EMPLOYING LAPLACIAN-GAUSSIAN DENSITIES FOR SPEECH ENHANCEMENT

Saeed Gazor

Department of Electrical and Computer Engineering, Queen's University, Kingston, Ontario K7L 3N6, Canada

Abstract–A new efficient Speech Enhancement Algorithm (SEA) is developed in this paper. A noisy speech is first decorrelated and then the clean speech components are estimated from the decorrelated noisy speech samples. The distributions of clean speech and noise are assumed to be Laplacian and Gaussian, respectively. The clean speech components are estimated either by Maximum Likelihood (ML) or Minimum-Mean-Square-Error (MMSE) estimators. These estimators require some statistical parameters that are adaptively extracted by the ML approach during the active speech or silence intervals, respectively. A Voice Activity Detector (VAD) is employed to detect whether the speech is active or not. The simulation results show that this SEA performs as well as a recent high efficiency SEA that employs the Wiener filter. The complexity of this algorithm is very low compared with existing SEAs.

1. INTRODUCTION

Most of Speech Enhancement (SE) research has focused on removing the corrupting noise. Usually it is assumed that speech is degraded by additive noise which is independent of clean speech. In early implementations, spectral subtraction approach was widely used. This approach estimates the Power Spectral Density (PSD) of a clean signal by subtracting the short-time PSD of the noise from the PSD of the noisy signal [1]. The Wiener and Kalman filters have been used for SE [3, 4]. The noisy speech is used to estimate an "optimum" filter adaptively, under the assumption that speech and noise are Gaussian, independent and have zero mean. Recently a signal subspace speech enhancement framework has been developed (see [5–7] and references therein). This signal subspace SE system decomposes the noisy signal into uncorrelated components by applying the Karhunen-Loève Transform (KLT).

The recent statistical modelling presented in [8] concludes that the clean speech components, in decorrelated domains (*e.g.*, in the KLT and the DCT domains) as random variables have Laplacian distributions, and noise components are accurately modelled by Gaussian distributions. Therefore, the speech decorrelated components could be accurately modelled as a multivariate Laplacian random vector, while for noise a multivariate Gaussian model is accurate. Based on these assumptions, we design a Bayesian SE system to estimate the clean speech signal components.

This paper is organized as follows. Section 2 reviews the basic principle of the decorrelation of speech signals. Section 3 provides the statistical modelling that will be used in this paper. A SEA is proposed in Section 4. The performance evaluation and conclusion are summarized in Section 5. Section 6 is the conclusion.

2. DECORRELATION OF SPEECH SIGNALS

Let $x(t)$ be the clean speech and the vector of samples of $x(t)$ be denoted by $X(m) = [x(m), x(m-1), \cdots, x(m-K+1)]^T$ where $(.)^T$ is the transpose operation. Also, let $Y(m) = X(m) + N(m)$ denote the corresponding K-dimensional vector of noisy speech, assuming that the noise vector $N(m)$ is additive. Applying a linear transformation to the noisy the speech signal we may approximately assume that signal components are uncorrelated in the transformed domain. Since the correlation between speech signals is commonly rather high, a speech data vector can be represented with a small error by a small number of components [7]. In this paper, the speech signals are transformed into uncorrelated components by using the DCT or the AKLT [6]. It can be easily seen that $v_i(m) = s_i(m) + u_i(m)$, where $v_i(m)$, $s_i(m)$ and $u_i(m)$ are transformed components of $Y(m)$, $X(m)$ and $N(m)$, respectively. In order to develop our SEA, we further assume that the uncorrelated components, *i.e.*, $\{s_i, u_i\}_{i=1}^K$, are independent [8]. This assumption is not required if random variables uncorrelated are Gaussian variables and are uncorrelated. As the KLT is complex to compute, harmonic transforms such as Discrete Cosine Transform (DCT) and Discrete Fourier Transform (DFT) are used as suboptimal alternatives. Another motivation for using these transforms instead of the AKLT is to avoid the subspace variations and errors of the AKLT [6]. Among the DCT and DFT, the DCT is cheaper and reduces better the correlation of the signal and compacts the energy of a signal block into some of the coefficients.

