A Risk-Estimation-Based Comparison of Mean Square Error and Itakura-Saito Distortion Measures for Speech Enhancement

Nagarjuna Reddy Muraka and Chandra Sekhar Seelamantula

Department of Electrical Engineering
Indian Institute of Science
Bangalore-560 012, India
nagarjuna@ee.iisc.ernet.in, chandra.sekhar@ieee.org

10.21437/Interspeech.2011-127

Abstract

The goal of speech enhancement algorithms is to provide an estimate of clean speech starting from noisy observations. In general, the estimate is obtained by minimizing a chosen distortion metric. The often-employed cost is the mean-square error (MSE), which results in a Wiener-filter solution. Since the ground truth is not available in practice, the practical utility of the optimal estimators is limited. Alternative, one can optimize an unbiased estimate of the MSE. This is the key idea behind Stein's unbiased risk estimation (SURE) principle. Within this framework, we derive SURE solutions for the MSE and Itakura-Saito (IS) distortion measures. We also propose parametric versions of the corresponding SURE estimators, which give additional flexibility in controlling the attenuation characteristics for maximum signal-to-noise-ratio (SNR) gain. We compare the performance of the two distortion measures in terms of attenuation profiles, average segmental SNR, global SNR, and spectrograms. We also include a comparison with the standard power spectral subtraction technique. The results show that the IS distortion consistently gives better performance gain in all respects. The perceived quality of the enhanced speech is also better in case of the IS metric.

Index Terms: Stein's unbiased risk estimator, Mean square error, Itakura-Saito distortion.

1. Introduction

The problem of enhancing speech corrupted by noise arises in many practical scenarios. The objective is to suppress the effect of noise and render speech more pleasing to listen to. In general, an estimate of clean speech is obtained by minimizing a chosen distortion metric that measures the closeness between the clean speech and its estimate. Classical solutions for speech enhancement are the spectral subtraction technique [1] and the Wiener filter [2], which are based on the assumption of stationary signal and noise. Priors about the signal can also be incorporated and such approaches give rise to Bayesian estimators [3]. These approaches, although optimal in the sense of the selected distortion metric, have the limitation that the ground truth is not available in practice. In practice, these solutions are replaced by approximate alternatives computed based on sample statistics. For example, the power spectrum density of speech that is required in a Wiener filter formulation is not known a priori, and has to be estimated. A frequently employed estimate is the difference between the noisy signal power spectrum estimate and the noise power spectrum estimate.

In this paper, we employ the minimization of risk estimates for speech enhancement. Specifically, we seek to minimize unbiased risk estimates of the mean square error (MSE) and Itakura-Saito (IS) distortion metrics [4]. A basic property of the risk estimators is that they do not depend on the unknown speech parameters. The risk estimate is a function of the noisy data alone, which makes it suitable to deploy in practical applications. This is the guiding principle behind the SURE methodology. We develop SURE-optimal solutions for the problem of speech enhancement with the chosen distortion metrics. The SURE formalism helps us replace the cost functions with their unbiased estimates, and results in corresponding denoising functions. We also consider parametric versions of the denoising functions, where the parameters give additional flexibility in further optimizing the performance of the enhancement algorithm.

2. Problem Formulation

Consider samples of clean speech s_n distorted additively by noise w_n resulting in the measurements:

$$x_n = s_n + w_n\,, n = 1, 2, \cdots, N. \tag{1}$$

The noise is assumed to be Gaussian, but we do not constrain it to be white. The signal s_n is assumed to be deterministic. Speech signals have a strong harmonic structure, which manifests strongly in the spectral domain. The properties of speech and noise are also distinct in the spectral domain. The standard approach to computing the spectrum is based on the discrete Fourier transform (DFT), but we use the discrete cosine transform (DCT) in view of its real-valued nature as well as its compactness of representation (decorrelation property).

