Speech Enhancement Employing Laplacian–Gaussian Mixture

Saeed Gazor, *Senior Member, IEEE,* and Wei Zhang

***Abstract*—A new, efficient speech enhancement algorithm (SEA) is developed in this paper. In this low-complexity SEA, a noisy speech signal is first decorrelated and then the clean speech components are estimated from the decorrelated noisy speech samples. The distributions of clean speech and noise signals 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 derived from speech and noise. These parameters are adaptively extracted by the ML approach during the active speech or silence intervals, respectively. In addition, a voice activity detector (VAD) that uses the same statistical model is employed to detect whether the speech is active or not. The simulation results show that our SEA approach performs as well as a recent high efficiency SEA that employs the Wiener filter. The computational complexity of this algorithm is very low compared with existing SEAs with low computational complexity.**

***Index Terms*—Adaptive Karhunen–Loève transform, adaptive signal detection, adaptive signal processing, colored noise, decorrelated domains, decorrelation, decorrelation transformation, discrete cosine transforms, Gaussian distribution, generalized GD, Karhunen–Loève transforms, Laplacian distribution, Laplacian-Gaussian Mixture, Laplacian random variables, linear minimum mean squared error estimation, marginal distributions, minimum mean squared error estimation, maximum likelihood estimation, multivariate distribution approximation, non-Gaussian distribution, nonlinear speech enhancement, speech activity detection, speech enhancement, speech probability distribution, speech processing, speech quality evaluation, speech samples distribution, speech signal statistics, time-varying speech components energy.**

I. Introduction

OVER the past four decades, the problem of speech enhancement (SE) has been discussed by many researchers [1], [2], [8], [9]. The main objective of SE is to improve the performance of speech communication systems in a noisy environment. Depending on the specific application, the objective of an enhancement system may be to improve the overall quality, increase intelligibility, reduce listener fatigue, or a combination of these.

Most of SE research has focused on removing the corrupting noise, which improves the overall quality of the speech signal. 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 PSD of the noise from the PSD of the noisy signal [2], [3]. The estimate of PSD is performed within short time segments, because the short-time spectral amplitude carries important information about both speech quality and intelligibility.

Almost all of the known speech enhancement algorithms, which operate in the transform domains, assume that the coefficients of both the noisy speech and noise are all jointly zero mean Gaussian distributed random variables in the transform domain. Such a assumption results in a linear estimator for the clean speech signal. However, it has been remarked that assuming some other distributions can result in better performance than the Gaussian model [16], [17]. A linear estimator is obviously suboptimal where the Gaussian distribution is not the best candidate for the data. For instance in [8], [9], the enhancement of speech in the log-domain (cepstrum) is considered. Due to a nonlinear blending of noise and clean speech, the distribution of speech and noise is non-Gaussian. Employing a mixture model as approximation for the pdfs considerable improvement is achieved in [8], [9].

The Wiener filter has been used for SE [4]. The noisy speech is used to estimate an "optimum" filter adaptively, under the assumption that speech and noise signals are independent and have zero mean Gaussian distributions. The Wiener filter could be applied in either the time domain or the frequency domain to obtain an estimate of the clean speech. The Kalman filter has also been used to estimate the clean speech signal [5]. It is "theoretically" optimal in the minimum-mean-square-error sense if the speech and noise have a joint linear Gaussian dynamic.

Recently a signal subspace speech enhancement framework has been developed (see [6], [7], [10], [18], and references therein). In this framework, the estimation is performed on a frame-by-frame basis under the assumption that the noise is additive and uncorrelated with the clean speech signal. This signal subspace SE system decomposes the noisy signal into uncorrelated components by applying the Karhunen–Loève Transform (KLT). So along each eigenvector, the component of noisy speech is the sum of the components of clean speech and noise. An estimation of each clean speech component is made. Then the clean signal is synthesized by applying the inverse KLT (IKLT) to the estimated clean speech vectors.

The recent statistical modeling presented in [11], [12] concludes that the clean speech components, in decorrelated domains (e.g., in the KLT and the Discrete Cosine Transform (DCT) domains [14]) as random variables have Laplacian distributions, and noise components are accurately modeled by Gaussian distributions. Therefore, the speech decorrelated components could be accurately modeled as a multivariate Laplacian

Manuscript received July 18, 2002; revised August 6, 2004. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Li Deng.

The authors are with the Department of Electrical and Computer Engineering, Walter Light Hall, Queen's University, Kingston, ON, K7L 3N6, Canada (e-mail: saeed.gazor@ece.queensu.ca; wzhang@ece.queensu.ca).

Digital Object Identifier 10.1109/TSA.2005.851943

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. Since speech signals are not stationary, the parameters of this system are adaptively calibrated. The Adaptive KLT (AKLT) is the first alternative that attempts to fully decorrelate the signals (see [6], [7], [10], and [18]). The DCT is another computationally inexpensive alternative that transforms acoustic signals into reasonably decorrelated components [14].

This paper is organized as follows. In Section II, we review the basic principle of the decorrelation of speech signals. Section III provides the statistical modeling that will be used in this paper. The main subject of this paper is a new estimation algorithm of clean speech components, proposed in Section IV. In Section V, we give the structure of a SEA system. The performance evaluation and conclusion are summarized in Section VI. Section VII is the conclusion.

II. Decorrelation of Speech Signals

In this section, we consider a speech signal in the decorrelated domain. Let $x(t)$ be the clean speech signal. A K-dimensional vector of samples of $x(t)$ at time m is denoted by

$$X(m) = [x(m), x(m-1), \ldots, x(m-K+1)]^T \quad (1)$$

where $(.)^T$ denotes the transpose operation. Also, let $Y(m)$ denote the corresponding K-dimensional vector of noisy speech. Assuming that the noise vector $N(m)$ is additive, we have

$$Y(m) = X(m) + N(m). \quad (2)$$

In [10], a *linear model* for the speech signal is described that approximates a vector of noisy signal with a linear combination of some basis vector. 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. In this paper, the speech signals are transformed into uncorrelated components by using the DCT or the AKLT [7], [14]. 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 assume that the uncorrelated components, i.e., $\{s_i, u_i\}_{i=1}^{K}$, are independent. This assumption is strongly justified by intuition for a set of uncorrelated Laplacian random variables. However, for the Gaussian case it can be easily proven, i.e., Gaussian random variables are independent if they are uncorrelated.

