{C}$. In our experiments, we selected a matrix $\mathbf{C}$ as follows. For the sake of convenience, we generated an upper triangular matrix with the elements in the j^{th} row being $10^{-7}j$ for $j = 1, 2, 3, \cdots, N$, made it symmetric, and used it as our $\mathbf{C}$. Note that in this case, the sufficient statistic $\mathbf{t}$ is given by (4) as $\mathbf{C}^{-1}\mathbf{x}$, which is Gaussian with mean $\mathbf{C}^{-1}\boldsymbol{\theta}$ and covariance matrix $\mathbf{C}^{-1}$. Therefore, $c(\mathbf{t})$ in (10) is equal to

$$B\exp\left(-\frac{1}{2}\mathbf{t}^T\mathbf{C}\mathbf{t}\right), \quad (33)$$

with B being a constant. Again, the maximum likelihood estimate of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_{\text{ML}} = \mathbf{x} = \mathbf{C}\mathbf{t}$. Stein [1] showed that one can dominate the classical maximum likelihood estimator by applying a suitable shrinkage, depending upon the noise strength. With this motivation, suppose we seek estimators of the form $\mathbf{f}(\mathbf{t}) = \mathbf{a} \,.\, \hat{\boldsymbol{\theta}}_{\text{ML}}$, with "." denoting element-wise multiplication, we can ask: "What are the GSURE-optimal a_ns?." For answering this question, we have to fully specify the objectives in (13) and (23). This is accomplished with the following expressions, which are easily verified:

$$\frac{\partial \log c(\mathbf{t})}{\partial t_n} = -x_n \quad (34)$$

$$\frac{\partial^2 \log c(\mathbf{t})}{\partial t_n^2} = -c_{nn} \quad (35)$$

$$\frac{\partial^3 \log c(\mathbf{t})}{\partial t_n^3} = \frac{\partial^4 \log c(\mathbf{t})}{\partial t_n^4} = 0, \quad (36)$$

with c_{nn} in (35) denoting the n^{th} diagonal element of $\mathbf{C}$. With these expressions, the costs in (13) and (23) get simplified as follows:

$$\epsilon_{\text{SE}} = a_n^2 x_n^2 - 2a_n x_n^2 + 2c_{nn}^2 a_n + \theta_n^2, \quad (37)$$

$$\epsilon_{\text{IS}} = a_n \left(1 + 60\frac{c_{nn}^6}{x_n^6} + 840\frac{c_{nn}^8}{x_n^8}\right) - \log(a_n x_n) - \log\theta_n - 1. \quad (38)$$

Thus, the respective optimum a_ns are given as:

$$a^*_{n,\text{SE}} = 1 - \frac{c_{nn}^2}{x_n^2}, \quad \text{and} \quad (39)$$

$$a^*_{n,\text{IS}} = \left[1 + 60\frac{c_{nn}^6}{x_n^6} + 840\frac{c_{nn}^8}{x_n^8}\right]^{-1}. \quad (40)$$

Defining the a-posteriori SNR ξ_n as $\xi_n = \dfrac{x_n^2}{c_{nn}^2}$, we see that the optimal a_ns can be represented equivalently as follows:

$$a^*_{n,\text{SE}} = 1 - \frac{1}{\xi_n} \quad (41)$$

$$a^*_{n,\text{IS}} = \left[1 + 60\frac{1}{\xi_n^3} + 840\frac{1}{\xi_n^4}\right]^{-1}. \quad (42)$$

Note that in (41), we make the $a^*_{n,\text{SE}}$s zero if they become negative, since the resulting estimator dominates the one with the coefficients being negative. Now that our algorithm is fully specified, we proceed to the presentation of results. We show the results from our risk-estimation algorithm on two signals–henceforth referred to as speech signals 1 and 2, respectively. The spectrograms of the clean, noisy, MSE-enhanced, and IS-enhanced speech, corresponding to the signals 1 and 2 are given in Figure 2 and Figure 3, respectively. All the spectrograms shown here, are constructed using a Hamming window of 256 samples with the overlap being 128 samples. Even though the global SNR is higher for the MSE-enhanced speech, the resulting signal suffers from musical noise, visible as red spots in the spectrogram. This undesirable noise is suppressed in the IS case, as seen from the IS-enhanced spectrograms. The variation of the IS-optimal spectral weighting coefficients as given in (42), with the a-posteriori SNR is presented in Figure (4). This variation is consistent with our intuition that, when the a-posteriori SNR is high, the weighting applied need only be close to one, and vice versa for suppression of noise. Also, we present the segmental SNR plots averaged over multiple noise realizations for the two signals with both MSE and IS enhancements in Figures (5) and (6), respectively. We understand that the MSE-based enhancement presents good performance in speech regions, whereas the residual noise in the time-frequency plane is suppressed to a considerable extent using the perceptually-motivated IS measure.

