An Adaptive KLT Approach for Speech Enhancement

Afshin Rezayee and Saeed Gazor, *Senior Member, IEEE*

***Abstract*—An adaptive Karhunen–Loeve transform (KLT) tracking-based algorithm is proposed for enhancement of speech degraded by colored additive interference. This algorithm decomposes noisy speech into its components along the axes of a KLT-based vector space of clean speech. It is observed that the noise energy is disparately distributed along each eigenvector. These energies are obtained from noise samples gathered from silence intervals between speech samples. To obtain these silence intervals, we proposed an efficient voice activity detector based on outputs of principle component eigenfilter; the greatest eigenvalue of speech KLT. Enhancement is performed by modifying each KLT component due to its noise and clean speech energies. The objective is to minimize the produced distortion when residual noise power is limited to a specific level. At the end, inverse KLT is performed and an estimation of the clean signal is synthesized. Our listening tests indicated that 71% of our subjects preferred the enhanced speech by the above method over former methods of enhancement of speech degraded by computer generated white Gaussian noise. Our method was preferred by 80% of our subjects when we processed real samples of noisy speech gathered from various environments.**

***Index Terms*—Adaptive estimation, adaptive filters, adaptive speech processing, adaptive voice activity detection, music quality enhancement, speech enhancement, speech subspace tracking.**

I. Introduction

TO PROVIDE better quality and performance of several applications, the enhancement of speech should be considered [1]–[14]. Speech enhancement attempts to improve one or more perceptual aspects of voice communication systems when the signal is corrupted by noise (e.g., overall quality, intelligibility for human or machine recognizers). The improvement is in the sense of minimizing the effects of the noise on the system performance.

The speech enhancement problem consists of a family of problems characterized by the type of noise source, the way the noise interacts with the clean signal, the number of voice channels or microphone outputs available for enhancement, and the nature of the speech communication system. This paper focuses on enhancement of speech signals that are degraded by statistically independent additive noise. Furthermore, the output of only one microphone which contains the noisy signal is assumed available.

In a pioneering contribution, Lim and Oppenheim [2] formulated the speech enhancement problem. There are two major classes of speech enhancement algorithms: 1) the class based on hidden Markov modeling (HMM) of noise and/or speech signals [6], [9], [11] and 2) the class based on transformation of signals.

The design and performance of the first generation of the HMM-based methods are highly dependent on speech signal syntax and noise characteristics. Hence, usage of first generation HMM-based systems was not found effective for the very noisy input SNR of 5 dB or less [1]. To solve this problem, the second generation of the HMM algorithms are developing by current researches (e.g., [9]–[11]) to estimate adaptively (usually based on expectation maximization) the characteristics of noise (and/or speech).

In the second class, the spectral subtraction estimation approach and its derivatives are efficient methods that were developed by various scientists such as [2]–[4]. A complete version of this approach was introduced by Kang and Fransen [5]. Spectral subtraction speech enhancement method suffers from a self producing noise, named musical noise. This noise is produced by the spectral subtraction method in low SNR environments. Of course, the implementation simplicity and good performance of the spectral subtraction method has caused it to be used in voice communication systems for a long time. Our paper belongs to this class.

Recently a signal subspace speech enhancement system has been proposed by Ephraim and Van-Trees [7]. The key idea in signal subspace speech enhancement system is to consider the signal as a vector in K-dimensional space and to decompose/transform noisy speech signal into uncorrelated components [7], [13]. In this way, one is able to obtain improved noise reduction. Each component contains a clean signal part plus a noise part. An estimation of the clean signal part is made for each component. Then the clean signal is synthesized by applying the Inverse KLT (IKLT) to the estimation of the clean signal parts. In [7], it is assumed that the additive noise is white. This system had good performance in simulations when noise was computer generated white Gaussian, but does not enhance as well for nonstationary colored noise as for white stationary noise. This is due to the fact that the decomposed components have various noise variances while they are assumed to be equal. So, we assumed noise to be colored and developed a more realistic clean signal estimator. In our proposed scheme, a vector of a sampled noisy speech signal is transformed by the KLT of the clean speech signal. The noise component of the transformed vector is assumed to have a diagonal covariance matrix. In practical situations, the assumed model in this paper is more accurate and matches noise behaviors better than the white noise model.

Manuscript received August 24, 1998; revised April 25, 2000. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jean-Claude Junqua.

A. Rezayee was with the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran. He is now with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada M5S 3G4 (e-mail: afshin@eecg.toronto.edu).

S. Gazor was with the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran. He is now with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, K7L 3N6 Canada.

Publisher Item Identifier S 1063-6676(01)00851-3.

Also, since the covariance matrix of the transformed noise is assumed to be diagonal, implementation of the proposed estimator will not be complicated. At the end, according to the proposed noise model an estimation of the clean signal is made for each component and the clean signal is synthesized using IKLT.

Another improvement we made was using an adaptive algorithm for performing KLT. In [7], a block of data is used to estimate noisy speech covariance matrix. Then, an Eigenvalue Decomposition (ED) is applied to perform KLT. This approach requires repeated ED computation that is a very time consuming task. In addition, speech and noise are nonstationary processes and hence adaptive KLT tracking algorithms are preferred. We used a new type of KLT tracking algorithm named projection approximation subspace tracking [15]–[17]. In the projection approximation method ED is considered as a constrained optimization problem and an adaptive algorithm is used to find a very close approximation of the eigenvectors of the clean speech covariance matrix using noisy speech signal samples. This algorithm is fast and has a simple construction. According to our simulations this kind of subspace tracking algorithm has a good behavior in speech enhancement.

This paper is organized as follows. In Section II the principles of the KLT approach for speech enhancement is introduced. Implementation of this method is reviewed in Section III. Some experimental results are discussed in Section IV. Conclusion and comments are given in Section V.