A. Karhunen–Loève Transform (KLT)

Let $R_X(m) = E[X(m)X^T(m)]$ be the covariance matrix of clean speech $X(m)$, and consider the eigendecomposition of $R_X(m)$ to be as follows:

$$R_X(m) = W(m)\Lambda_X(m)W^T(m) \quad (3)$$

where $\Lambda_X(m)$ is a diagonal matrix containing the eigenvalues of the clean speech covariance matrix $R_X(m)$. The matrices $W^T(m)$ and $W(m)$ are called the KLT and the Inverse KLT (IKLT) of the clean signal $X(m)$, respectively [7]. In fact, the main property of the KLT is that the covariance matrix of the transformed signal, $X_W(m) = W^T(m)X(m)$, is diagonal, i.e., $\Lambda_X(m) = E[X_W(m)X_W^T(m)]$. The column span of $W(m)$ corresponding to nonzero eigenvalues is referred to as signal subspace. The KLT and IKLT are unitary transformations, i.e., $W^T(m)W(m) = I$.

In most subspace-based speech-processing algorithms, the noise subspace components are first assumed to be white, i.e., $R_N(m) \simeq \lambda_N(m)I$, where $\lambda_N(m)$ is the variance of noise [10]. In this case, the covariance matrix of $W^T(m)N(m)$ should also equal to $\lambda_N(m)I$ because the matrix $W^T(m)$ is unitary. Practical results show that each component of $W^T(m)N(m)$ has a different variance. A better approximation for the covariance matrix of the noise components in the KLT domain $W^T(m)N(m)$ is as follows [7]:

$$\begin{aligned}\Lambda_N(m) &= W^T(m)R_N(m)W(m) \\ &= \mathrm{diag}(\lambda_{1,N}(m),\ \lambda_{2,N}(m), \ldots,\ \lambda_{K,N}(m))\end{aligned} \quad (4)$$

where $\lambda_{i,N}(m)$ is the variance of noise along the i^{th} eigenvector at time m. It is reasonable to assume that the noise $N(m)$ is uncorrelated with and independent of the speech $X(m)$. In this case, the covariance matrix of noisy speech $Y(m)$ is given by

$$\begin{aligned}R_Y(m) &= E[Y(m)Y^T(m)] = R_X(m) + R_N(m) \\ &= W(m)\left(\Lambda_X(m) + \Lambda_N(m)\right)W^T(m).\end{aligned} \quad (5)$$

This means that the eigenvectors of $R_X(m)$ and $R_Y(m)$ are the same in the presence of speech.

Accurate estimates of eigenvectors and eigenvalues of $R_Y(m)$ are required in the speech processing algorithms based on subspace approaches. As the speech signals are not stationary processes, adaptive subspace tracking algorithms could be applied here, e.g., the algorithm suggested in Table I [7].

B. Discrete Cosine Transform (DCT)

The KLT is optimal for transform coding of Gaussian sources. It is an orthonormal transform that produces uncorrelated coefficients. As the KLT is complex to compute, harmonic transforms such as 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. Among the DCT and DFT, the DCT yields a better performance, and is computationally less expensive. Thus, it is preferred in practice.

The DCT also reduces the correlation of the signal and compacts the energy of a signal block into some of the DCT coefficients. The results in [14] illustrate that the efficiency of the DCT in whitening an autoregressive signal (or speech) is as good as the KLT if the data size is large enough.

III. Statistical Modelling

The statistical modeling in [11], [12] leads to the conclusion that the DCT and the KLT components of speech follow Laplacian distributions more accurately than Gaussian distributions. Our experimental results show that most noise signals can be precisely modeled by Gaussian distributions in the transformed domain. In this section, we review the model of noise components and clean speech components.

TABLE I
ADAPTIVE KLT TRACKING ALGORITHM FOR SEA [7]

Initialize: $d_i(0) = 0,\ \beta = 0.9132,$
$W(0) = [w_1(0)|w_2(0)|\cdots|w_K(0)] = I_K,$
The β is chosen for a time constant of 10msec in the following.
For each time step m do

$$\begin{aligned}
Y_1(m) &= Y(m)\\
N_1(m) &\leftarrow \text{Read from noise memory}\\
&\text{For } i\text{= 1, 2,}\cdots\text{, } K \text{ do}\\
v_i(m) &= w_i^T(m-1)Y_i(m)\\
u_i(m) &= w_i^T(m-1)N_i(m)\\
d_i(m) &= \beta d_i(m-1) + |v_i(m)|^2\\
E_i(m) &= Y_i(m) - w_i(m-1)v_i(m)\\
w_i(m) &= w_i(m-1) + E_i(m)\frac{v_i(m)}{d_i(m)}\\
Y_{i+1}(m) &= Y_i(m) - w_i(m)v_i(m)\\
&\text{end;}\\
W(m) &= [w_1(m)|w_2(m)|\cdots|w_K(m)]
\end{aligned}$$

end;

A. Noise Distribution

We assume that the noise components in uncorrelated domains, $\{u_i(m)\}_{i=1}^K$, are Gaussian, i.e.,

$$f_{\mathbf{u}_i}(u_i(m)) = \frac{1}{\sqrt{2\pi\sigma_i^2(m)}} e^{-u_i^2(m)/2\sigma_i^2(m)},\ \forall i = 1, 2, \ldots K \tag{6}$$

where $\sigma_i^2(m)$ is the variance of the i^{th} noise component. If the successive samples of $u_i(m)$, during the silence interval between $(m - M_N + 1)$ and m, are independent and the variations of their variances are very small, then the Maximum Likelihood estimate of σ_i^2 is given by

$$\widehat{\sigma_i^2} = \frac{1}{M_N} \sum_{t=m-M_N+1}^{m} |u_i(t)|^2. \tag{7}$$

We use the following low-complexity substitute for (7):

$$\widehat{\sigma_i^2}(m) = \beta_N \widehat{\sigma_i^2}(m-1) + (1-\beta_N)|u_i(m)|^2. \tag{8}$$

In our simulations, β_N is chosen to let the time constant of the above filter be 0.5 s, assuming that the variation of the noise spectrum is negligible over a time interval of 0.5 s. Because the speech signal includes some silence intervals, the noise samples can be always separated from these time intervals and stored using a VAD.