3. STATISTICAL MODELLING

The noise samples can be separated during silence intervals using a VAD. For a Gaussian noise component $u_i(m)$ with variance $\sigma_i^2(m)$, we have: $f_{\mathbf{u}_{i,m}}(u_i(m)) = \frac{\exp(-\frac{u_i^2(m)}{2\sigma_i^2(m)})}{\sqrt{2\pi\sigma_i^2(m)}}$. If the samples $\{u_i(i)\}_{i=m-M_N+1}^m$ are iid, the ML estimate of σ_i^2 is $\widehat{\sigma_i^2} = \frac{1}{M_N}\sum_{t=m-M_N+1}^m |u_i(t)|^2$. A lower-complexity estimator is

$$\widehat{\sigma_i^2}(m) = \beta_N \widehat{\sigma_i^2}(m-1) + (1-\beta_N)|u_i(m)|^2, \qquad (1)$$

where β_N is chosen to let the time constant of the above filter be 0.5 second, assuming that the variation of the noise spectrum is negligible over a time interval of 0.5 second.

Assuming that the components of the clean speech in the decorrelated domains, $\{s_i(m)\}_{i=1}^K$, follow zero-mean Laplacian are uncorrelated, we have $f_{\mathbf{s}_{i,m}}(s_i(m)) = \frac{1}{2a_i(m)} e^{-\frac{|s_i(m)|}{a_i(m)}}$, where $a_i(m)$ is the Laplacian factor [8]. Similarly, the ML estimate of the Laplacian Factor a_i yields $\hat{a}_i(m) = \frac{1}{M_S}\sum_{t=m-M_S+1}^{m} |s_i(t)|$. Similarly, we use the following low-complexity substitute:

$$\hat{a}_i(m) = \beta_S \hat{a}_i(m-1) + (1-\beta_S)\,|s_i(m)|\,. \qquad (2)$$

In our simulations, β_S is chosen to let the time constant of the above adaptive process be 10msec, because the speech signal can

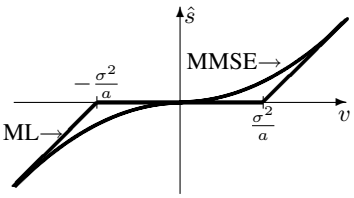

Fig. 1. Comparison of MMSE and ML estimators of the clean speech component s from the noisy input component $v = s + u$.

be assumed to be stationary over such a period. The estimation of a_i in (2) is equivalent to the expected value of $|s_i(m)|$. Note that there is no access to the clean signal, $s_i(m)$. If the noise power is small, we may use $|v_i(m)|$ as an approximation for $|s_i(m)|$ in (2).

4. ESTIMATION OF CLEAN SPEECH COMPONENTS

In this section for simplicity of notation, we drop the time index m and eigenvector index i. Here, the problem is to estimate the clean speech component s when the noisy speech component $v = s+u$ is given. Assuming that the speech is detected as present and the speech s and noise u components are independent, the joint distribution of s and v is given by

$$f_{\mathbf{s},\mathbf{v}}(s,v) = \frac{1}{2a\sqrt{2\pi\sigma^2}} e^{-\frac{|s|}{a} - \frac{|v-s|^2}{2\sigma^2}}, \tag{3}$$

and the conditional distribution of s given v is

$$f_{\mathbf{s}|\mathbf{v}}(s|v) = \frac{f_{\mathbf{s},\mathbf{v}}(s,v)}{\int_{s=-\infty}^{+\infty} f_{\mathbf{s},\mathbf{v}}(s,v)\, ds}. \tag{4}$$