II. KLT-Based Speech Enhancement System

The problem considered in this paper is to extract the clean speech signal $x(t)$ from the degraded speech signal $y(t)$. A K-dimensional vector of speech samples $X(n)$ is defined as follows:

$$X(n) = \left[x\left(\frac{n}{f_s}\right), x\left(\frac{n-1}{f_s}\right), \cdots, x\left(\frac{n-K+1}{f_s}\right)\right]^T \tag{1}$$

where f_s is the sampling frequency and $(.)^T$ denotes transpose operation. Also, let $Y(n)$ denote the corresponding K-dimensional vector of noisy speech. Since noise is assumed to be additive, we have

$$Y(n) = X(n) + N(n) \tag{2}$$

where $N(n)$ is the K-dimensional noise vector. Now, let $H(n)$ be a $K \times K$ linear estimator of clean speech vector as follows:

$$\hat{X}(n) = H(n)Y(n). \tag{3}$$

The error signal obtained in this estimation is given by

$$\begin{aligned}\mathcal{E}(n) &= \hat{X}(n) - X(n), \\ &= (H(n) - I)X(n) + H(n)N(n), \\ &\triangleq \mathcal{E}_X(n) + \mathcal{E}_N(n)\end{aligned} \tag{4}$$

where $\mathcal{E}_X(n) \triangleq (H(n) - I)X(n)$ represents signal distortion and $\mathcal{E}_N(n) \triangleq H(n)N(n)$ represents residual noise [7]. Define the energies of signal distortion $\overline{\epsilon_X^2}(n)$ and residual noise $\overline{\epsilon_N^2}(n)$, respectively, as follows:

$$\begin{aligned}\overline{\epsilon_X^2}(n) &= \mathrm{tr}(E[\mathcal{E}_X(n)\mathcal{E}_X^T(n)]), \\ &= \mathrm{tr}((H(n) - I)R_X(n)(H(n) - I)^T)\end{aligned} \tag{5}$$

and

$$\begin{aligned}\overline{\epsilon_N^2}(n) &= \mathrm{tr}(E[\mathcal{E}_N(n)\mathcal{E}_N^T(n)]), \\ &= \mathrm{tr}(H(n)R_N(n)H^T(n))\end{aligned} \tag{6}$$

where $R_X(n)$ and $R_N(n)$ are covariance matrixes of clean signal and noise vector, respectively. Now, assuming that $R_X(n)$ and $R_N(n)$ are provided, we minimize the distortion energy while maintaining the noise energy less than a positive threshold level in each time step. Thus, the optimum linear estimator is defined as follows [7]:

$$\begin{aligned}&H_{\mathrm{opt}}(n) \triangleq \arg\{\min_{H(n)} \overline{\epsilon_X^2}(n)\}, \\ &\text{Subject to: } \frac{1}{K}\overline{\epsilon_N^2}(n) \le \sigma^2\end{aligned} \tag{7}$$

where σ^2 is a positive constant. Reducing noise in a given noisy speech signal causes some distortion in speech component. In the above optimization, by decreasing noise threshold level σ^2, we can decrease the amount of residual noise and therefore, increase the amount of distortion and vise versa. In this case we are allowed to decide whether residual noise or amount of distortion be decreased at the expense of increasing the other one. The optimal estimator in the sense of (7) can be found using Kuhn–Tucker necessary conditions for constrained minimization [19]. Specifically, $H(n)$ is a stationary feasible point if it satisfies the gradient equation of Lagrangian

$$L(H(n), \mu) \triangleq \overline{\epsilon_X^2}(n) + \mu\left(\overline{\epsilon_N^2}(n) - K\sigma^2\right) \tag{8}$$

and

$$\left(\overline{\epsilon_N^2}(n) - K\sigma^2\right) = 0 \qquad \text{for } \mu \ge 0 \tag{9}$$

where μ is the Lagrangian multiplier. From $\nabla_{H(n)}L(H(n), \mu) = 0$, and (5) and (6) we obtain

$$H_{\mathrm{opt}}(n) = R_X(n)(R_X(n) + \mu R_N(n))^{-1}. \tag{10}$$

From (9), it is obtained that μ should satisfy

$$\begin{aligned}\sigma^2 = \frac{1}{K}\,\mathrm{tr}(&R_X(n)(R_X(n) + \mu R_N(n))^{-1}R_N(n) \\ &\cdot (R_X(n) + \mu R_N(n))^{-1}R_X(n)).\end{aligned} \tag{11}$$

It is seen from (11) that varying μ from 0 to ∞ causes σ^2 to vary from zero until $(1/K)\,\mathrm{tr}(R_N(n))$. It is clear that the proposed estimator does not add noise to the estimated signal. So, the noise threshold level σ^2 should be smaller than the input noise energy level $(1/K)\,\mathrm{tr}(R_N(n))$. Now let eigenvalue decomposition of $R_X(n)$ be defined as follows:

$$R_X(n) = U(n)\Lambda_X(n)U^T(n) \tag{12}$$

where

$\Lambda_X(n)$ diagonal $K \times K$ matrix that contains clean speech covariance matrix eigenvalues;

$U(n)$ contains its eigenvectors called in literature the IKLT and the unitary matrix;

$U^T(n)$ called KLT.