B. Speech Signal Modeling

In [11], [12], we demonstrated that the clean speech components in the decorrelated domains, $\{s_i(m)\}_{i=1}^K$, have zero-mean Laplacian distributions and are uncorrelated, i.e., the pdfs of $s_i(m)$ are given by

$$f_{\mathbf{s}_i}(s_i(m)) = \frac{1}{2a_i(m)} e^{-|s_i(m)|/a_i(m)},\quad \forall i = 1, 2, \ldots K \tag{9}$$

where $a_i(m)$ is the Laplacian factor for the i^{th} clean speech component. Similarly, by applying the Maximum Likelihood estimate of the Laplacian Factor a_i yields

$$\hat{a}_i(m) = \frac{1}{M_S} \sum_{t=m-M_S+1}^{m} |s_i(t)|. \tag{10}$$

Similarly, we use the following low-complexity substitute for (10)

$$\hat{a}_i(m) = \beta_S \hat{a}_i(m-1) + (1-\beta_S)\,|s_i(m)|\,. \tag{11}$$

In our simulations, β_S is chosen to let the time constant of the above adaptive process be 10 ms, because the speech signal can be assumed to be stationary over such a period.

The estimation of a_i in (11) 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 (11). For low SNRs, our experimental results show that the substraction of the estimated noise variance $\sigma_i^2(m-1)$ from $|v_i(m)|^2$ and using the following estimator lead to further SNR enhancement:

$$\begin{aligned}
a_i(m) = {} & \beta_S a_i(m-1)\\
& + (1-\beta_S)\sqrt{\max\{|v_i(m)|^2 - \sigma_i^2(m-1), 0\}}.
\end{aligned} \tag{12}$$

IV. ESTIMATION OF CLEAN SPEECH COMPONENTS

In this section, the estimators of clean speech components are presented, based on the statistical distributions given in the previous section.

Assuming that different components are independent, we can process different components in parallel. 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 is given. Assuming that the speech is detected as present, we have

$$v = s + u \tag{13}$$

where

$$f_{\mathbf{s}}(s) = \frac{1}{2a} e^{-|s|/a} \quad \text{and} \quad f_{\mathbf{u}}(u) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-u^2/2\sigma^2}. \tag{14}$$

It is reasonable to assume that the speech s and noise u components are independent; therefore, the joint distribution of s and v is given by

$$f_{\mathbf{s},\mathbf{v}}(s, v) = \frac{1}{2a\sqrt{2\pi\sigma^2}} e^{-|s|/a - |v-s|^2/2\sigma^2} \tag{15}$$

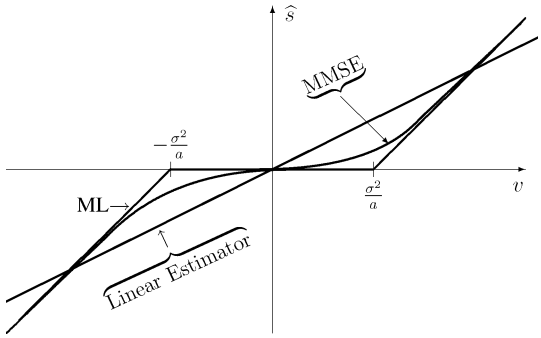

Fig. 1. Comparison of a linear estimator with two proposed nonlinear estimators (MMSE and ML) of the clean speech signal component s from the noisy input component $v = s + u$.

and the conditional distribution of s given v is

$$f_{\mathbf{s}|\mathbf{v}}(s|v) = \frac{f_{\mathbf{s},\mathbf{v}}(s,v)}{\int_{s=-\infty}^{+\infty} f_{\mathbf{s},\mathbf{v}}(s,v)\,ds}. \tag{16}$$

Subsequently, we derive two estimators for s.

A. Minimum Mean Square Error (MMSE) Estimator

This MMSE estimator is the conditional mean of s with $f_{\mathbf{s}|\mathbf{v}}(s|v)$ as the pdf. Using (15), the MMSE estimator of the clean speech component s, is given as a nonlinear function of three inputs: 1) noisy speech component v, 2) noise variance σ^2, and 3) speech Laplacian factor a

$$\textbf{MMSE:}\quad \hat{s} \triangleq E\{s|v\} = \int_{-\infty}^{+\infty} s f_{\mathbf{s}|\mathbf{v}}(s|v)\,ds$$
$$= ae^{\psi/2}\left[\frac{(\psi+\xi)e^{\xi}\mathrm{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right)-(\psi-\xi)e^{-\xi}\mathrm{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right)}{e^{\xi}\mathrm{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right)+e^{-\xi}\mathrm{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right)}\right] \tag{17}$$

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

B. Maximum Likelihood (ML) Estimator

The ML estimate 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 obviously equivalent to maximizing (15). Thus, we have

$$\textbf{ML:}\quad \hat{s} \triangleq \arg\max_s f_{\mathbf{v}|\mathbf{s}}(v|s) = \arg\max_s f_{\mathbf{s},\mathbf{v}}(s,v)$$
$$= \arg\min_s \left(\frac{|s|}{a} + \frac{|v-s|^2}{2\sigma^2}\right)$$
$$= \begin{cases} v - \frac{\sigma^2}{a}, & \text{if } v \geqslant \frac{\sigma^2}{a}, \\ 0, & \text{if } |v| \leqslant \frac{\sigma^2}{a}, \\ v + \frac{\sigma^2}{a}, & \text{if } v \leqslant -\frac{\sigma^2}{a}. \end{cases} \tag{18}$$

The MMSE and ML estimators in (17) and (18) are depicted in Fig. 1 versus the noisy input signal v for a given value of σ^2/a. We find that these estimators operate very similarly; if the amplitude of the noisy input is large (*i.e*, $|v| \gg 2\sigma^2/a$), then the magnitude of the output is almost equal to the magnitude of the input minus σ^2/a, *i.e*, $|\hat{s}| \simeq \max\{0, |v| - \sigma^2/a\}$. If the magnitude of the input v is smaller than σ^2/a the ML interprets it as noise and projects it to zero, while MMSE attenuates the input. To avoid the computational complexity involved in MMSE, one may suggest using a simpler nonlinear function.