The Minimum Mean Square Error (MMSE) Estimator is the conditional mean of s with $f_{\mathbf{s}|\mathbf{v}}(s|v)$ as the pdf. Using (3), the MMSE estimator of the clean speech component s, is given as a non-linear function of three inputs: 1) noisy speech component v, 2) noise variance σ^2 and 3) speech Laplacian factor a:

$$\mathbf{MMSE}: \quad \hat{s} \triangleq E\{s|v\} = \int_{-\infty}^{+\infty} s f_{\mathbf{s}|\mathbf{v}}(s|v)\, ds \tag{5}$$

$$= a e^{\frac{\psi}{2}} \left[\frac{(\psi+\xi)e^{\xi}\mathrm{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right) - (\psi-\xi)e^{-\xi}\mathrm{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right)}{e^{\xi}\mathrm{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right) + e^{-\xi}\mathrm{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right)} \right]$$

where $\xi = \frac{v}{a}$, $\psi = \frac{\sigma_i^2}{a_i^2}$ and $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt$.

The ML Estimator of s given the observation v is the value for which the likelihood function $f_{\mathbf{v}|\mathbf{s}}(v|s)$ is maximum. Maximizing $f_{\mathbf{v}|\mathbf{s}}(v|s)$ is equivalent to maximizing (3). Thus, we have

$$\mathbf{ML}: \quad \hat{s} \triangleq \arg\max_s f_{\mathbf{v}|\mathbf{s}}(v|s) = \arg\max_s f_{\mathbf{s},\mathbf{v}}(s,v) \tag{6}$$

$$= \arg\min_s \left(\frac{|s|}{a} + \frac{|v-s|^2}{2\sigma^2} \right) = \begin{cases} v - \frac{\sigma^2}{a}, & \text{if } v \geq \frac{\sigma^2}{a}, \\ 0, & \text{if } |v| \leq \frac{\sigma^2}{a}, \\ v + \frac{\sigma^2}{a}, & \text{if } v \leq -\frac{\sigma^2}{a}. \end{cases}$$

Figure 1 depicts the MMSE and ML estimators (5) and (6) versus the noisy input signal v for a given value of $\frac{\sigma^2}{a}$. We find that these estimators operate very similarly; if the amplitude of the noisy input is large (*i.e*, $|v| \gg 2\frac{\sigma^2}{a}$), then the magnitude of the output is almost equal to the magnitude of the input minus $\frac{\sigma^2}{a}$, *i.e*, $|\hat{s}| \simeq \max\left\{0, |v| - \frac{\sigma^2}{a}\right\}$. If the magnitude of the input v is smaller than $\frac{\sigma^2}{a}$ the ML interprets it as noise and projects it to zero, while MMSE attenuates the input.

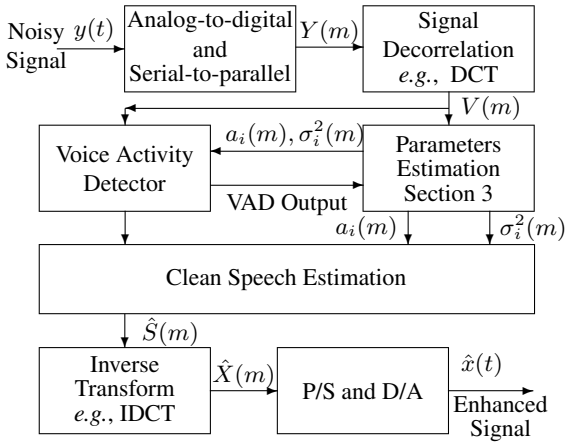

Fig. 2. The block diagram of the proposed SE system.

Figure 2 and Table 2 is a summarize the proposed SEA. We use the VAD in [9], since its design structure and assumptions are as same as those of this paper; therefore, the transformation and the estimation of the parameters could be shared between the proposed SE and this VAD. Providing the speech and noise statistic parameters, $\widehat{a}_i(m)$ and $\widehat{\sigma}_i^2(m)$, each DCT component will be used to estimate the clean speech component along the corresponding eigenvector using either ML estimation (5) or MMSE estimation (6). The enhanced signal is obtained via the inverse transform of $\hat{S}(m)$. The S/P is a buffer that takes the new portion of each new input vector and feeds it to a shift register in order to produce the stream of the enhanced signal $\hat{x}(t)$.