In fact, the property of KLT is that the covariance matrix of $U^T(n)X(n)$ is diagonal, i.e., $\Lambda_X(n)$. The column span of $U(n)$ corresponding to nonzero eigenvalues is called signal subspace and is a unitary transformation matrix. Substituting (12) in (10) we have

$$\begin{aligned} H_{\text{opt}}(n) =& U(n)\Lambda_X(n)(\Lambda_X(n) + \mu U^T(n)R_N(n)U(n))^{-1} \\ &\cdot U^T(n). \end{aligned} \tag{13}$$

In [7] noise is assumed to be white, i.e., $R_N(n) \simeq \lambda_N(n)I$, where $\lambda_N(n)$ is the variance of white noise in each time step. From this assumption, one can say that the covariance matrix of $U^T(n)N(n)$ (transformation of the noise vector) is also equal to $\lambda_N(n)I$. But, using computer simulations we found out that each component of $U^T(n)N(n)$ has a different variance. Hence, we assumed a more realistic approximation model for the noise as follows:

$$\begin{aligned} \Lambda_N(n) &\triangleq E[U^T(n)N(n)N^T(n)U(n)] = U^T(n)R_N(n)U(n). \\ \Lambda_N(n) &\simeq \text{diag}(\lambda_N^1(n),\, \lambda_N^2(n),\, \cdots,\, \lambda_N^k(n)) \end{aligned} \tag{14}$$

where $\lambda_N^i(n)$ is the variance of noise along the ith eigenvector. This approximation is less restrictive and covers the case of white noise [i.e., for all i: $\lambda_N^i(n) = \lambda_N(n)$]. Moreover, it means that the unitary matrix $U(n)$ applied on $R_N(n)$ make the covariance matrix $U^T(n)R_N(n)U(n)$ more close to a diagonal matrix. This is a known result in matrix computation theory when $R_N(n)$ is close to a Toeplitz matrix [20]. In fact, the best choice for $U(n)$ is a unitary transformation which transforms jointly both $R_X(n)$ and $R_N(n)$ to a diagonal matrix. Substituting (14) in (13), we derive a suboptimal estimator denoted by $H(n)$, as follows:

$$H(n) = U(n)\Lambda_X(n)(\Lambda_X(n) + \mu\Lambda_N(n))^{-1}U^T(n). \tag{15}$$

Let $G(n)$ be defined as follows:

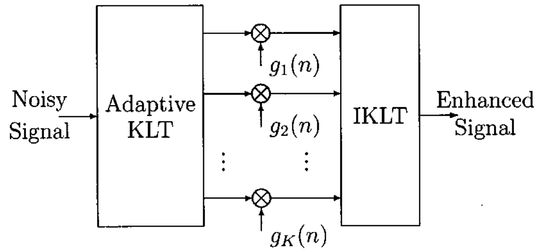
$$G(n) = \Lambda_X(n)(\Lambda_X(n) + \mu\Lambda_N(n))^{-1}. \tag{16}$$

Since $\Lambda_X(n)$ and $\Lambda_N(n)$ are diagonal matrices, we have

$$\begin{aligned} G(n) &= \text{diag}(g_1(n),\, g_2(n),\, \cdots,\, g_K(n)), \\ g_i(n) &= \frac{\lambda_X^i(n)}{\lambda_X^i(n) + \mu\lambda_N^i(n)}. \end{aligned} \tag{17}$$

The above parallel inherent property will simplify the implementation of the algorithm. Choosing the noise covariance matrix as (14) means that noise and speech signal KLT eigenvectors are approximately equal. In such a case, since noise and clean speech are assumed to be uncorrelated, the KLT eigenvectors of noisy speech will also be the same as the KLT eigenvectors of clean speech. This property will simplify the implementation of the estimator.

In Fig. 2, noise energies along the clean speech KLT eigenvectors are plotted. A noise signal has been recorded in a car (Jeep) moving with speed of 60 km/h and it contains engine, road and wind noises. It is seen from the figure that noise energies along KLT eigenvectors are not equal and also each one varies as time passes. Note that not only statistical characteristics of noise varies in time, but also clean speech eigenvectors vary as time passes.



$$g_i(n) = \frac{\lambda_X^i(n)}{\lambda_X^i(n)+\mu\lambda_N^i(n)}, \quad i = 1, 2, \cdots, K$$

Fig. 1. KLT-based speech enhancement system.

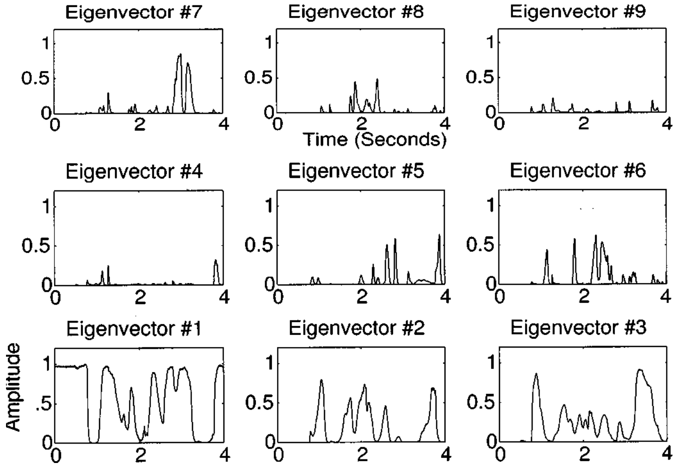


Fig. 2. Noise energies along each eigenvector (normalized on the averaged value $\{(d_i(n)/\sum_{p=1}^{K} d_p(n))\}_{i=1}^{9}$).

A block diagram of the enhancement system is shown in Fig. 1. This estimator is somehow similar to the spectral subtraction speech enhancement system used by Kang and Fransen [5]. They used DFT instead of KLT for decomposition. In the next section we will show how $\Lambda_X(n)$ and $\Lambda_N(n)$ can be obtained from noisy speech and noise intervals between speech signals. Also, we will show how KLT eigenvectors are estimated from noisy speech samples, adaptively.