The equivalent version of (1) in the DCT domain is

$$X_k = S_k + W_k, \quad k = 1, 2, \cdots, N. \tag{2}$$

Since the DCT is an orthonormal transform, and noise is additive Gaussian, the transformed noise also has a Gaussian distribution.

The goal is to estimate S_k given X_k. Let the estimate be denoted by $\hat{S}_k$. The estimation error is $e_k = S_k - \hat{S}_k$. The estimates are computed by minimizing a non-negative distortion metric or risk functional: $d(e_k) = d(S_k, \hat{S}_k)$. For example, $d(e_k) = \mathcal{E}\{e_k^2\}$, where $\mathcal{E}$ denotes the expectation operator. This cost is the mean-square error (MSE). In general, optimization of the risk requires a priori knowledge about the ground truth S_k or its statistics. In the SURE formalism, we first construct an unbiased estimator of the risk and minimize that. This approach results in an estimator of $\hat{S}_k = f(X_k)$, which depends only on the noisy data X_k, and f is the denoising function. We use linear denoising functions f, although,

Copyright © 2011 ISCA

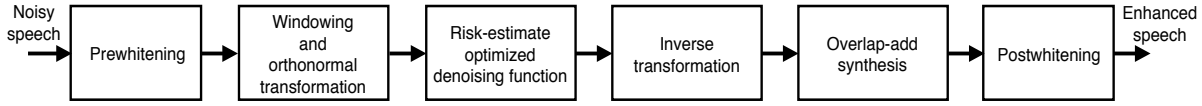


Figure 1: *Block diagram representation for speech enhancement using risk estimation.*

in principle, one may also use nonlinear functions. This procedure is carried out on a frame-by-frame basis to obtain enhanced speech. A block diagram representation of the method is shown in Figure 1. The first block is a pre-whitener, the role of which is to whiten colored noise (in practice, an estimate of the colored noise spectrum is obtained from the initial silence regions). The last block is a post-whitening filter, which is the inverse of the pre-whitener. This formalism helps us to handle colored Gaussian noises, but it suffices to formulate the developments of the intermediate blocks only for white noise.

3. SURE

We recall the following theorem from [5] to obtain the risk estimates.

Theorem 1. *Let Y be a $\mathcal{N}(0,\sigma^2)$ real random variable and let $f:\mathbb{R}\to\mathbb{R}$ be an indefinite integral of the Lebesgue measurable function f', essentially the derivative of f. Suppose also that $\mathcal{E}\{|f'(Y)|\}<\infty$. Then*

$$\mathcal{E}\{Yf(Y)\}=\sigma^2\mathcal{E}\{f'(Y)\}.$$

The proof of the theorem is available in [5]. We state a corollary of the above result:

Corollary 1.

$$\mathcal{E}\{Y^n f(Y)\}=\sigma^2\mathcal{E}\{f'(Y)Y^{n-1}\}+\sigma^2(n-1)\mathcal{E}\{f(Y)Y^{n-2}\}.$$

Proof.

$$\mathcal{E}\{Y^n f(Y)\}=\mathcal{E}\{Yg(Y)\},$$

where $g(Y)=Y^{n-1}f(Y)$. Applying Theorem 1, we get that

$$\begin{aligned}\mathcal{E}\{Yg(Y)\}&=\sigma^2\mathcal{E}\{g'(Y)\}\\&=\sigma^2\mathcal{E}\left\{f'(Y)Y^{n-1}+(n-1)f(Y)Y^{n-2}\right\}.\end{aligned}$$

□

The theorem and corollary stated above will be used in subsequent analysis.

