SURE-MSE speech enhancement for robust speech recognition

Nengheng ZHENG and Xia LI
College of Information Engineering
Shenzhen University, Shenzhen
Email: {nhzheng;lixia}@szu.edu.cn

Thierry BLU and Tan LEE
Department of Electronic Engineering
The Chinese University of Hong Kong, Hong Kong
Email:{tblu;tanlee}@ee.cuhk.edu.hk

***Abstract*—This paper presents a new approach to enhancing noisy (white Gaussian noise) speech signals for robust speech recognition. It is based on the minimization of an estimate of denoising MSE (known as SURE) and does not require any hypotheses on the original signal. The enhanced signal is obtained by thresholding coefficients in the DCT domain, with the parameters in the thresholding functions being specified through the minimization of the SURE. Thanks to a linear parametrization, this optimization is very cost-effective. This method also works well for non-white noise with a noise whitening processing before the optimization. We have performed automatic speech recognition tests on a subset of the AURORA 2 database, to compare our method with different denoising strategies. The results show that our method brings a substantial increase in recognition accuracy.**

***Index Terms*—Speech enhancement, Stein's unbiased risk estimate, MMSE, automatic speech recognition**

I. Introduction

Speech enhancement aims at improving the performance of speech communication between humans (e.g., in telecommunication systems, for hearing-aids users, etc.), or between human and machines (e.g., automatic speech recognition). In these communication systems, the interference sources could be either additive or convolutive, and sometimes both. In this paper, we focus on additive wide-band Gaussian noise (white or colored noise). Although there are many approaches developed for estimating clean speech from noisy input, e.g., [1][2][3][4][5][6], the most effect approach is the Wiener filtering-based speech estimation, which aims at minimizing the mean square error (MSE) between the estimated signal and the original one. This category of approaches include spectral subtraction [1] and its derivatives such as the signal subspace approach [4] and the estimation of short-time spectral magnitude [2] or log-magnitude [3]. The MMSE estimator, however, requires prior knowledge of the second-order statistics (the variance or covariance) of the noise and the clean signal. In practice, although the noise statistics can be obtained from the non-speech portion by voice activity detection (VAD), the clean speech signal is generally not available and should be estimated from noisy signals. Thus the MMSE criteria is not guaranteed. More importantly, the error in clean speech statistics estimation leads to a certain degree of speech distortion, which would reduce the accuracy of automatic speech recognition. As a matter of fact, most of the existing speech enhancement approaches improve the signal-to-noise ratio (SNR) at the expense of decreasing speech intelligibility. But a significant improvement in SNR does not necessarily lead to better recognition performance [7].

In this paper, we present a new MMSE speech enhancement approach. Instead of minimizing the true denoising MSE, our approach minimizes a statistically unbiased MSE estimate – Stein's unbiased risk estimate (SURE), which depends on the noise and the noisy signal, but not on the clean signal. Blu *et al.* [8][9] presented an effective image denoising approach based on the SURE MSE estimate in the DWT domain. In this paper, the enhanced speech signal is obtained by thresholding the coefficients in the DCT domain. The thresholding parameters are specified through the minimization of the SURE-MSE. Thanks to the linear parametrization, the optimization process can be made very cost-effective. We demonstrate the effectiveness of this method in enhancing speech corrupted by additive noise, and its performance as an front-end processor for robust speech recognition. Speech recognition experiments on the AURORA 2 database [10] show that our method brings a substantial increase in word accuracy.

II. SURE-based MMSE approach for signal denoising

Consider a clean signal $\mathbf{x}$ contaminated by additive white Gaussian noise $\mathbf{b} \sim \mathcal{N}(0, \sigma^2)$. The observed noisy signal is $\mathbf{y} = \mathbf{x} + \mathbf{b}$. The basic idea of MMSE denoising is to find an estimate of $\mathbf{x}$, $\hat{\mathbf{x}}$, which minimizes the mean square error

$$\mathbf{MSE} = \langle |\hat{\mathbf{x}} - \mathbf{x}|^2 \rangle = \frac{1}{N} \sum_{n=1}^{N} |\hat{x}_n - x_n|^2, \qquad (1)$$

where N is the number of samples. The task can be solved as finding a transformed function of $\mathbf{y}$, i.e., $\hat{\mathbf{x}} = \theta(\mathbf{y})$ that minimizes

$$\mathbf{MSE} = \langle |\theta(\mathbf{y}) - \mathbf{x}|^2 \rangle = \langle \theta(\mathbf{y})^2 \rangle - 2\langle \mathbf{x}\theta(\mathbf{y}) \rangle + \langle \mathbf{x}^2 \rangle. \quad (2)$$

A. Stein's unbiased risk estimate (SURE)

Since we do not have access to $\mathbf{x}$, the MSE form (2) can not be computed directly. Nevertheless, following Stein's theorem, an unbiased estimate of the MSE, which does not depend on the clean signal, can be found.