III. Implementation

For implementation of $H(n)$ (15), in each time step the SVD of clean speech signal and the noise energy level over each eigenvector is required. By the fact that noise is assumed to be uncorrelated with clean signal and from (12) and (14), we have

$$\begin{aligned} R_Y(n) &= R_X(n) + R_N(n), \\ &= U(n)(\Lambda_X(n) + \Lambda_N(n))U^T(n) \end{aligned} \tag{18}$$

where $R_Y(n)$ is the covariance matrix of noisy speech. It is clear in (18) that eigenvectors of $R_Y(n)$ are the same as $R_X(n)$'s. Also, eigenvalues of $R_Y(n)$ are equal to the sum of eigenvalues of $R_X(n)$ and noise variances. Hence for finding $U(n)$ and $\Lambda_X(n)$ we may compute the eigenvalue

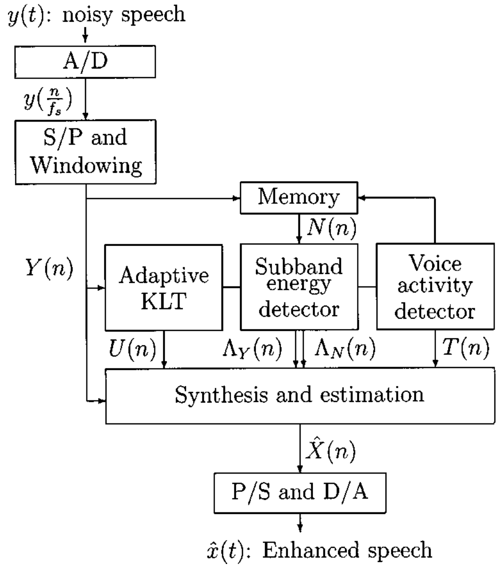


Fig. 3. Block diagram of the proposed system.

decomposition of $R_Y(n)$. In this paper $U(n)$ is obtained directly from $R_Y(n)$ and the diagonal values of $\Lambda_X(n)$ are obtained by subtracting $\Lambda_N(n)$ from $\Lambda_Y(n)$, the same as classical spectral subtraction methods. The diagonal elements of $\Lambda_N(n)$ are computed using noise samples gathered from the silence intervals between speech dialogs using a speech activity detector. Since noise statistical characteristics do not vary rapidly in time, these characteristics may be considered constant from one silence interval until the next one arrives. But clean speech KLT eigenvectors vary in time. So, noise samples are stored in a memory and are projected over the newest KLT eigenvectors. Then whenever a new string of noise samples is detected, these samples are stored in memory and are used for the next speech duration. A block diagram of this speech enhancement system is plotted in Fig. 3. In this section, we describe briefly how each block can be implemented.

A. Serial to Parallel Conversion and Windowing

In this block, using a digitized speech signal produced by an analog to digital converter, a serial to parallel converter and a windowing block, speech vectors $Y(n)$ are produced (1). We set the sampling frequency equal to 8 kHz and chose K, the length of $Y(n)$, equal to 20 as a proper value for this sampling frequency. It is obvious from (1) that we have used rectangular window over $X(n)$. In simulations we used Kaiser *et al.* and triangular moving window but we found out rectangular window performing better than the others.

The overlap between respective speech vectors in (1) is set to $K - 1$. It is obvious that a larger sampling frequency needs larger vector size and results in more correlation between speech vectors. In this case, overlap length may be reduced and this will reduce the algorithm computational complexity at the expense of some performance degradation.

B. KLT Tracking Algorithm

As it was discussed in Section II, the proposed estimator requires an accurate estimation of eigenvalues and eigenvectors of noisy speech covariance matrix. In [7], an estimation of covariance matrix is made, then it's eigenvalue decomposition is computed. And a block processing method which is very time consuming is used. Also, since speech is not a stationary process the performance of KLT tracking algorithm may be improved by using adaptive subspace tracking algorithms. Thus we propose to use projection approximation subspace tracking method introduced by Yang [15] instead. This is an adaptive method that tracks eigenvectors of covariance matrix using a recursive least square (RLS) type algorithm. The performance of this method is extensively analyzed in [16], [17]. Projection approximation is a less expensive and more accurate algorithm for subspace tracking. In [15], a constrained optimization problem is defined such that it's optimum point is located at the set of eigenvectors of covariance matrix, i.e., KLT. Define $J(u(n))$ as follows:

$$J(u(n)) = \sum_{i=1}^{n} \beta^{n-i} \|Y(i) - u(n)u^T(n)Y(i)\|^2 \tag{19}$$

where $u(n)$ is a K-dimensional vector and $0 \leq \beta \leq 1$ is a forgetting factor. It has been proved that $J(u(n))$ has no local minimum and just one global minimum. Also, it's global minimum is located at the dominant eigenvector of the empirical covariance matrix defined as follows [15]:

$$\hat{R}_Y(n) = \sum_{i=1}^{n} \beta^{n-i} Y(i)Y(i)^T. \tag{20}$$

Now, define $J'(u(n))$ as follows:

$$J'(u(n)) = \sum_{i=1}^{n} \beta^{n-i} \|Y(i) - u(n)u^T(i-1)Y(i)\|^2. \tag{21}$$

The only difference between $J'(u(n))$ and $J(u(n))$ is using $u^T(i-1)$ instead of $u^T(n)$. It is obvious that $J'(u(n))$ is a good approximation of $J(u(n))$. Since when $i \ll n$, the forgetting factor β^{n-i} will become very small and will reduce the difference between two cost functions and when i is close to n, since speech statistical characteristics varies slowly in time, $u^T(i-1) \simeq u^T(n)$. In contrast with $J(u(n))$, the new cost function can be minimized using recursive least mean square techniques. We conclude that adaptive minimization of $J'(u(n))$ approximately tracks the dominant eigenvector of noisy speech covariance matrix which is the same as the dominant eigenvector of clean speech covariance matrix. For tracking the rest of eigenvectors we use deflation technique. The basic idea of deflation is the sequential estimation of the principle components. First the most dominant eigenvector is updated by optimizing $J'(u(n))$. Then the projection of the current data vector $Y(n)$ on this eigenvector is removed. Now, the second eigenvector of $\hat{R}_Y(n)$ becomes the most dominant one and it can be extracted in the same way as before. Applying this procedure repeatedly, all eigenvectors are estimated sequentially. Using RLS and deflation techniques the KLT eigenvectors tracking algorithm will be summarized in Table I.