3.1. SURE for MSE distortion

The risk function for the square error is

$$d(S_k,\hat{S}_k)=(\hat{S}_k-S_k)^2,\quad \hat{S}_k=f(X_k). \tag{3}$$

We seek to minimize the MSE:

$$\mathcal{E}\{d(S_k,\hat{S}_k)\}=\mathcal{E}\{f(X_k)^2+S_k^2-2f(X_k)X_k+2f(X_k)W_k\}. \tag{4}$$

Applying Stein's lemma, we get

$$\mathcal{E}\{f(X_k)W_k\}=\sigma^2\mathcal{E}\{f'(X_k)\}. \tag{5}$$

Substituting (5) in (4), we get that

$$\mathcal{E}\{d(S_k,\hat{S}_k)\}=\mathcal{E}\{f(X_k)^2-2f(X_k)X_k+2\sigma^2f'(X_k)\}+S_k^2. \tag{6}$$

Hence, we conclude that

$$\epsilon=\underbrace{f(X_k)^2-2f(X_k)X_k+2\sigma^2f'(X_k)}_{\tilde{\epsilon}}+S_k^2, \tag{7}$$

is an unbiased risk estimator for the average square error distortion. This is the SURE solution for the distortion metric in (4). Minimization of the risk estimate results in the minimization of the square error. Although the risk estimate ϵ contains the signal term S_k^2, it does not affect the process of minimization since the minimization is carried out with respect to f. Hence it is sufficient to minimize $\tilde{\epsilon}$.

3.2. SURE for Itakura-Saito distortion

The Itakura-Saito (IS) metric is quite popular in the speech coding community. It is not a symmetric metric, but its popularity is due to its perceptual relevance in matching two spectra. Here, we consider the IS distortion between the DCT coefficients of noise free speech and its estimate. The IS distortion is defined as

$$d_{IS}(S_k,\hat{S}_k)=\frac{\hat{S}_k}{S_k}-\log\left(\frac{\hat{S}_k}{S_k}\right)-1. \tag{8}$$

Since we consider only linear denoising functions in this paper and the associated gain functions are constrained to be positive, effectively, $\hat{S}_k$ and S_k have the same sign. Therefore, the logarithm in the above equation does not require a modulus on its argument. The average risk is given by

$$\mathcal{E}\{d_{IS}(S_k,\hat{S}_k)\}=\mathcal{E}\left\{\frac{\hat{S}_k}{S_k}\right\}-\mathcal{E}\{\log(\hat{S}_k)\}+\log(S_k)-1. \tag{9}$$

The first term of the preceding equation after substituting $S_k=X_k-W_k$ is

$$\mathcal{E}\left\{\frac{\hat{S}_k}{S_k}\right\}=\mathcal{E}\left\{\frac{\hat{S}_k}{X_k}\left(1-\frac{W_k}{X_k}\right)^{-1}\right\}. \tag{10}$$

Assuming that $\left|\frac{W_k}{X_k}\right|<1$ and replacing $\hat{S}_k=f(X_k)$, we get

$$\begin{aligned}\mathcal{E}\left\{\frac{\hat{S}_k}{S_k}\right\}&=\mathcal{E}\left\{\frac{f(X_k)}{X_k}\sum_{n=0}^{\infty}\left(\frac{W_k}{X_k}\right)^n\right\}\\&=\sum_{n=0}^{\infty}\mathcal{E}\{J_n(X_k)W_k^n\},\end{aligned}$$

where $J_n(X_k)=f(X_k)/X_k^{n+1}$. We rewrite the average risk as

$$\begin{aligned}\mathcal{E}\{d_{IS}(S_k,\hat{S}_k)\}&=\sum_{n=0}^{\infty}\mathcal{E}\{J_n(X_k)W_k^n\}-\mathcal{E}\{\log(f(X_k))\}\\&\quad-\log(S_k)-1.\end{aligned} \tag{11}$$

Further simplification of the above expression in the case of a general denoising function f is not feasible. However, for linear denoising functions, a truncated version of the infinite series can be used to derive the SURE-optimal solution using Corollary 1. This aspect is addressed in the following section.

4. SURE-optimal solutions

Consider a linear denoising function of the form $f(X_k) = a_k X_k$. We derive the optimal value of the parameter a_k that minimize the SURE for MSE and IS distortion measures.