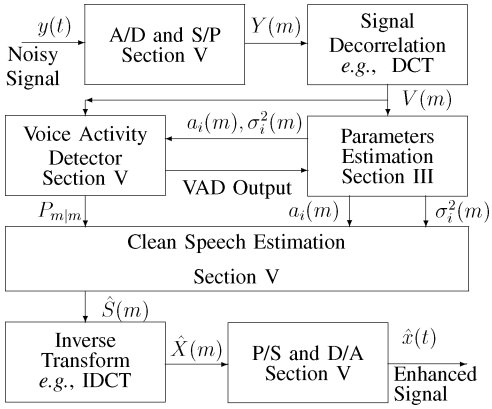

Fig. 2. Block diagram of the proposed SE system.

V. Proposed SE System

In this section, components of this system are introduced. Fig. 2 is a block diagram for the proposed SEA. In Section III, two Bayesian estimator of the clean speech component $s_i(m)$ are proposed based on the following assumptions.

1) Speech components in the transformed domains (i.e., DCT or AKLT) have Laplacian distributions and are independent.[1]
2) Noise components have Gaussian distributions.
3) Signal and noise are independent.

The following information needs to be provided to these estimators:

- noisy speech component $v_i(m)$;
- speech Laplacian factor $a_i(m)$;
- noise variance $\sigma_i^2(m)$.

Both $a_i(m)$ and $\sigma_i^2(m)$ can be estimated using the ML estimation method given in (8) and (12) (for high SNRs, one may use (11) instead of (12)). The clean speech component can be estimated with MMSE or ML estimations as shown in Section IV. The two proposed nonlinear estimators are derived based on above assumptions. It is reported that the first assumption is considerably better than a Gaussian model if the transformation is applied to time frames of at least longer than 10 ms [11], [12]. The second assumption is valid if the noise power is almost stationary over several successive frames. For instance, this is a reasonable assumption for computer fan noise and car noise and is not reasonable for bobble noise [15]. In Fig. 1, these clean speech estimators are also compared with a linear suboptimal one, i.e., with the Winner filter that is the MMSE under a Gaussian-Gaussian model which is a suboptimal estimator under our assumptions.

[1]It is demonstrated that speech signal components are sufficiently uncorrelated in these domains [14].

TABLE II
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

$$V(m) = [v_1(m), v_2(m), \cdots, v_K(m)]^T \leftarrow \begin{cases} \mathrm{dct}\{Y(m)\} \\ \text{or} \\ \mathrm{klt}\{Y(m)\} \end{cases}$$

For i= 1, 2,$\cdots$, K do
if speech is present:
$a_i(m) \leftarrow$ from (12) and $\sigma_i^2(m) = \sigma_i^2(m-1)$
if speech is absent:
$\sigma_i^2(m) \leftarrow$ from (8) and $a_i(m) = a_i(m-1)$
end if;
MMSE: $\hat{s}_i(m) \leftarrow$ from (17),
or ML: $\hat{s}_i(m) \leftarrow$ from (18)
end;
end; $\hat{X}(m) \leftarrow \mathrm{idct}\{[\hat{s}_1(m), \hat{s}_2(m), \cdots, \hat{s}_K(m)]\}$

The SEA is shown in Table II. The first step is to decompose the noisy speech vectors to uncorrelated components. Providing the speech and noise statistic parameters, each DCT component will be used to estimate the clean speech component along the corresponding eigenvector using (17) and (18). Then the the enhanced speech signal is obtained via the inverse transform. The synthesis process applies the inverse transform of the decorrelation process to the cleaned signal vector $\hat{S}(m)$. The enhanced speech vector is given by

$$\hat{X}(m) = W(m)\,\hat{S}(m) \tag{19}$$

where $\hat{S}(m) = [\hat{s}_1(m), \hat{s}_2(m), \ldots, \hat{s}_K(m)]^T$ is the vector of estimated (enhanced) speech components in the decorrelated domain which could be obtained either from ML estimation or MMSE estimation. The overlap between successive input vectors $X(m)$ controls a trade off between computational complexity and the performance. To create the stream of the output, we only need some samples of $\hat{X}(m)$; therefore, to reduce the computational complexity, only required samples in (19) are needed to be calculated.

The estimations of the speech and noise parameters are performed on a frame-by-frame basis. The speech Laplacian factors $a_i(m)$ and the noise variance $\sigma_i^2(m)$ may be obtained with the ML approach given in Section III. Using these parameters, SE is performed with (17) or with (18) in MMSE and ML senses, respectively.

Analog-to-Digital and Serial-to-Parallel: After analog-to-digital (A/D) conversion, the signal is passed through a serial-to-parallel (S/P) convertor, by using a tap delay line to obtain the noisy speech signal vector $Y(m)$. This vector will be then transformed by the DCT, KLT, or another decorrelation transformation.

Voice Activity Detector: A voice activity detector (VAD) is required to separate silence intervals from voice activity intervals. We suggest using the VAD in [13] that processes the decorrelated speech samples. This soft VAD is obtained from a Bayesian hypotheses test by assuming the same statistical modeling that is a Laplacian distribution for speech and a Gaussian distribution for the additive noise. In addition, this VAD employs a hidden Markov model with two states representing silence and active speech. The probability of speech being active is estimated recursively by this soft VAD that is summarized in Table III. The *a priori* state probability provided from the previous time instance, is combined with the new observation to calculate the probability of speech being active at a given moment. A typical simulation result for a noisy speech signal (5 dB white noise) is shown in Fig. 3. We adopt this VAD because it provides very good and robust performance for a wide range of signal-to-noise-ratios (SNR) and noise types.

Another advantage of this VAD algorithm is that its design structure and assumptions are the 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.