C. Subband Energy Estimation and Memory Unit

To implement the proposed estimator, it is necessary to compute noise and clean signal energies along each KLT eigenvector. It can be seen from Table I, that $d_i(n)$ is equal to an

TABLE I
KLT TRACKING ALGORITHM

```
Initialize   $d_i(0) = 0, \beta = 0.95$
  $U(0) = [u_1(0) \mid u_2(0) \mid \cdots \mid u_K(0)] = I_K$
$T(n)$: Voice activity detector output
For each time step do
  $Y_1(n) = Y(n)$
  For $i = 1, 2, \cdots, K$ do
    $\nu_i(n) = u_i^T(n-1)Y_i(n)$
    $d_i(n) = \beta\, d_i(n-1) + |\nu_i(n)|^2$
    $E_i(n) = Y_i(n) - u_i(n-1)\nu_i(n)$
    $u_i(n) = u_i(n-1) + T(n)E_i(n)\frac{\nu_i(n)}{d_i(n)}$
    $Y_{i+1}(n) = Y_i(n) - u_i(n)\nu_i(n)$
  end
$U(n) = [u_1(n) \mid u_2(n) \mid \cdots \mid u_K(n)]$
```

TABLE II
ADAPTIVE VOICE ACTIVITY DETECTION ALGORITHM

Initialize:	$\gamma = 0.9995$
Max–filter:	$M(n) = \gamma \max\{d_1(n), M(n-1)\} + (1-\gamma)d_1(n)$
Min–filter:	$m(n) = \gamma \min\{d_1(n), m(n-1)\} + (1-\gamma)d_1(n)$
Voice activity detector:	$T(n) = \begin{cases} 1, & \text{for } d_1(n) \geq m(n) + \frac{M(n)-m(n)}{12} \\ 0, & \text{else.} \end{cases}$

exponentially averaged energy of the noisy speech signal projected on the ith eigenvector. As noise and speech signals are supposed to be uncorrelated, noisy speech energy along each eigenvector, is equal to its clean signal energy plus noise signal energy. So, we have

$$\Lambda_Y(n) = \Lambda_X(n) + \Lambda_N(n). \tag{22}$$

If somehow noise energies along KLT eigenvectors are computed, clean speech one's will be obtained by a simple subtraction (22).

Noise process is not stationary in general. To achieve an estimation for noise level $\lambda_N^i(n)$, we assume that the statistics (i.e., autocorrelation function) of ordinary noises vary slowly with time. Of course, the noise process is not assumed stationary, but the practical results show that statistics of most of noises gathered from various environments, do not vary rapidly in time. Also, a speech dialogue consists of separate sentences that has some noise gaps located between them. In some languages, there are lots of noise gaps between separate words in a sentence. If we assume that noise statistical characteristics approximately do not change from one silence interval until the next one arrives, we can use the latest noise samples from silence intervals between speech samples to implement the proposed estimator. In the speech activity detector unit, these silence gaps are detected and noise is stored in a memory unit. So, when speech is detected to be active, we compute an exponentially averaged energy of the noise signal on the i'th eigenvector as follows:

$$\lambda_N^i(n) = \beta\lambda_N^i(n-1) + |u_i^T(n)N(n)|^2 \tag{23}$$

where $0 \leq \beta \leq 1$ is smoothing factor and $N(n)$ is chosen from the noise memory hold from just the previous silence interval. It will be seen in the next section that when the silence intervals are detected by the speech activity detector and noise samples are stored in the memory unit. These samples will be used by a subband energy detector unit for estimating noise energies along KLT eigenvectors during the next speech intervals. Noise samples are renewed every time that a new noise gap arrives. Noise interval length may be shorter than the next speech interval, at this case saved noise samples are used again and again until the next silence interval begins.

One may say that we could compute noise energies along KLT eigenvectors during noise intervals and use these energies for implementing the estimator during the next speech active interval without using a memory unit. But, note that KLT eigenvectors vary with time when new speech samples arrive. So, we use a block of saved noise samples to compute the noise energies along the newest KLT eigenvectors in each time step.

D. Adaptive Voice Activity Detector

The role of the Voice activity detector (VAD) is mostly described in the next section. Using VAD, the silence intervals between speech signals are detected and as these intervals only contain noise, noise samples are stored in memory unit and will be used for enhancement of next speech active duration. In this section we propose a simple speech activity detector like one Kang and Fransen used in [5]. The most energy part of a speech signal is located along the dominant eigenvector of the speech correlation matrix, called principle component of speech [18]. It is seen from Table I that $d_1(n)$ is somewhat representing the principle component of the speech signal in each time step. For speech activity detection we computed the past maximum and minimum history of the principle component and using a simple rule, we detected speech activity and silence intervals. As indicated in Table II, noise interval is detected when the current principle component energy ($d_1(n)$) is below a threshold set of one-twelfth of the distance between its past minimum and maximum. Since speech and noise energies may vary significantly as time passes, we use a low pass filter to let the maximum and minimum of the principle component track the real values. The performance of proposed VAD in high SNR is ideal but in small SNR when speech interval starts with a low energy formant the probability of detection decreases. In case of VAD error the low energy formant of the beginning of the incoming word is omitted. In our simulations, almost 10% of the listeners indicated the existence of the above problem in enhancement of poor speech signals (SNR ≤ 5 dB) but no one indicated that due to this phenomena he or she did not prefer the proposed enhancement system to the Ephraim system or noninhanced speech signal. In general a modified version of this speech activity detector or even a more complicated one will improve the performance of the enhancement system as indicated in Section IV.

In Table II, the value of $T(n)$ is equal to 1, if speech duration is detected and $T(n) = 0$ if silence duration is detected. In Fig. 4, the performance of the proposed VAD is investigated.