4.1. SURE-optimal solution for MSE

Substituting $f(X_k) = a_k X_k$ in (7), we get that

$$\tilde{\epsilon} = a_k^2 X_k^2 - 2a_k X_k^2 + 2\sigma^2 a_k. \tag{12}$$

The optimal value of a_k is obtained as follows:

$$\frac{\partial \tilde{\epsilon}}{\partial a_k} = 0 \quad \Rightarrow \quad a_k = \left[1 - \frac{\sigma^2}{X_k^2}\right]. \tag{13}$$

To avoid negative a_k, the estimator is revised as follows: $\hat{S}_k = \left[1 - \frac{\sigma^2}{X_k^2}\right]_+ X_k$, where $[x]_+ = \max(0, x)$. Let $\xi_k = X_k^2/\sigma^2$ denote the a posteriori or measured SNR based on the observed data. The above estimate can be expressed in terms of the a posteriori SNR as

$$\hat{S}_k = \left[1 - \frac{1}{\xi_k}\right]_+ X_k = G_{MS}(\xi_k) X_k, \tag{14}$$

where $G_{MS}(\xi_k)$ denotes the corresponding gain function.

4.2. SURE-optimal solution for Itakura-Saito distortion

Substituting $f(X_k) = a_k X_k$ in (11), we get that

$$\begin{aligned}\mathcal{E}\{d_{IS}(S_k, \hat{S}_k)\} \quad &= \quad \left(\sum_{n=0}^{\infty} \mathcal{E}\{a_k W_k^n / X_k^n\}\right) \\ &\quad -\mathcal{E}\{\log(a_k X_k)\} - \log(S_k) - 1.\end{aligned}$$

The terms in the infinite series of the above equation decay with respect to n. Therefore, a truncated series can be used for further calculations. Truncating the infinite series beyond $n = 4$ and applying Corollary 1, we get the risk estimate as

$$\epsilon = \underbrace{a_k\left(1 + 60\frac{\sigma^6}{X_k^6} + 840\frac{\sigma^8}{X_k^8}\right) - \log(a_k X_k)}_{\tilde{\epsilon}} - \log\, S_k - 1.$$

The proof is lengthy and the details are omitted here. Differentiating $\tilde{\epsilon}$ with respect to a_k and setting the derivative to zero, we have

$$\frac{\partial \tilde{\epsilon}}{\partial a_k} = 0 \quad \Rightarrow \quad a_k = \left[1 + 60\frac{\sigma^6}{X_k^6} + 840\frac{\sigma^8}{X_k^8}\right]^{-1}. \tag{15}$$

The resulting estimate is $\hat{S}_k = \left[1 + 60\frac{\sigma^6}{X_k^6} + 840\frac{\sigma^8}{X_k^8}\right]^{-1} X_k$.
In terms of a posteriori SNR, the above estimate is

$$\hat{S}_k \quad = \quad \left[1 + \frac{60}{\xi_k^3} + \frac{840}{\xi_k^4}\right]^{-1} X_k = G_{IS}(\xi_k) X_k, \tag{16}$$

where G_{IS} denotes the corresponding gain function.

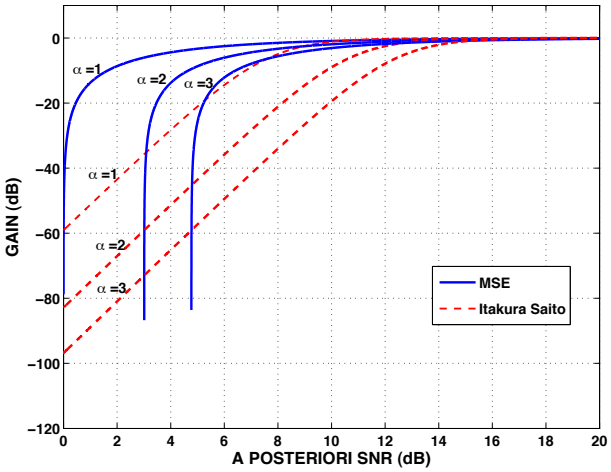


Figure 2: *(Color online) Gain functions associated with the SURE-optimal solutions for MSE and IS distortions for different values of α.*