Parallel-to-Serial (P/S) and Digital-to-Analogue (D/A): Overlapping vectors of the signal are enhanced. Then a P/S and an A/D converter are used to produce the stream of the enhanced signal. The P/S is a buffer that takes the new portion of each new input vector and feeds it to a shift register followed by an D/A in order to produce the stream of the enhanced signal $\hat{x}(t)$.

VI. PERFORMANCE EVALUATION

The compromise between signal distortion and the level of residual noise is a well known problem in SE [4], [6]. 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 7 ms. 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 are updated. This represents about 1 ms 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 10 dB. The computer fan noise is picked up by a microphone, as is the lab noise which is mainly the sound of a network switch.

A. Time Domain and Spectrogram Evaluation

First, the results of the proposed SE algorithm are evaluated and compared with the SEA in [7] in the time domain and the frequency domain by means of the spectrogram. Fig. 4 shows the results of enhanced speech corrupted by a white noise with 5 dB SNR. From this figure we observe that the enhanced speech

TABLE III
SUMMARY OF A SOFT VAD ALGORITHM SIMILAR TO THE PROPOSED ONE IN [13]

Initialize: $\beta_S = 0.913, \quad \beta_N = 0.983, \quad P_{1|0} = \frac{1}{2},$

For each time step, $m = 1, 2, \cdots$ do

For i= 1, 2,$\cdots$, K do

$v_i(m) \leftarrow i^{\text{th}}$ component of the DCT of $Y(m)$

if $P_{m|m-1} \geqslant 0.5$ then

$a_i(m) = \beta_S a_i(m-1) + (1-\beta_S)\sqrt{\max\{|v_i(m)|^2 - \sigma_i^2(m-1), 0\}}$

$\sigma_i^2(m) = \sigma_i^2(m-1)$

else

$\sigma_i^2(m) = \beta_N \sigma_i^2(m-1) + (1-\beta_N) z_i^2(m)$

$a_i(m) = a_i(m-1)$

end;

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

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

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

end;

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

$P_{m|m} = \frac{L(m) P_{m|m-1}}{L(m) P_{m|m-1} + (1 - P_{m|m-1})}, \qquad \leftarrow$ soft-output

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

end;

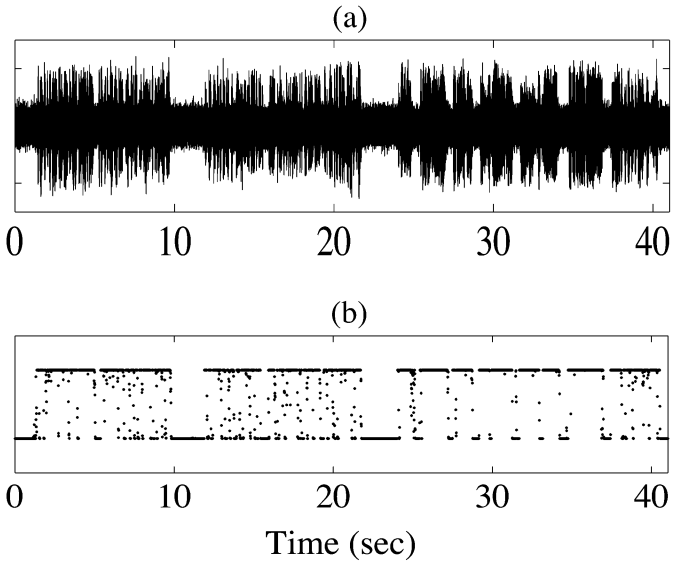

Fig. 3. Results of the proposed VAD in the presence of white noise with 5 dB SNR. (a) Noisy speech and (b) soft detection result, $P_{m|m}$ of the proposed VAD.

has a lower noise level in the time domain, where the ML approach results in a lower residual noise level. From the spectrograms in Fig. 5 it can be seen also that the background noise is very efficiently reduced, while the energy of most of the speech components remained unchanged.

Figs. 5 and 6 illustrate the results for a nonstationary, colored lab noise. From our simulations and Figs. 4–6, 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 Fig. 7, 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 Fig. 4. This illustrates that the effectiveness of the proposed SE could be further improved.

B. Spectral Distortion

We use Spectral Distortion (SD) as a criterion for the performance of SEAs [6]. The SD between two signals $x(t)$ and $y(t)$ with length N is calculated as follows. First, both signals are normalized, i.e., $\tilde{x}(t) = x(t)/\|x(t)\|$ and $\tilde{y}(t) = y(t)/\|y(t)\|$. Signals are normalized to neglect the gain of the algorithm that could be compensated simply by an amplifier. Then $\tilde{x}(t)$ and $\tilde{y}(t)$ are both divided into frames of length 64 samples