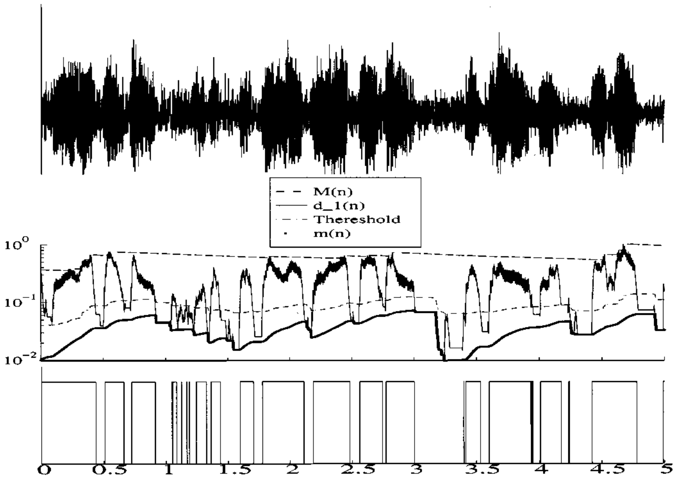


Fig. 4. Adaptive voice activity detection (a) noisy speech, (b) normalized principle component energy $d_1(n)$, Max-filter output $M(n)$, Min-filter output $m(n)$, threshold level, and (c) voice activity detectors output $T(n)$.

TABLE III
SUMMARY OF SYNTHESIS AND ESTIMATION PROCESS

```
Initialize    $\mu = 1$
For each time step do
  If speech interval detected $(T(n) = 1)$ then
    For $i = 1, 2, \cdots, K$ Do
      $\lambda_N^i(n) = \beta \lambda_N^i(n-1) + |u_i^T(n) N(n)|^2$
      $\lambda_X^i(n) = \max\{d_i(n) - \lambda_N^i(n), 0\}$
      $g_i(n) = \frac{\lambda_X^i(n)}{\lambda_X^i(n) + \mu\ \lambda_N^i(n)}$
    end
    $G(n) = \mathrm{diag}(g_1(n), g_2(n), \cdots, g_K(n))$
    $\hat{X}(n) = U(n) G(n) U^T(n) Y(n)$
    $\hat{x}(\frac{n}{f_s}) = [\hat{X}(n)]_1$
  If silence detected $(T(n) = 0)$ then
    attenuate input to -30dB.
```

Although, these kind of figures are common in the literature of speech enhancement, they do not give much information about the performance of the VAD in various SNRs, noise types, etc. But Fig. 4 somewhat represents the basis of performance of the proposed VAD. The test signal is a sentence lasting for five seconds. Noise is chosen from real samples of recorded sounds in a lab environment and SNR is approximately equal to 5 dB. The amplitude of $d_1(n)$, maximum and minimum envelopes of $d_1(n)$, threshold level and the output of the detector output are plotted. We tested this speech activity detector in various conditions such as the SNR and the noise gathered from various environments, and the results are acceptable.

E. Synthesis and Estimation Process

As noise statistics are necessary for performing $H(n)$, when silence is detected by the speech activity detector, the adaptation of KLT eigenvectors is stopped because in such a case the algorithm will track unbiased noise subspace. When the input signal only contains noise, it seems that the enhanced signal should be zero. As noted by Berouti *et al.* [8], a small amount of spectral floor in noise intervals improves speech quality. Thus when a silence interval is detected by the speech activity detector, we reduce the noise power to -30 dB. And, when speech is detected, using eigenvalues and eigenvectors estimated in Sections III-B and III-C, enhancement is performed as indicated in Table III. In this manner, vectors of K samples of noisy speech are enhanced. Since the overlap between input noisy speech vectors $Y(n)$ is selected as $K-1$, the overlap between enhanced blocks will be equal to $K-1$. To reconstruct the main speech in parallel to serial block we just picked up the first element of the enhanced speech vector. Of course, we used overlap-add method and windowing but the best results were obtained with the above reconstruction method.

To compute clean signal energies $\lambda_X^i(n)$ along KLT eigenvectors, noise energies $\lambda_N^i(n)$ are subtracted from noisy signal energies $d_i(n)$ as indicated in Table III. Because of errors in energy estimation filters (Section III-C), specially when clean signal along an eigenvector is weak in comparison with noise, some times the answer of above subtraction becomes negative. In such cases the negative number will be forced to zero as indicated in Table III. So, some part of speech information is missed and enhanced speech is degraded. This phenomenon is similar to musical noise generation of spectral subtraction speech enhancement system [5], [7]. In spectral subtraction method, speech signal is projected on DFT vectors instead of KLT eigenvectors. In this case in low SNR when negative errors like the above problem happens, some part of speech signal information is missed and also some degradation is added to enhanced speech. This will lead to presence and absence of speech information in spectral domain randomly and will cause a special noise which is named musical noise. In the proposed enhancement system, when KLT is used instead of DFT, the probability of losing most part of speech signal information decreases. Because, KLT eigenvectors are ordered according to clean speech energy along each eigenvector and in dominant eigenvectors no negative error happens. So, greatest part of speech energy which is located along dominant eigenvectors is not degraded and negative error does not cause serious problem. It is obvious that in spectral subtraction method whole information of speech is divided on DFT vectors according to speech type and somewhat randomly. So, the probability of negative error increase and musical noise is generated. In simulations we did not observe any kind of musical noises generated by the proposed enhancement system.