4.3. Parametric refinement

In standard spectral subtraction algorithms, there is an oversubtraction factor that takes into account the errors in the estimation of the noise power spectral density. The factor also allows us to tradeoff residual noise levels with speech distortion. Taking this as motivation, we also propose parametric versions of the estimators in (14) and (16), where we attach a parameter α to control estimation errors in the a posteriori SNR ξ_k. These are given as

$$\text{SURE-MSE:} \quad \hat{S}_k = \left[1 - \frac{\alpha}{\xi_k}\right]_+ X_k. \tag{17}$$

$$\text{SURE-IS:} \quad \hat{S}_k = \left[1 + \frac{60\alpha^3}{\xi_k^3} + \frac{840\alpha^4}{\xi_k^4}\right]^{-1} X_k. \tag{18}$$

For $\alpha > 1$ the noisy coefficients are attenuated severely. This leads to significant reduction of noise but it may introduce speech distortion. The attenuation curves for both the estimators for different values of α are shown in Figure 2. The attenuation curves shift from left to right with increase in α for both the estimators as shown in Figure 2. We see that for a given α, the attenuation is more for SURE-IS estimator compared to the SURE-MSE estimator.

5. Experimental results

We next present experimental results of speech enhancement using risk estimation techniques. The speech files used for the experiments are taken from the NOIZEUS and CMU–ARCTIC database. The sampling frequency is 8 kHz. The denoising operation was performed using a short-time approach, where the frame size is 256 samples (32 milliseconds), and the overlap between successive frames is 50%. Noisy speech was generated from noise-free speech by adding white Gaussian noise and pink noise at a specified global SNR. The noise variance is estimated from the initial few frames corresponding to no speech activity. The denoising is performed in DCT domain using the estimators given in previous section. The enhancement performance is measured as a function of the input global SNR. The

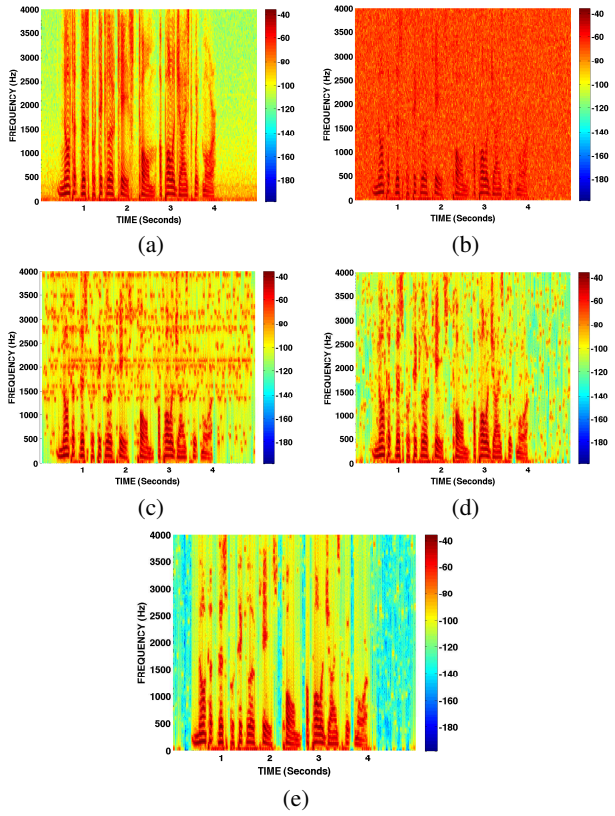


Figure 3: *(Color online) Spectrograms of* (a) *noise-free speech,* (b) *noisy speech(SNR=5dB),* (c) *power spectral subtraction output,* (d) *SURE-optimal solution for MSE*, and (e) *SURE-optimal solution for IS distortion*.