5. PERFORMANCE EVALUATION

The compromise between signal distortion and the level of residual noise is a well known problem in SE [3, 5]. In this section, the performance of the proposed SEA is evaluated using objective criteria, such as noise reduction criterion and distortion criterion.

The sampling frequency of the noisy speech signal is 11,025 Hz. The vector length, K, is chosen to be 80 samples, which corresponds to approximately 7msec. The overlap between signal vectors is set at 70 samples. The overlap can be reduced to reduce the computation complexity, at the expense of some performance degradation. In this case, at each iteration 10 samples of the enhanced signal is updated. This represents about 1msec of the signal. Further reduction of this updating time interval provides only a very slight improvement. Software generated white Gaussian noise, computer fan noise and lab noise are added to the original speech signal with different SNRs, namely 0, 5, and 10dB. The computer fan noise is picked up by a microphone, as is the lab noise which is mainly the sound of a network switch.

Time Domain and Spectrogram Evaluation: First, the results of the proposed SE algorithm are evaluated and compared with the SEA in [6] in the time domain and the frequency domain by means of the spectrogram. Figure 3 shows the results of enhanced speech corrupted by a 5dB white noise. From this figure we observe that the enhanced speech has a lower noise level in the time

Table 1. Summary of the soft VAD algorithm in [9]

Initialize: $\beta_S = 0.913, \quad \beta_N = 0.983, \quad P_{1|0} = \frac{1}{2},$
For each time step m do: $v_i(m) = [\mathrm{dct}\{Y(m)\}]_i,$

For i= 1, 2,$\cdots, K$, if $P_{m|m-1} \geq 0.5$ do

$$a_i(m) = \beta_S a_i(m-1) + (1-\beta_S)\,|v_i(m)|$$

if else do $\sigma_i^2(m) = \beta_N \sigma_i^2(m-1) + (1-\beta_N) u_i^2(m)$ end;

$$f_{0i}(m) = \frac{1}{\sqrt{2\pi\sigma_i^2(m)}} e^{-\frac{v_i^2(m)}{2\sigma_i^2(m)}}$$

$$f_{1i}(m) = \frac{e^{\frac{\psi_{i,m}}{2}}}{4a_i(m)} \left[e^{\xi_{i,m}} \mathrm{erfc}\left(\frac{\psi_{i,m} + \xi_{i,m}}{\sqrt{2\psi_{i,m}}} \right) + \right.$$

$$\left. + e^{-\xi_{i,m}} \mathrm{erfc}\left(\frac{\psi_{i,m} - \xi_{i,m}}{\sqrt{2\psi_{i,m}}} \right) \right],$$

where $\xi_{i,m} = \dfrac{v_i(m)}{a_i(m)}$, and $\psi_{i,m} = \dfrac{\sigma_i^2(m)}{a_i^2(m)}$

end;

$$L(m) = \prod_{i=1}^{K} \frac{f_{1i}(m)}{f_{0i}(m)}$$

$$P_{m|m} = \frac{L(m) P_{m|m-1}}{L(m) P_{m|m-1} + \left(1 - P_{m|m-1}\right)},$$

$$P_{m+1|m} = \Pi_{01}\left(1 - P_{m|m}\right) + \Pi_{11} P_{m|m}.$$

end;

Table 2. The structure of the proposed SEA.

