14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

MINIMA CONTROLLED NOISE ESTIMATION FOR KLT-BASED SPEECH ENHANCEMENT

Adam Borowicz and Alexander Petrovsky

Department of Real-Time Systems, Bialystok Technical University
Wiejska Str. 45A, 15-351 Bialystok, Poland
phone: +(48) 085 746-90-50, fax: +(48) 085 746-90-57, email: borowicz@ii.pb.bialystok.pl[1], palex@it.org.by[2]
web: www.pb.bialystok.pl

ABSTRACT

This paper addresses the problem of noise estimation for the Karhunen-Loeve transform (KLT) based speech enhancement. The eigenvalues and eigenvectors of the noise covariance matrix are tracked using recursive averaging algorithm. This process is controlled by the noise power minima obtained from the noisy signal even during the speech activity periods. The proposed approach is especially recommended for a class of signal subspace methods where a whitening transformation is required. Experiments show that the noise tracking algorithm offers similar performance as the method based on idealized voice activity detector (VAD).

1. INTRODUCTION

The noise estimator is one of the most important component of the practical speech enhancement system. An inaccurate estimation of the noise process statistics can lead to significant speech distortions and unwanted artifacts. Although, the noise reduction techniques evolve rapidly, the noise estimation problem is frequently neglected. The simplest solution is to use a VAD for a classification of signal frames. The statistics of the noise process are then gathered during the non-speech activity periods. Conventional VADs are very difficult to adjust and often fail in adverse noise environment, especially at low signal to noise ratio (SNR). The most promising solutions, intended to be used in the case of non-stationary noises, consists of minimum statistics [1] and minima controlled recursive averaging (MCRA) techniques [2]. In the first case the noise is obtained as the minima values of the smoothed periodograms. The second solution is based on recursive averaging the spectral power estimates. This process is controlled by power minima obtained in a similar way as in the minimum statistics method. These approaches are relatively simple and efficient in the adverse noise environment. Unfortunately they have been adopted only to the frequency domain speech enhancement scheme.

In recent years we observe a growing interest in the KLT-based approach for speech enhancement [3]. These methods are closely related to conventional spectral weighting techniques. The main difference is that the noisy signal is processed in the KLT domain. In addition, covariance matrices are estimated instead of power spectral densities (PSDs). In this paper we restrict our considerations to the non-trivial case of colored noise. The most advanced KLT-based speech enhancement schemes [4], [5], [6] exploit a joint diagonalization of the noise and clean speech covariance matrices. These approaches assume prewhitening that is realized using square root of the inverse noise covariance matrix. Other solutions [7], [8] diagonalize only the clean signal covariance matrix. However the noise covariance matrix is required for estimation of that matrix. All these methods have a great theoretical importance, but its practical realization is still a challenge. The existing schemes use conventional VADs and from this reason their noise tracking capabilities are insufficient.

In fact we do neither need direct estimation of the noise covariance matrix nor the matrix inversion, since this task can be viewed as the noise subspace filtering. If the eigenvectors and eigenvalues of the noise signal are known we can easily perform whitening transformation. Therefore we propose an estimation of the noise KLT basis using a modified version of a projection approximation subspace tracking (PASTd) algorithm [9]. We will show that the noise subspace tracking can be controlled by the noise power minima computed in transformed domain.

2. PROBLEM FORMULATION

Let $\mathbf{y}(t)$ and $\mathbf{n}(t)$ be k-dimensional zero-mean independent random vectors representing clean speech and noise respectively at time instant t. The corresponding noisy speech vector is $\mathbf{x}(t) = \mathbf{y}(t) + \mathbf{n}(t)$ with the covariance matrix defined as

$$\mathbf{R}_{\mathbf{x}}(t) = E\left\{\mathbf{x}(t)\mathbf{x}^T(t)\right\} = \mathbf{R}_{\mathbf{y}}(t) + \mathbf{R}_{\mathbf{n}}(t), \qquad (1)$$

where $\mathbf{R}_{\mathbf{y}}(t)$, $\mathbf{R}_{\mathbf{n}}(t)$ are the covariance matrices of the noise and clean speech process, respectively. It is also assumed that the matrix $\mathbf{R}_{\mathbf{n}}(t)$ is positive definite. Consider the eigenvalue decomposition (ED) of the matrix $\mathbf{R}_{\mathbf{x}}(t)$ to be

$$\mathbf{R}_{\mathbf{x}}(t) = \mathbf{U}_{\mathbf{x}}(t)\Lambda_{\mathbf{x}}(t)\mathbf{U}_{\mathbf{x}}^T(t), \qquad (2)$$