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

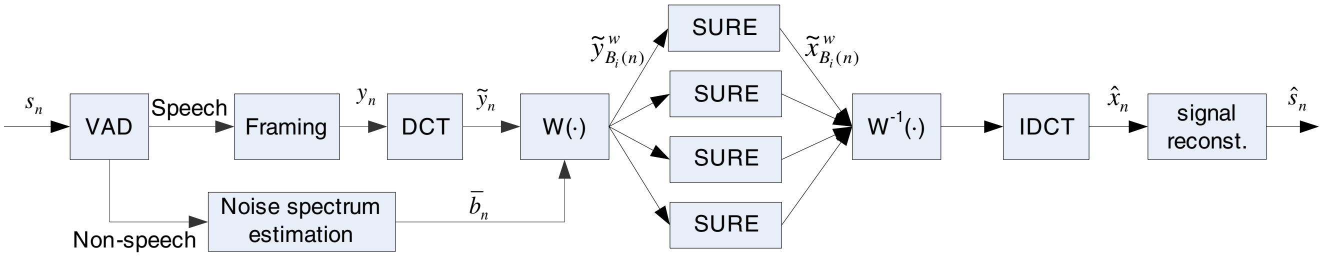


Fig. 1. Block diagram for the SURE approach for speech enhancement.

Theorem: Let $\theta : \mathbf{R} \rightarrow \mathbf{R}$ be a smooth function. The random variable

$$\epsilon = \langle \theta^2(\mathbf{y}) - 2\mathbf{y}\theta(\mathbf{y}) + 2\sigma^2\theta'(\mathbf{y})\rangle + \langle \mathbf{y}^2 \rangle - \sigma^2 \tag{3}$$

is an unbiased estimate of the MSE, i.e.,

$$\mathcal{E}\{\epsilon\} = \mathcal{E}\{\langle |\theta(\mathbf{y}) - \mathbf{x}|^2\rangle\}. \tag{4}$$

Refer to Blu [8] and Stein [11] for the proof of this theorem.

B. Denoising in a transformed domain

Suppose $\tilde{\mathbf{y}}$ is a transformed version of $\mathbf{y}$, i.e.,

$$\tilde{\mathbf{y}} = \mathbf{A}\mathbf{y}, \tag{5}$$

where $\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_N]^T$, $\mathbf{y} = [y_1, y_2, \cdots, y_N]^T$. $\mathbf{A}$ is supposed to be orthonormal, i.e., $\mathbf{A}^T = \mathbf{A^{-1}}$, in order to preserve the MSE in the transformed domain. Similarly, we have the transformed version of the clean speech $\tilde{\mathbf{x}} = \mathbf{A}\mathbf{x}$, $\tilde{\mathbf{x}} = [\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_N]^T$, $\mathbf{x} = [x_1, x_2, \cdots, x_N]^T$. The MSE in the transformed domain can be written as

$$\mathbf{MSE} = \langle |\theta(\tilde{\mathbf{y}}) - \tilde{\mathbf{x}}| \rangle^2. \tag{6}$$

Note that θ is a pointwise denoising function. That is, $\theta(\tilde{\mathbf{y}}) = [\theta(\tilde{y}_1), \theta(\tilde{y}_2), \cdots, \theta(\tilde{y}_N)]^T$. In that case, the MSE estimate (SURE) is

$$\epsilon = \sum_n (\theta^2(\tilde{y}_n) - 2\tilde{y}_n\theta(\tilde{y}_n) + 2\sigma^2\theta'(\tilde{y}_n) + \tilde{y}_n^2) - \sigma^2. \tag{7}$$

C. Minimizing the SURE-MSE

Now the task is to find an appropriate denoising function θ to minimize ϵ. To do so, we use a linearly parameterized pointwise thresholding function as follows

$$\theta(\tilde{y}_n) = \sum_{k=1}^{K} a_k \varphi_k(\tilde{y}_n), \tag{8}$$

where K is the number of parameters, and

$$\varphi_k(\tilde{y}_n) = \tilde{y}_n e^{-(k-1)\frac{\tilde{y}_n^2}{12\sigma^2}}. \tag{9}$$