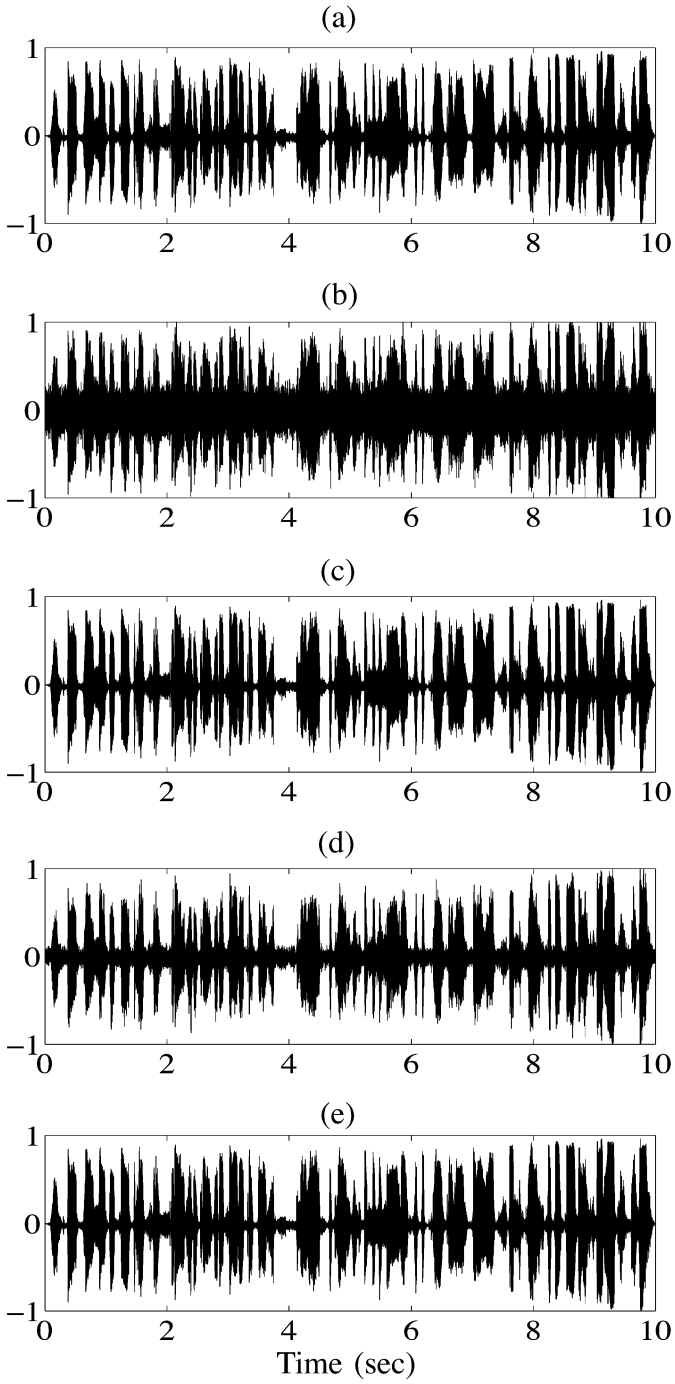

Fig. 4. Enhanced speech corrupted by white Gaussian noise ($\mathrm{SNR} = 5$ dB). (a) Speech signal, (b) noisy signal, (c) enhanced signal (ML), (d) enhanced signal (MMSE), and (e) enhanced signal (SEA in [7]).

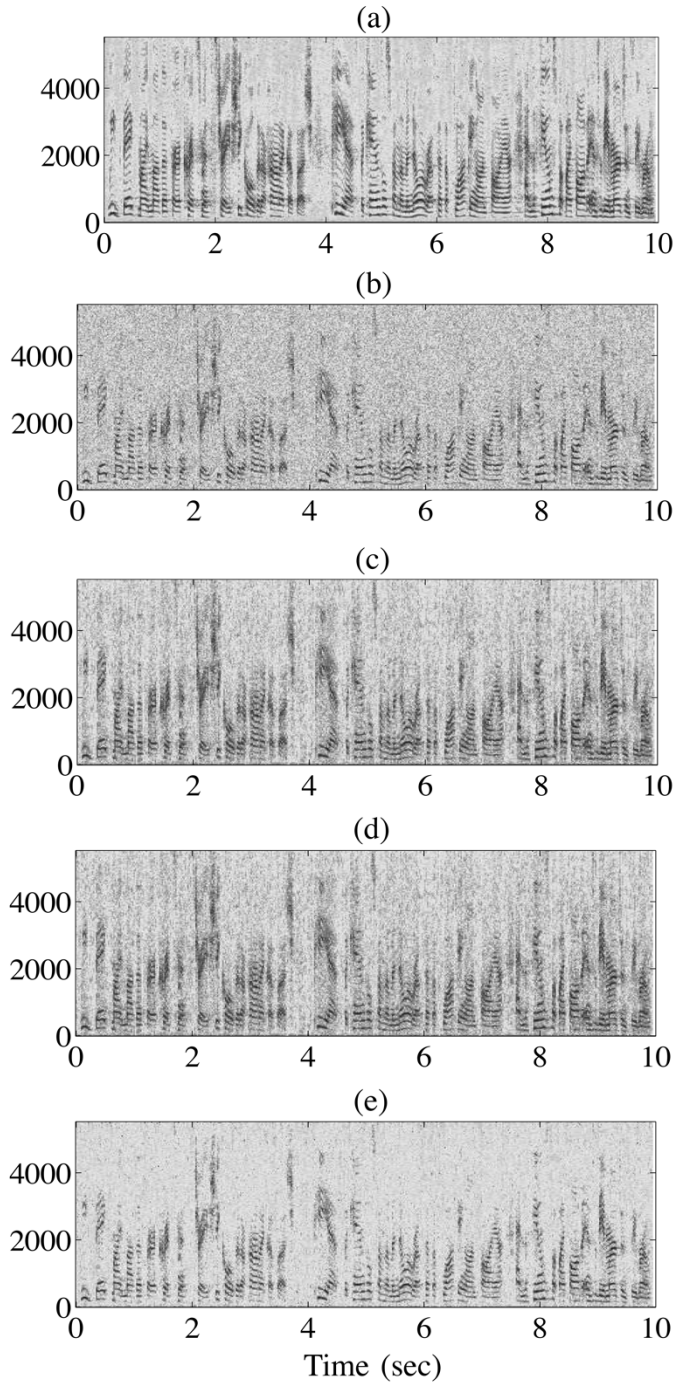

Fig. 5. Spectrograms of enhanced speech corrupted by white Gaussian noise ($\mathrm{SNR} = 5$ dB). (a) Speech signal, (b) noisy signal, (c) enhanced signal (ML), (d) enhanced signal (MMSE), and (e) enhanced signal (SEA in [7]).

without overlapping. After padding 192 zeros into each frame, the 256-point FFT is calculated for each frame. Let $X_p(k)$ and $Y_p(k)$ be the kth frequency components of the pth frame of $\hat{x}(t)$ and $\tilde{y}(t)$, respectively. The SD in decibels is defined as follows:

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

Table IV 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 [7]. 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 very low frequency that results in a strong SD around this frequency.