IV. PERFORMANCE EVALUATION

The major goal is to improve quality and intelligibility of a degraded speech signal. The quality of speech signal is a subjective measure which reflects the way the signal is perceived by listeners. It can be expressed in terms of how pleasant the signal sounds or how much effort is required to understand the message. Intelligibility, on the other hand is an objective measure of the amount of information that can be extracted by listeners from the given signal, whether the signal is clean or noisy. A given signal may be of high quality and low intelligibility and

TABLE IV
PERFORMANCE EVALUATION TESTS, PREFERENCE PERCENTAGE OF PEOPLE

SNR	Noise type	Comparison with non-processed	Comparison with Ephraim's	Remarks
10dB	white	85%	55%	software generated
5dB	white	75%	69%	white Gaussian noise
0dB	white	64%	89%	
10dB	quiet room	75%	73%	Hiss in the original
5dB	quiet room	85%	79%	analog tapes
0dB	quiet room	68%	89%	
10dB	lab	75%	73%	Chirp tones of
5dB	lab	85%	79%	computer fans
0dB	lab	68%	89%	
10dB	office	75%	73%	Human voices,
5dB	office	85%	79%	telephone ring and
0dB	office	68%	89%	type writer noise
10dB	car	75%	73%	Engine, road and wind
5dB	car	85%	79%	noises; speed: 60km/h
0dB	car	68%	89%	and engine: 3000rpm

vice versa. Since the human perceptual domain is not well understood, both the quality and the intelligibility of a set of given signals must be evaluated based on tests performed on human listeners. A careful subjective test can be tedious and time consuming, and generally requires processing a large amount of data. Because of the difficulties involved in the evaluation, only a few systems have been carefully evaluated. A few others have only been evaluated based on an objective measure such as SNR improvement even though such an objective measure does not correlate well with a subjective measure.

To evaluate the performance of the proposed estimator and compare it with enhancement system introduced by Ephraim and Van-Trees, we used a test algorithm similar to what they used in [7]. The speech material consisted of six sentences which was spoken by three male and three female speakers. The evaluation was performed by a group of 25 listeners. All of them were engineering students but none of them was dealing with speech processing area problems, as we wanted them to have no bias on what they choose. Subjects were asked to choose one of the two sentences that they listened to. Each of these two sentences was enhanced by proposed speech enhancement system or by Ephraim's system or was not enhanced. White Gaussian noise signal was produced using software and real noise samples were gathered from various places as indicated in Table IV. Adding real noise and computer generated white Gaussian noise to clean speech with various SNR's, we produced noisy speech signals. Then above tests were performed. Results are summarized in Table IV.

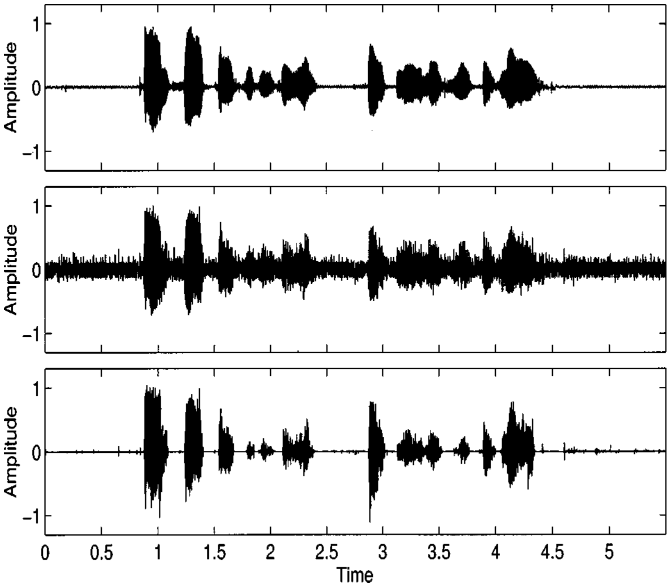


Fig. 5. (a) Clean speech, (b) noisy speech, and (c) enhanced speech.

It is seen from the table that when noise is white Gaussian and SNR is 10 dB, almost both Ephraim's and proposed enhancement system performance is the same since nearly half of the test subjects chose proposed estimator and half chose Ephraim system. It is obvious that in this case both systems should perform the same since our noise model covers white noise model. As SNR decreases, the number of persons who preferred our estimator to Ephraim's increase. This is due to the KLT tracking algorithm which we use a projection approximation type algorithm and Ephraim used a block processing type. It is interesting to note that the amount of the people who prefer nonenhanced speech to enhanced speech increase as SNR decreases. This is due to the produced distortion by the enhancement algorithm. According to (17) when noise energy along a KLT eigenvector increase, $g_i(n)$ decreases so, some part of noise is omitted and also clean speech signal is distorted. In tests we observed that some people prefer produced distortion to noise persistence and others prefer noise persistence to produced distortion. As noise energy increases and so the produced distortion increases, percent of people who prefer enhanced speech to nonenhanced decreases. When speech signal is degraded by real noise, it is seen from Table IV, that proposed estimator is preferred to Ephraim's. It is due to the fact that in Ephraim's enhancement system, noise model is supposed to be white Gaussian and so, does not contains color noises. Of course, wideband hiss noise in analog tapes, which is seen in noisy speech signals is approximately white. This fact has caused the two systems almost perform equivalently. The proposed system elevated the SNR of the noisy signals by 1.7–6.3 dB and Ephraim system by -0.5–6.1 in various noise types and SNRs. Notice that when noise energy in enhanced signal decreases, produced distortion increases also; therefore SNR improvement can not be a good measure for judging a speech enhancement system. In Fig. 5, clean, noisy and enhanced signal of a test sentence are plotted respectively. SNR is equal to 5 dB and additive noise is gathered from the lab environment. In Fig. 6, spectrograms of clean, noisy and enhanced signal of former test sentence are drawn. As there is a background noise in each of the signals and this make gray level figures ruined, only contours of spectrograms are plotted. It is seen from the figure that noise contains a chirp tone in it's spectrogram and this tone is also deleted by the system.