Initialization: $d_i(0) = 0$, $\beta_S = 0.913$, $\beta_N = 0.983$,
β_S and β_N are chosen to let the time constants of filters to be 10msec and 0.5sec, respectively.
For each time step m do
$V(m) = [v_1(m), v_2(m), \cdots, v_K(m)]^T \leftarrow \mathrm{dct}\{Y(m)\}$
For i= 1, 2,$\cdots, K$ do
$\widehat{a}_i(m) \leftarrow$ from (2)
if speech is absent: $\widehat{\sigma}_i^2(m) \leftarrow$ from (1) end;
MMSE: $\hat{s}_i(m) \leftarrow$ from (5),
or ML: $\hat{s}_i(m) \leftarrow$ from (6)
end; end; $\hat{X}(m) \leftarrow \mathrm{idct}\{[\hat{s}_1(m), \hat{,} \cdots, \hat{s}_K(m)]\}$

domain, where the ML approach results in a lower residual noise level. From the spectrograms in Figure 3 it can be seen also that the background noise is very efficiently reduced, while the energy of most of the speech components remained unchanged.

Figures 3 and 4 illustrate the results for a nonstationary, colored lab noise. From our simulations and Figures 3, 3 and 4, we conclude that the proposed methods perform very well for various noise conditions such as for colored and/or nonstationary noises.

The estimation of a_i has an important impact on the performance of the proposed SEAs. To illustrate this impact, we estimate the Laplacian factor a_i using the clean speech signal and call this estimate as "*best value*" of a_i. In Figure 5, noisy speech is enhanced with these so-called *best values*. We will use the term "*best value*" to refer to the SEA that processes the noisy speech with these so-called *best values*, which theoretically provides the "best" performance that can be achieved with this SE framework under the Laplacian-Gaussian assumption. We can clearly see that the residual noise level of this ideal case is much lower than the results from Figure 3. This illustrates the effectiveness of the SEA.

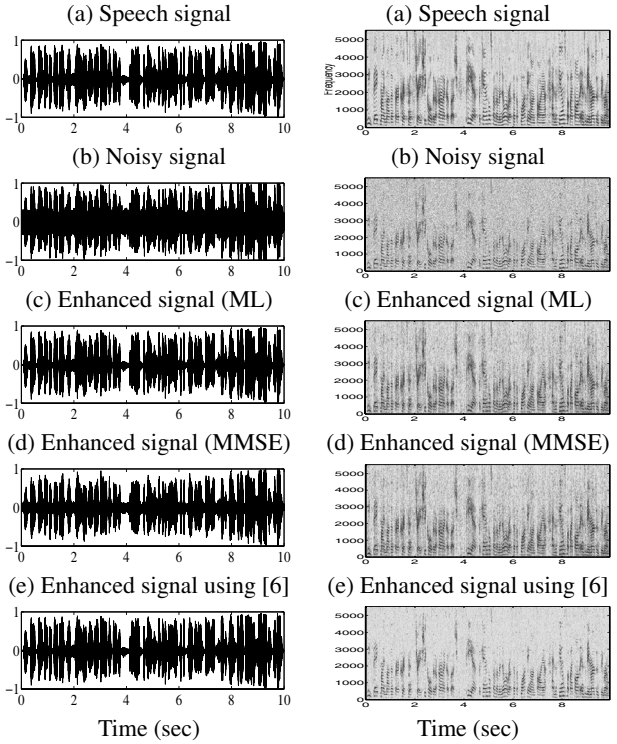

Fig. 3. Enhanced speech corrupted by white Gaussian noise (SNR=5dB), and corresponding spectrograms.

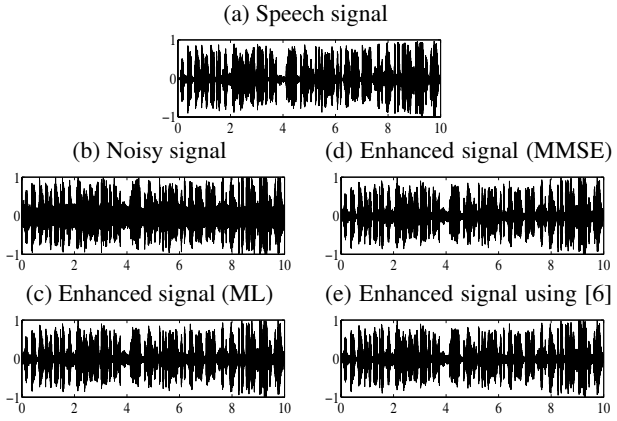

Fig. 4. Enhanced speech corrupted by nonstationary colored lab noise (SNR=5dB).