The derivative of Gaussian is selected in (9) for its fast decay, which ensures a linear behavior of large signal coefficients. The thresholding parameters a_k can be computed by minimizing the MSE estimate ϵ. That is, for each $k \in [1, K]$, performing differentiation of ϵ over a_k, we obtain

$$\frac{1}{2}\frac{\partial \epsilon}{\partial a_k} = \sum_n \{\theta(\tilde{y}_n)\varphi_k(\tilde{y}_n) - \tilde{y}_n\varphi_k(\tilde{y}_n) + \sigma^2\varphi_k'(\tilde{y}_n)\} = 0. \tag{10}$$

The following equations are obtained for each k,

$$\sum_{l=1}^{K}\sum_n \varphi_k(\tilde{y}_n)\varphi_l(\tilde{y}_n)a_l = \sum_n \{\tilde{y}_n\varphi_k(\tilde{y}_n) - \sigma^2\varphi_k'(\tilde{y}_n)\}. \tag{11}$$

These equations can be summarized in a matrix form as $\mathbf{M}\mathbf{a} = \mathbf{c}$, where $\mathbf{a} = [a_1 \cdots a_k]^T$ and $\mathbf{c} = [c_1 \cdots c_k]^T$ are vectors of size $K \times 1$, and $\mathbf{M} = [M_{k,l}]_{1 \le k,l \le K}$ is a matrix of size $K \times K$. The linear system is solved for $\mathbf{a}$ by

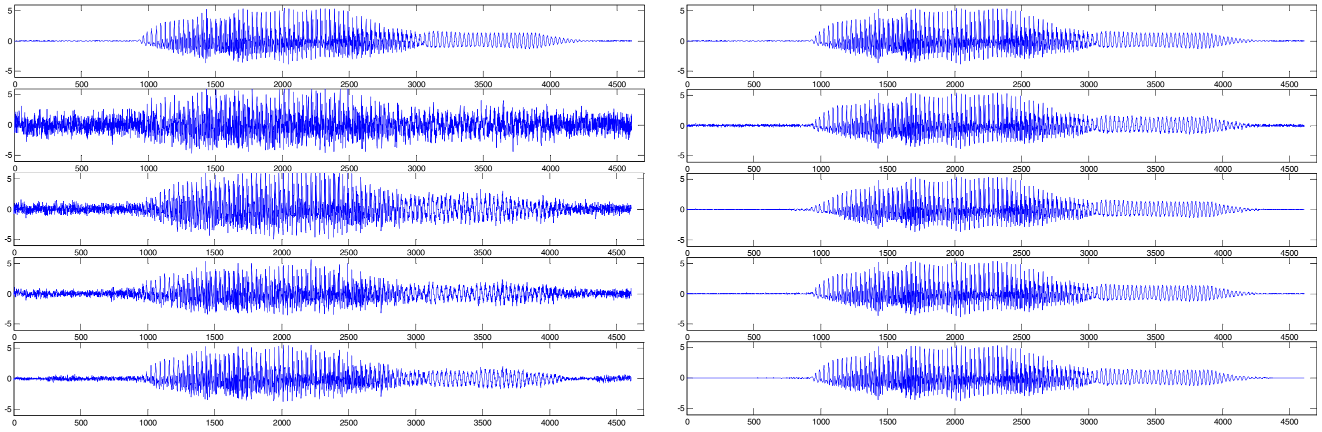
$$\mathbf{a} = \mathbf{M}^{-1}\mathbf{c}, \tag{12}$$

which makes the approach very simple to implement.

III. SURE APPROACH TO SPEECH ENHANCEMENT

In this study, the enhanced speech signal is obtained by thresholding the DCT coefficients of the noisy signal. In practical speech communication environments, the contaminating noise is usually not white noise. To deal with the non-white noise, a pre-whitening processing is implemented before denoising. The procedures are illustrated as in Fig. 1 and summarized below:

- Voice activity detection: an energy-based VAD is applied to separate speech from non-speech.
- Frame blocking: the signal is divided into frames of 512 samples.
- Noise spectrum estimation: a smooth noise spectrum $\bar{b}_n, n \in [1, 512]$ is estimated from detected non-speech segments. The noise spectrum $\tilde{b}_n$ is first computed by the averaging periodograms method. Then, $\tilde{b}_n$ is further smoothed by a seven-point mean processing. That is, $\bar{b}_n = \mathbf{mean}(\tilde{b}_{n-3} : \tilde{b}_{n+3})$.
- Noise whitening: on each frame of noisy speech, DCT spectrum is computed and divided by the noise spectrum, i.e., $\tilde{y}_n^w = \mathbf{W}(\tilde{y}_n) = \tilde{y}_n/\bar{b}_n$.
- Subband SURE denoising: the SURE denoising process is applied to $\tilde{y}_n^w$ in 4 subbands, i.e., 0∼0.5 kHz, 0.5∼1 kHz, 1∼2 kHz and 2∼4 kHz, respectively. The enhanced coefficients in each subband are obtained by $\tilde{x}^w_{B_i(n)} =$



(a) SNR=0 dB (b) SNR=20 dB

Fig. 2. Comparison of speech waveforms obtained from different denoising methods. (a) SNR = 0 dB, (b) SNR = 20 dB. From top to bottom are clean speech, noisy speech, and enhanced speech by E-M, subspace and SURE approach, respectively.

$\theta(\tilde{y}^w_{B_i(n)})$, where $i = 1, 2, 3, 4$, θ takes the form as in (8) and K is selected to be 3. Note that subband processing would not change the MSE estimate since the MSE in the DCT domain is the weighted sum of MSEs from individual subbands.

- Signal reconstruction: the estimated coefficients are multiplied by noise spectrum $\bar{b}_n$ and then the speech signal is obtained by inverse DCT, i.e., $\hat{x}_n = \mathbf{IDCT}(\tilde{x}^w_n \cdot \bar{b}_n)$. Finally the denoised speech signal $\hat{s}_n$ is reconstructed from each frame of $\hat{x}_n$.

IV. Speech Enhancement Results

We evaluate the performances of three MMSE-based algorithms: 1) Ephraim-Malah' short-time spectral amplitude estimator (E-M); 2) Ephraim-Van Tree's signal subspace approach (SubSp); and 3) our proposed SURE-MSE approach. We compare them in the cases of additive white Gaussian and non-white noises.

The evaluation is conducted on the AURORA 2 Set A test data. The database has been widely used for developing and evaluating noise-robust speech recognition techniques [10]. It contains both clean and noisy speech data. The clean speech is the 8 kHz down-sampled TIDIGIT utterances. The noisy speech were obtained by artificially adding different types of noise to the clean data at various SNRs.

A. Speech enhancement performance on white noise

Fig. 2 shows the waveforms of clean speech, white noise contaminated speech, and the enhanced speech by different approaches. We can see from the figure that the SURE approach performs better than the other two methods in suppressing white noise. Fig. 3 illustrates the performance of the three methods in terms of segmental SNR improvement. The results are the average over the 1001 test sentences from AURORA 2 Test Set A. The noises are white Gaussian noise artificially added to clean speech at SNRs of 0, 5, 10, 15 and 20 dB. Here the computation of segmental SNR improvement covers only the speech frames detected by VAD. The SURE approach outperforms the E-M and the subspace approaches at low SNRs. For SNR of 15 dB and 20 dB, SURE and subspace method have comparable performance and both are significantly better than the E-M method.

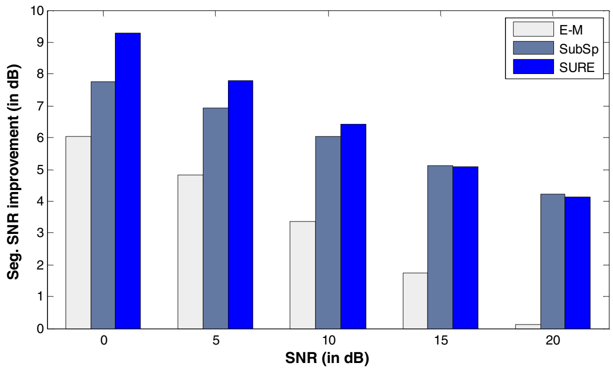


Fig. 3. Improvement of segmental SNR obtained by different approaches at various SNRs.

B. Speech recognition accuracy

The effectiveness of the SURE approach as a front-end processor for speech recognition is evaluated with the AURORA 2 database. We follow the standard AURORA experimental framework as described in [10], except that VAD and speech enhancement are performed before feature extraction.