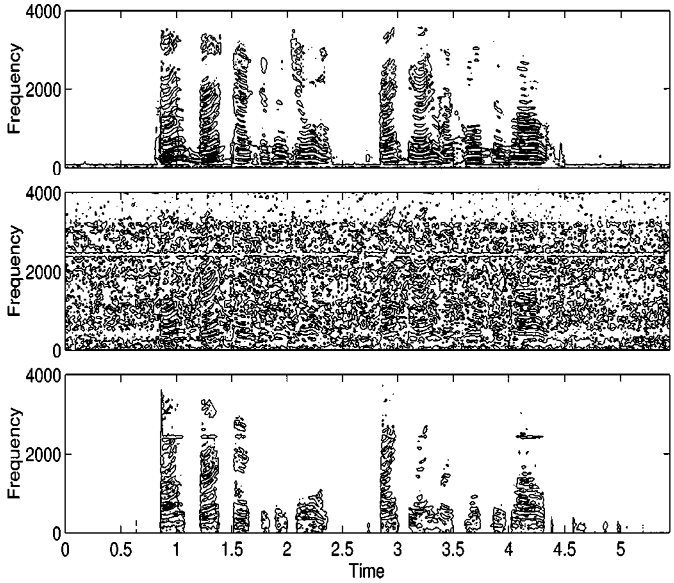


Fig. 6. (a) Clean speech, (b) noisy speech, and (c) enhanced speech.

V. Summary and Conclusion

A comprehensive framework for nonparametric speech enhancement is developed (see Tables I–III). The basic principle is to decompose noisy speech vectors to uncorrelated components. The signal decomposition is performed by a unitary transform say KLT, adaptively. Each component is modified via a linear filter. Then enhanced speech is synthesized via IKLT. In contrast to the white noise model assumption in former methods, noise is assumed to be colored. The eigenvectors of the clean speech covariance matrix are tracked using a an adaptive algorithm. It seems that using adaptive method the tracking is more efficient than block processing method used in former systems [7].

Linear estimation of the clean signal is performed using a perceptually = meaningful estimation criteria. First, two parameters are considered: 1) residual noise $\overline{\epsilon_N^2(n)}$ and 2) signal distortion $\overline{\epsilon_X^2(n)}$. Then, the signal distortion is minimized while the residual noise is maintained below some given threshold. This criterion results in a linear filter with adjustable input noise level. For each component, if noise has a great variance and clean speech is weak, the signal is attenuated. The attenuation gain is computed as a function of the variances of the noise and the clean speech signal (17). The noise variance of each component is estimated during the silence intervals. The voice activity and silence intervals are detected using a simple speech activity detector proposed in this paper. The proposed VAD could be tuned for any language that contains silence intervals (such as Persian, English, etc.) between separate words. During silence intervals, the signal samples (i.e., noise) are stored in a finite length memory and the variances of the noise for each component is computed adaptively using a simple energy detector filter at the output of each eigenfilter. It is assumed that the statistical characteristics of noise do not vary too much until the next noise interval arrives and noise samples are renewed. The proposed speech enhancement system was judged better than the Ephraim's system in our simulations.

The computational complexity of the proposed algorithm is K times less than the algorithm proposed by Ephraim and Van-Trees.

References

[1] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1524–1555, Oct. 1992.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 62, pp. 292–293, Apr. 1974.

[3] M. R. Weiss, E. Aschkenasy, and T. W. Parsons, "Processing speech signal to attenuate interference," in *Proc. IEEE Symp. Speech Recognition*, Apr. 1974, pp. 292–293.

[4] J. S. Lim, "Evaluation of correlation subtraction method for enhancing speech degraded by additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 471–472, Oct. 1978.

[5] G. S. Kang and L. J. Fransen, "Quality improvement of LPC-processed noisy speech by using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 939–943, June 1989.

[6] J. D. Ferguson, *Applications of Hidden Markov Models to Text and Speech*. Princeton, NJ: IDA-CRD, 1980.

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

[8] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 208–211, 1979.

[9] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 36, pp. 445–455, Sept. 1998.

[10] C. E. Mokbel and G. Chollet, "Automatic word recognition in cars," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 346–356, Sept. 1995.

[11] O. Cappe, C. E. Mokbel, D. Jouvet, and E. Moulines, "An algorithm for maximum likelihood estimation of hidden Markov models with unknown state-tying," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 61–70, Jan. 1998.

[12] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 373–385, July 1998.

[13] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 439–448, Nov. 1995.

[14] B. H. Juang and C. Tsuhan, "The past, present, and future of speech processing," *IEEE Signal Processing Mag.*, vol. 15, pp. 24–48, May 1998.

[15] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Processing*, vol. 43, pp. 95–107, Jan. 1995.

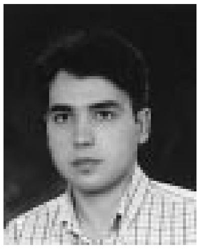
[16] V. Solo and X. Kong, "Performance analysis of adaptive eigenanalysis algorithms," *IEEE Trans. Signal Processing*, vol. 46, Mar. 1998.

[17] T. Gustafson, "Instrumental variable subspace tracking using projection approximation," *IEEE Trans. Signal Processing*, vol. 46, Mar. 1998.

[18] S. Y. Kung, *Digital Neural Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[19] D. J. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1984.

[20] G. H. Gloub and C. F. Van Loan, *Matrix Computations*, 2nd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1989.



Afshin Rezayee was born in Isfahan, Iran, in 1972. He received the B.S. and M.S. degrees in electrical engineering from Isfahan University of Technology (IUT) in 1995 and 1997, respectively. He is currently pursuing the Ph.D. degree in the Electrical and Computer Engineering Department, University of Toronto, Toronto, ON, Canada.

From 1997 to 1999 he was with the Electrical and Computer Engineering Research Center, IUT. His field of research is focused on analog integrated circuit design.

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

From 1995 to 1998, he was an Assistant Professor with the Department of Electrical and Computer Engineering, IUT. He was a Research Associate with the University of Toronto, Toronto, ON, Canada, from January 1999 to July 1999. He is now an Assistant Professor with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada. His research interests are now mainly in statistical and adaptive signal processing, speech processing (over IP), array signal processing for wireless communication systems, and analog adaptive circuits.