6. CONCLUSIONS

In this work, we have addressed the speech enhancement problem using a risk-estimation approach. In particular, we considered the well-known Stein's risk estimator, and derived an unbiased estimator for the perceptually motivated IS distortion, which is non-quadratic using an approximate Taylor series analysis. The observation model was assumed to fall within the exponential family, with there being no assumptions of noise following a particular pdf or that of independence. Using an additive noise model, we noted that in the context of speech, our development is of significance since it serves to enhance speech corrupted with correlated Gaussian or double-exponential noise. We used a multiplicative factor a_n, in the spectral domain for enhancement, which can be seen as a filtering process. Even though the estimator was chosen to be pointwise linear, the optimum form turned out to be highly non-linear in the observations. Again, we observed the variation of the optimum gain factor (dependent only upon the a-posteriori SNR) with a-posteriori SNR and found that in the high SNR regions, the factor is close to unity, and vice versa in the low SNR regions. This form for the optimum estimator, derived using a risk-estimation approach, helps us to draw certain commonalities between the classical Wiener filtering and our approach. Experimental results showed that the undesirable musical noise is suppressed considerably using the IS metric, though the global SNR is better for the MSE-enhanced speech. Due to space constraints, we have presented some preliminary results in this paper, which confirmed the validity of our theoretical developments. An extended set of results, considering different perceptual measures and different noise models will be provided in a journal version of this paper.

7. REFERENCES

[1] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 9, no. 6, pp. 1135–1151, Nov. 1981.

[2] F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding," *IEEE Trans. Image Process.*, vol 16, no. 3, pp. 593–606, Mar. 2007.

[3] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.

[4] A. Benazza-Benyahia and J.-C. Pesquet, "Building robust wavelet estimators for multicomponent images using Stein's principle," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1814–1830, Nov. 2005.

[5] J.-C Pesquet, A. Benazza-Benyahia, and C. Chaux, "A SURE approach for digital signal/image deconvolution problems," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4616–4632, Dec. 2009.

[6] R. Giryes, M. Elad, and Y. C. Eldar, "The projected GSURE for automatic parameter tuning in iterative shrinkage methods," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 3, pp. 407–422, May 2011.

[7] Y. C. Eldar, "Generalized SURE for exponential families: Applications to regularization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 471–481, Feb. 2009.

[8] N. R. Muraka and C. S. Seelamantula, "A risk-estimation-based comparison of mean square error and Itakura-Saito distortion measures for speech enhancement," in *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 349–352.

[9] N. Zheng, X. Li, T. Blu, and T. Lee, "SURE-MSE speech enhancement for robust speech recognition," in *Proc. 7th Int. Symp. Chinese Spoken Language Process.*, Shenzhen, China, Nov. 2010, pp. 271–274.

[10] H. M. Hudson, "A natural identity for exponential families with applications in multiparameter estimation," *Ann. Stat.*, vol. 6, no. 3, pp. 473–484, May 1978.

[11] J. Berger, "Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters," *Ann. Stat.*, vol. 8, no. 3, pp. 545–571, May 1980.

[12] J. T. Hwang, "Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases," *Ann. Stat.*, vol. 10, no. 3, pp. 857–867, Sep. 1982.

[13] M. Raphan and E. P. Simoncelli, "Learning least squares estimators without assumed priors or supervision," Tech. Rep. Computer Science TR2009-923, Courant Inst. of Mathematical Sciences, New York University, Aug. 2009.

[14] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, 1st ed., NJ: Prentice-Hall, 1993.

[15] Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, 2007.