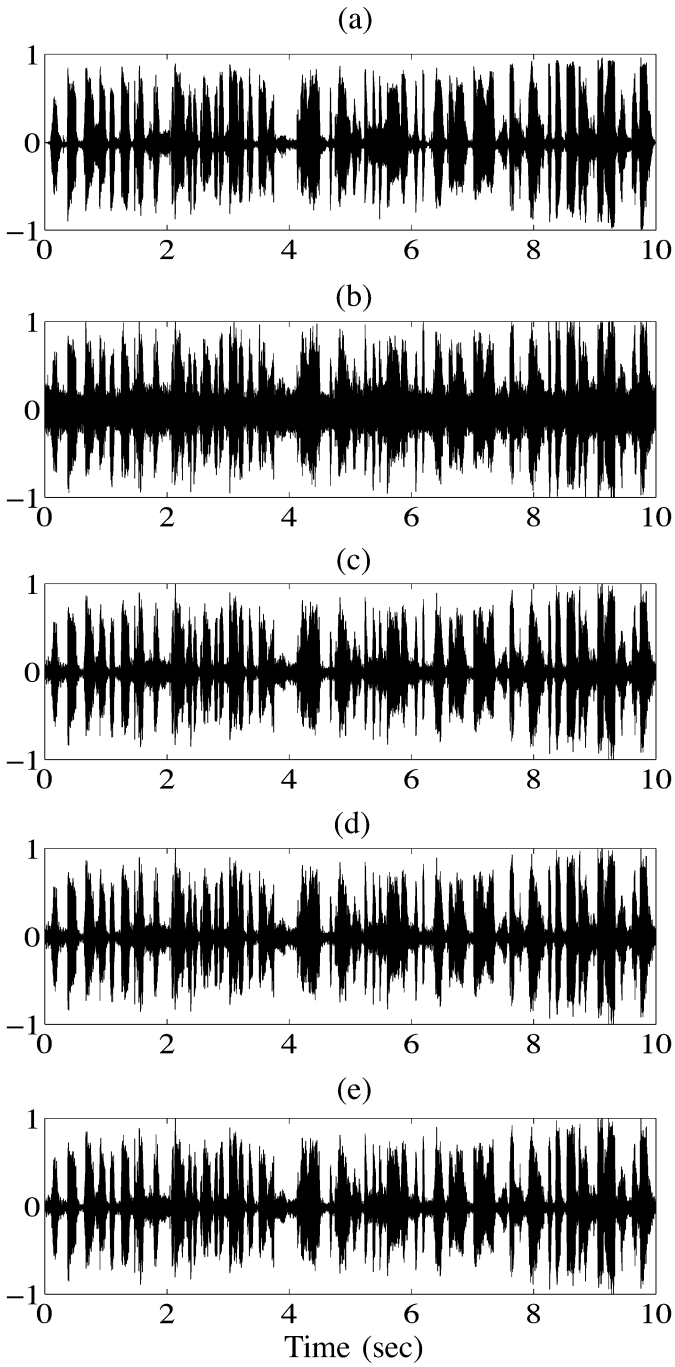

Fig. 6. Enhanced speech corrupted by nonstationary colored lab noise ($\mathrm{SNR} = 5$ dB). (a) Speech signal, (b) noisy signal, (c) enhanced signal (ML), (d) enhanced signal (MMSE), and (e) enhanced signal (SEA in [7]).

C. Output SNR

The SNR of the enhanced signal (or the noisy signal) $y(t)$ is defined by

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{k=1}^{N} x^2(k)}{\sum_{k=1}^{N}\left(y(k)-x(k)\right)^2} \tag{21}$$

where $x(t)$ is the clean speech signal and N is the number of signal samples. Table V compares the SNR of 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.

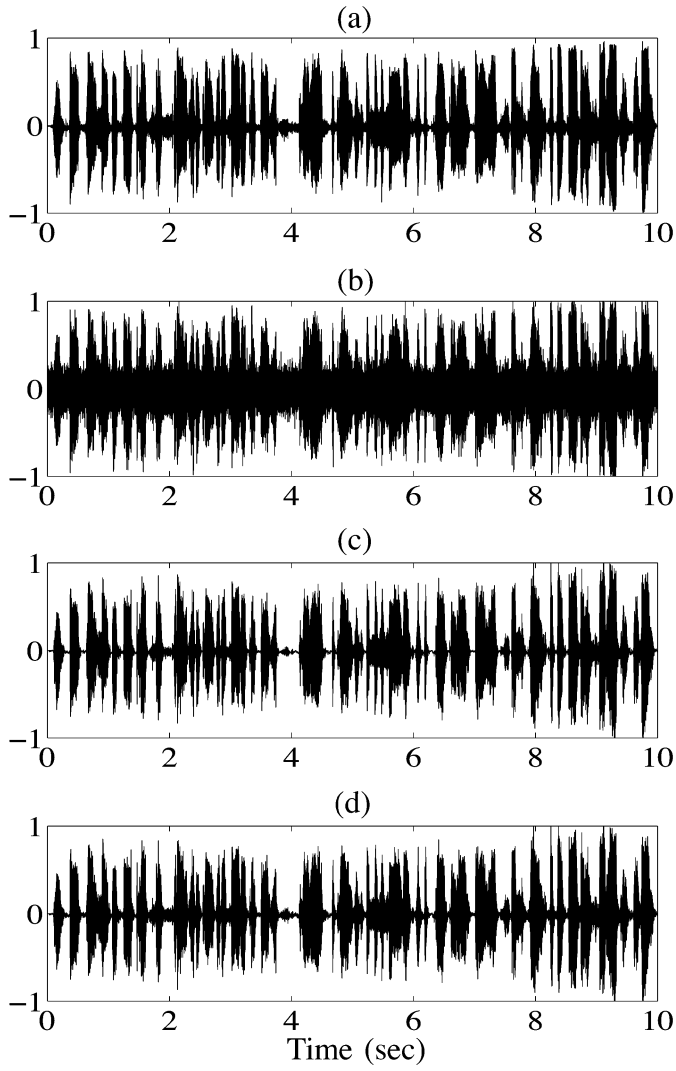

Fig. 7. Results of SEA with the "*best value*" (white noise, $\mathrm{SNR} = 5$ dB). (a) Speech signal, (b) noisy signal, (c) enhanced signal (ML), and (d) enhanced signal (MMSE).