Word accuracies on AURORA 2 Test Set A are given in Table I. The baseline front-end refers to that described in [10] plus a VAD. The front-end with SURE denoising is as described in Section III. We can see that the SURE front-end brings an overall absolute improvement of 7% in word accuracy. The improvement is more noticeable for low SNR conditions. At 20 dB SNR, the SURE front-end causes a slight degradation of 0.43% in word accuracy. Table I also shows that

TABLE I
WORD ACCURACIES FOR THE AURORA 2 RECOGNITION TASK (SET A). THE ACOUSTIC MODELS ARE TRAINED ON CLEAN DATA.

(a) Word accuracy (in %) by the baseline front-end.

	Subway	Babble	Car	Exhibition	Average
20dB	97.39	97.61	97.64	97.72	97.59
15dB	93.71	94.98	95.50	94.63	94.71
10dB	83.36	84.19	83.66	83.83	83.76
5dB	63.22	59.16	53.06	57.91	58.34
0dB	36.48	22.91	21.00	23.05	25.86
Average	74.83	71.77	70.17	71.43	72.05

(b) Word accuracy (in %) by the SURE denoising front-end.

	Subway	Babble	Car	Exhibition	Average
20dB	97.24	96.70	97.49	97.22	97.16
15dB	95.36	94.07	95.76	94.75	94.99
10dB	88.03	86.28	90.69	87.38	88.10
5dB	69.82	67.32	78.20	71.27	71.65
0dB	42.25	30.59	56.34	44.83	43.50
Average	78.54	74.99	83.70	79.09	79.08

TABLE II
WORD ACCURACIES FOR THE AURORA 2 RECOGNITION TASK WITH DIFFERENT DENOISING FRONT-ENDS.

(a) Average word accuracy (in %) for different SNRs.

	0 dB	5 dB	10 dB	15 dB	20 dB	Average
Baseline	25.86	58.34	83.76	94.71	97.59	72.05
E-M	34.98	64.49	84.09	93.28	96.49	74.67
SubSp	36.40	66.26	85.87	93.99	96.85	75.88
SURE	43.50	71.65	88.10	94.99	97.16	79.08

(b) Average word accuracy (in %) for different noise types.

	Subway	Babble	Car	Exhibition	Average
Baseline	74.83	71.77	70.17	71.43	72.05
E-M	76.50	71.62	76.81	73.74	74.67
SubSp	77.91	72.01	78.41	75.18	75.88
SURE	78.54	74.99	83.70	79.09	79.08

the SURE front-end brings more performance improvement for car and exhibition noises than for subway and babble noises. The car and exhibition noises are found to be relatively stationary as compared with the other two types of noise. In our experiments, we made an assumption of stationary noise and did not performance noise spectrum updating. We expect that greater benefit would be seen from the SURE approach if the noise statistics are updated dynamically.

Table II compares the word accuracy obtained from different denoising front-ends. The SURE front-end consistently outperforms the other two approaches at all SNRs and for all noise types.

V. CONCLUSION

We present a new speech enhancement approach based on the minimization of SURE-MSE, which does not require any hypothesis on the clean speech. The enhanced signal is obtained by thresholding signal coefficients in the DCT domain and the thresholding parameters are specified by minimizing the SURE-MSE. The effectiveness of the new approach is evaluated on a subset of the AURORA 2 database, in comparison with other two MMSE based enhancement approaches relying on the prior knowledge of the clean speech. Experimental results show that our method brings a substantial increase in both segmental SNR and speech recognition accuracy.

ACKNOWLEDGEMENT

This work is jointly supported by National Science Foundation of China (Project 60901061), National Science Foundation of Guangdong Province (9151806001000025).

REFERENCES

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

[2] Y. Ephraim and M. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, 1984.

[3] Y. Ephraim and M. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, 1985.

[4] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, 1995.

[5] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.

[6] I. Cohen, "Relaxed statistical model for speech enhancement and *a priori* SNR estimation," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 45, pp. 870–881, 2005.

[7] Collin Breithaupt and Rainer Martin, "DFT-based speech enhancement for robust automatic speech recognition," in *Proc. 8th ITG Conf. Speech Communication*, 2008.

[8] Thierry Blu and Florian Luisier, "The sure-let approach to image denoising," *IEEE Trans. Image Processing*, vol. 16, no. 11, pp. 2778–2786, 2007.

[9] Florian Luisier, Thierry Blu, and Michael Unser, "A new SURE approace to image denoising: interscale orthonormal wavelet thresholding," *IEEE Trans. Image Processing*, vol. 16, no. 3, pp. 593–606, 2007.

[10] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, 2000, pp. 181–188.

[11] C. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, pp. 1135–1151, 1981.