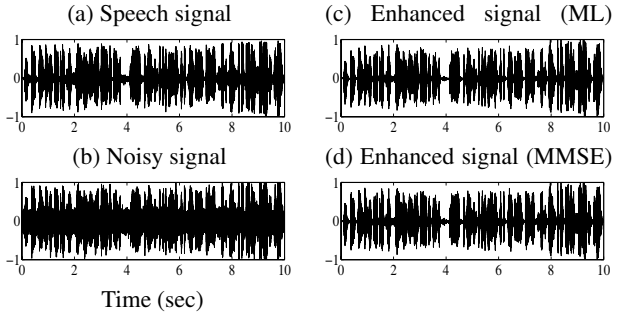

Fig. 5. SEA results with the "*best value*", white noise, SNR=5dB.

Table 3. Spectral Distortion between the clean signal and signals enhanced using different SEAs for various noise conditions.

Noise Type	Input SNR	Input SD	Proposed SEA MMSE	Proposed SEA ML	*best value* of a_i MMSE	*best value* of a_i ML	SEA in [6]
White	0dB	5.85	5.67	5.78	5.41	5.68	5.70
Gaussian	5dB	4.84	4.58	4.67	4.27	4.73	4.59
Noise	10dB	3.78	3.53	3.57	3.40	3.75	3.53
Lab	0dB	5.98	5.58	5.52	5.01	5.53	5.52
Noise	5dB	4.91	4.58	4.54	4.11	4.51	4.49
	10dB	3.81	3.54	3.55	3.24	3.53	3.49
Computer	0dB	4.88	4.12	4.34	4.96	4.59	4.29
Fan	5dB	3.73	3.26	3.53	3.50	3.77	3.47
Noise	10dB	2.71	2.56	2.82	2.81	3.02	2.77

Table 4. Comparison of SNR (in dB) of enhanced signals for various noise conditions.

Noise Type	Input SNR	Proposed SEA MMSE	Proposed SEA ML	*best value* SEA MMSE	*best value* SEA ML	SEA in [6]
White	0dB	4.30	4.87	6.05	5.38	5.22
Gaussian	5dB	8.26	8.51	9.24	8.67	8.72
Noise	10dB	12.47	12.47	12.87	12.32	12.57
Lab	0dB	4.26	5.10	6.49	5.73	5.40
Noise	5dB	8.27	8.64	9.45	8.64	8.78
	10dB	12.33	12.32	12.84	11.99	12.37
Computer	0dB	5.74	5.90	6.48	6.15	6.04
Fan	5dB	9.87	9.82	10.26	9.84	9.94
Noise	10dB	13.46	13.29	13.57	13.16	13.37

Spectral Distortion: We use Spectral Distortion (SD) as a criterion for the performance of SEAs. The SD in decibels (dB) between two signals $x(t)$ and $y(t)$ with length N is defined by

$$\mathrm{SD}(x(t);y(t)) = \frac{1}{4N}\sum_{i=1}^{\frac{N}{64}}\sum_{k=0}^{255} 20\,|\log_{10}|X_p(k)| - \log_{10}|Y_p(k)||.$$

where $X_p(k)$ and $Y_p(k)$ are the kth FFT frequency components of the pth frame of $\frac{x(t)}{||x(t)||}$ and $\frac{y(t)}{||y(t)||}$, respectively. Signals are divided into frames of length 64 samples without overlapping. After padding 192 zeros into each frame, the 256-point FFT is calculated. Table 3 presents SDs where the clean speech signal is compared with the noisy input signal, two proposed enhanced signals and the enhanced signal using the algorithm in [6]. In white and lab noise conditions, the SD values for all these approaches are better than the noisy speech. The result of the "*best value*" MMSE approach is slightly better than those of the others. Only for the fan noise in a high SNR condition does the SD result seem to be unexpected. The reason is that the PSD of the fan noise has an strong peak at a low frequency that results in a strong SD around this frequency.