TABLE IV
SPECTRAL DISTORTION (IN dB) BETWEEN THE CLEAN SIGNAL AND SIGNALS ENHANCED USING DIFFERENT SEAS FOR VARIOUS NOISE CONDITIONS

Noise Type	Input SNR	Input SD	Proposed SEA		*best value* of a_i		SEA in [7]
			MMSE	ML	MMSE	ML	
White	0dB	5.85	5.61	5.75	5.41	5.68	5.70
Gaussian	5dB	4.84	4.51	4.68	4.27	4.73	4.59
Noise	10dB	3.78	3.50	3.62	3.40	3.75	3.53
Lab	0dB	5.98	5.46	5.52	5.01	5.53	5.52
Noise	5dB	4.91	4.48	4.53	4.11	4.51	4.49
	10dB	3.81	3.48	3.54	3.24	3.53	3.49
Computer	0dB	4.88	4.16	4.36	4.96	4.59	4.29
Fan	5dB	3.73	3.28	3.57	3.50	3.77	3.47
Noise	10dB	2.71	2.58	2.85	2.81	3.02	2.77

TABLE V
COMPARISON OF SNR (IN dB) OF ENHANCED SIGNALS FOR VARIOUS NOISE CONDITIONS

Noise Type	Input SNR	Input SD	Proposed SEA		*best value* of a_i		SEA in [7]
			MMSE	ML	MMSE	ML	
White Gaussian Noise	0dB	5.85	5.61	5.75	5.41	5.68	5.70
	5dB	4.84	4.51	4.68	4.27	4.73	4.59
	10dB	3.78	3.50	3.62	3.40	3.75	3.53
Lab Noise	0dB	5.98	5.46	5.52	5.01	5.53	5.52
	5dB	4.91	4.48	4.53	4.11	4.51	4.49
	10dB	3.81	3.48	3.54	3.24	3.53	3.49
Computer Fan Noise	0dB	4.88	4.16	4.36	4.96	4.59	4.29
	5dB	3.73	3.28	3.57	3.50	3.77	3.47
	10dB	2.71	2.58	2.85	2.81	3.02	2.77

VII. CONCLUSION

A comprehensive framework for SE is developed based on a Laplacian distribution for speech and a Gaussian distribution for additive noise signals. The enhancement is performed by estimating the clean speech components from a Laplacian plus Gaussian mixture in a decorrelated domain. Each component is estimated from the corresponding noisy speech component by applying a nonlinear memoryless filter.

The speech signal is decomposed into uncorrelated components by the DCT or adaptively by the KLT. The estimates are obtained based on the information of statistical models for speech and noise components. This assumes that the speech is stationary within 20–40 ms and the noise is stationary over a longer period of about 0.5 s. 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 [7] that uses a complex Adaptive KLT. The comparison of results with the method in [7] 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.

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[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] S. M. McOlash, R. J. Niederjohn, and J. A. Heinen, "A spectral subtraction method for the enhancement of speech corrupted by nonwhite, non-stationary noise," in *Proc. 1995 IEEE IECON 21st Int. Conf. Industrial Electronics, Control, and Instrumentation*, vol. 2, 1995, pp. 872–877.

[4] I. Y. Soon and S. N. Koh, "Low distortion speech enhancement," *Proc. Inst. Elect. Eng.*, vol. 147, no. 3, pp. 247–253, Jun. 2000.

[5] Z. Goh, K.-C. Tan, and B. T. G. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 510–524, Sep. 1999.

[6] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 2, pp. 159–167, Mar. 2000.

[7] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 2, pp. 87–95, Feb. 2001.

[8] L. Deng, J. Droppo, and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 3, pp. 218–233, May 2004.

[9] ——, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 2, pp. 133–143, Mar. 2004.

[10] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[11] S. Gazor and R. R. Far, "Probability distribution of speech signal envelope," in *Proc. IEEE Can. Conf. Electrical and Computer Engineering, (CCECE'04)*, May 2004.

[12] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Lett.*, vol. 10, no. 7, pp. 204–207, Jul. 2003.

[13] ——, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 498–505, Sep. 2003.

[14] I. Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech Commun.*, vol. 24, no. 3, pp. 249–257, 1998.

[15] J.-H. Chang, S. Gazor, N. S. Kim, and S. K. Mitra, "Soft decision speech enhancement using a multiple statistical modeling," J. Signal Process., Apr. 2004, submitted for publication.

[16] J.-H. Chang and N. S. Kim, "Speech enhancement using warped discrete cosine transform," in *Proc. IEEE Speech Coding Workshop*, Tsukuba, Japan, Oct. 2002.

[17] C. Breithaupt and R. Martin, "MMSE estimation of magnitude-squared DFT coefficients with superGaussian priors," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, ICASSP '03*, vol. 1, Apr. 2003, pp. 896–899.

[18] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 700–708, Nov. 2003.

Saeed Gazor (S'94–M'95–SM'98) received the B.Sc. degree in electronics and the M.Sc. degree in communication systems both from Isfahan University of Technology in 1987 and 1989, respectively. He received the Ph.D. degree in signal and image processing from Telecom Paris, Département Signal (École Nationale Supérieure des Télécommunications/ENST PARIS), France, in 1994 (all with highest honors).

From 1995 to 1998, he was with the Department of Electrical and Computer Engineering, Isfahan University of Technology. From January 1999 to July 1999, he was with the Department of Electrical and Computer Engineering, University of Toronto. He is currently Associate Professor with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada. His main research interests are array signal processing, statistical and adaptive signal processing, speech processing, analog adaptive circuits, communication systems and information theory.

Dr. Gazor received the Premier's Research Excellence Award of the Province of Ontario in 2004. He is a Registered Professional Engineer in Ontario, Canada.

Wei Zhang received the B. Eng. degree in information engineering from the Harbin Institute of Technology, Harbin, China, and the M.S. (Eng.) degree in electrical engineering from Queen's University, Kingston, ON, Canada, in 1999 and 2002, respectively. He is currently pursuing the Ph. D. degree in the School of information technology and engineering, University of Ottawa, Ottawa, ON. His research interests include speech processing, especially for speech modeling and speech enhancement, adaptive signal processing, and general area in digital signal processing.