performance metrics used are: (i) the global SNR improvement, which is the difference between the output SNR and the input SNR, and (ii) average segmental SNR gain, that is, the difference between average input segmental SNR and average output segmental SNR, which is a local metric. The noisy speech is enhanced using the SURE criteria for both MSE and IS distortions. The SNR gains (global as well as local) are calculated as a function of the input SNR for different values of α, and averaged over all thirty speech files in the NOIZEUS corpus. The results of global SNR gain are shown in Figure 4. For both estimators, the gains increase as α increases and beyond a certain value of α (2 in case of SURE-IS and 5 in case of SURE-MSE), the gains decrease. The SNR gains are significantly higher for input SNR less than 0 dB than for large values of input SNR. The gain curves in case of average segmental SNR gain are shown in Figure 5. It is evident that for a chosen α, the improvement in average segmental SNR gain for the SURE-optimal IS estimator is significantly better than that of the MSE estimator.

In Figure 3, we show the corresponding spectrograms. For the purpose of comparison, we have also included the results of the standard power spectrum subtraction technique. The power spectrum subtraction technique contains lot of random spurious components (source of musical noise) compared with the SURE-optimal MSE and IS estimators. The SURE-optimal IS risk gives the best results in terms of reduction in noise, musical noise artifacts, and faithful reproduction of the signal spectral components.

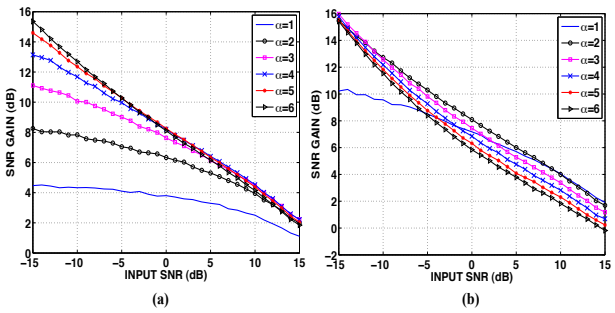


Figure 4: *(Color online) Global SNR gain versus input SNR as a function of* α *for* (a) *SURE-optimal MSE estimator*, and (b) *SURE-optimal IS estimator*.

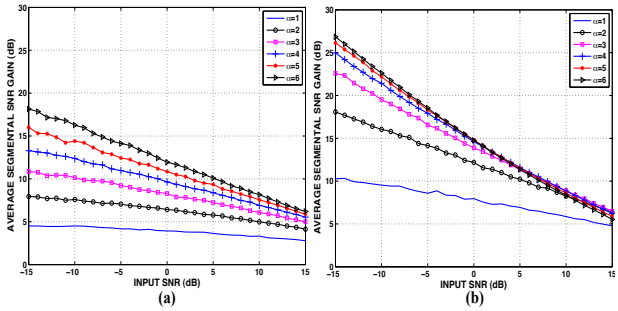


Figure 5: *(Color online) Average segmental SNR gain versus input SNR as a function of* α *for* (a) *SURE-optimal MSE estimator*, and (b) *SURE-optimal IS estimator*.

6. Conclusions

We formulated the problem of speech enhancement in the DCT domain within the framework of Stein's unbiased risk estimation. We employed two different distortion metrics; one is based on the mean-square error and the other is based on the Itakura-Saito distortion metric. We derived the unbiased risk estimators in each case and also optimized them assuming a linear denoising function. The resulting attenuation functions depend only on the noisy observations. We compared the two metrics in terms of their attenuation profiles, the global SNR gain, and the average segmental SNR gain. We found that the SURE-optimal IS estimator outperforms the SURE-optimal MSE solution. In terms of the time-frequency structure also, the SURE-optimal IS is better than the MSE solution and the power spectrum subtraction technique.

7. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Sig. Proc.,* vol. ASSP-27, pp. 113-120, Apr. 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586-1604, Dec. 1979.

[3] P. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech and Audio Proc.,* vol. 13, no. 5, pp. 857-869, Sep. 2005.

[4] L. R. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition, *Prentice Hall*, 1993.

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