The Output SNR of the enhanced signal (or the noisy signal) $y(t)$ is defined by $\mathrm{SNR} = 10\log_{10}\frac{\sum_{k=1}^{N} x^2(k)}{\sum_{k=1}^{N}(y(k)-x(k))^2}$. Table 4 compares the enhanced signals using different approaches versus the input SNR. As expected, the SNR performance of the "*best value*" MMSE is the best (highest) in all noise conditions. The SNR improvement in the MMSE approach for high SNRs is higher than that of other approaches.

6. CONCLUSION

An SEA is developed based on a Laplacian distribution for speech and a Gaussian distribution for additive noise signals. The enhancement is performed in a decorrelated domain. Each component is estimated from the corresponding noisy speech component by applying a non-linear memoryless filter. The speech signal is decomposed into uncorrelated components by the DCT or by the adaptive KLT. It is assumd that the speech is stationary within 20-40msec and the noise is stationary over a longer period of about 0.5sec. The proposed SEAs are based on the MMSE and the ML approaches, respectively. The speech is then synthesized by the IDCT or IKLT. Overall, proposed SEAs effectively reduce the additive noise. At the same time, the proposed SEAs produce a lower level of distortion in the enhanced speech when compared with the method in [6] that uses a complex Adaptive KLT. The comparison of results with the method in [6] shows that the proposed SEAs provide a better (or similar) performance. The performance criteria of the proposed SEAs give similar results. The fact that the SEAs with "best value" outperformed all the others, indicates that the new proposed framework for SE could be further improved.

The computational complexity of the proposed SEAs is very low compared with the existing algorithms because of the use of fast DCT. In fact, most of the computationally complex parts are the DCT and IDCT (the computational complexity the DCT and IDCT is of the order of $K\log_2(K)$, where K is the size of the vectors). All our simulations and listening evaluations confirm that the proposed methods are very useful for SE.

7. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[2] S. M. McOlash, R. J. Niederjohn, and J. A. Heinen, "A Spectral Subtraction Method for the Enhancement of Speech Corrupted by Nonwhite, Nonstationary Noise," *Proc. of IEEE Int. Conf. on Industrial Electronics, Control, and Instrumentation,* vol. 2, pp. 872-877, 1995.

[3] I. Y. Soon and S. N. Koh, "Low Distortion Speech Enhancement," *IEE Proceedings-Vision, Image and Speech Processing*, vol. 147, no. 3, pp. 247-253, June 2000.

[4] Z. Goh, K.-C. Tan, and B. T. G. Tan, "Kalman-Filtering Speech Enhancement Method Based on a Voiced-unvoiced Speech Model," *IEEE Transactions on Speech and Audio Processing,* vol. 7, no. 5 , pp. 510-524, Sept. 1999.

[5] U. Mittal and N. Phamdo, "Signal/Noise KLT Based Approach for Enhancing Speech Degraded by Colored Noise," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 2 , pp. 159-167, March 2000.

[6] A. Rezayee and S. Gazor, "An Adaptive KLT Approach for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 2, pp. 87-95, Feb. 2001.

[7] Y. Ephraim and H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 251-266, July, 1995.

[8] S. Gazor and W. Zhang, "Speech Probability Distribution", *IEEE, Signal Processing Letters*, vol. 10, no. 7, pp. 204-207, July 2003.

[9] S. Gazor and W. Zhang "A soft voice activity detector based on a laplacian-gaussian model", *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 498-505, Sept. 2003.