where $\mathbf{U}_{\mathbf{x}}(t) = [\mathbf{u}_{\mathbf{x},1}(t), ..., \mathbf{u}_{\mathbf{x},k}(t)]^T$ is eigenvector matrix and $\Lambda_{\mathbf{x}}(t)$ is a diagonal matrix containing the eigenvalues $\{\lambda_{\mathbf{x},i}(t), i = 1, ..., k\}$. Assuming the same notation as above, the covariance matrices of the clean speech and noise can be written as follows

$$\mathbf{R}_{\mathbf{y}}(t) = \mathbf{U}_{\mathbf{y}}(t)\Lambda_{\mathbf{y}}(t)\mathbf{U}_{\mathbf{y}}^T(t), \qquad (3)$$

$$\mathbf{R}_{\mathbf{n}}(t) = \mathbf{U}_{\mathbf{n}}(t)\Lambda_{\mathbf{n}}(t)\mathbf{U}_{\mathbf{n}}^T(t). \qquad (4)$$

Accurate estimates of the eigenvectors and eigenvalues of the matrix $\mathbf{R}_{\mathbf{x}}(t)$ are required in the KLT-based speech processing algorithms. In the basic approaches the noise is assumed

Work supported by Bialystok Technical University under the grant W/WI/2/05.

to be white i.e. $\mathbf{R}_{\mathbf{n}}(t) = \sigma_{\mathbf{n}}^2(t)\mathbf{I}$, where $\sigma_{\mathbf{n}}^2(t)$ is the variance of noise. In this trivial case, we have $\mathbf{U}_{\mathbf{x}}(t) = \mathbf{U}_{\mathbf{y}}(t)$, thus

$$\mathbf{R}_{\mathbf{x}}(t) = \mathbf{U}_{\mathbf{x}}(t)\left(\Lambda_{\mathbf{y}}(t) + \sigma_{\mathbf{n}}^2(t)\mathbf{I}\right)\mathbf{U}_{\mathbf{x}}^T(t). \quad (5)$$

The eigenstructure of the matrix $\mathbf{R}_{\mathbf{x}}(t)$ can be easily obtained from the noisy data and only the noise variance should be estimated. However, in the case of colored noise the matrix $\mathbf{R}_{\mathbf{n}}(t)$ is no longer diagonal and $\mathbf{U}_{\mathbf{x}}(t) \neq \mathbf{U}_{\mathbf{y}}(t)$. Since the *i*-th eigenvector of clean speech $\mathbf{u}_{\mathbf{y},i}(t)$ and noise $\mathbf{u}_{\mathbf{n},i}(t)$ are not parallel, the corresponding eigenvalues are not additive, i.e.

$$\lambda_{\mathbf{x},i}(t) \neq \lambda_{\mathbf{y},i}(t) + \lambda_{\mathbf{n},i}(t). \quad (6)$$

Thus in this case a subspace separation is not possible. However, in the most advanced KLT-based applications the noisy speech vectors $\mathbf{x}(t)$ are whitened using a square root of the inverse noise covariance matrix. Let's denote whitened noisy vector by

$$\hat{\mathbf{x}}(t) = \mathbf{R}_{\mathbf{n}}^{-0.5}(t)\mathbf{x}(t) \quad (7)$$

and the corresponding covariance matrix

$$\mathbf{R}_{\hat{\mathbf{x}}}(t) = E\left\{\hat{\mathbf{x}}(t)^T\hat{\mathbf{x}}(t)\right\} = \mathbf{R}_{\mathbf{n}}^{-0.5}(t)\mathbf{R}_{\mathbf{x}}(t)\mathbf{R}_{\mathbf{n}}^{-0.5}(t). \quad (8)$$

Since the processing is performed in the signal-subspace of the whitened noisy signal, the ED of the matrix $\mathbf{R}_{\hat{\mathbf{x}}}(t)$ is considered instead of (5), i.e.

$$\mathbf{R}_{\hat{\mathbf{x}}}(t) = \mathbf{U}_{\hat{\mathbf{x}}}(t)\left(\Lambda_{\hat{\mathbf{x}}}(t) + \mathbf{I}\right)\mathbf{U}_{\hat{\mathbf{x}}}^T(t). \quad (9)$$

Now its clear that inexact whitening may deteriorate a performance of whole system. Not only signal distortions may be introduced but also the KLT matrices may be estimated incorrectly, which results in suboptimal signal decorrelation efficiency. Typically, the noise covariance matrix is estimated in the speech absent frames using a voice activity detector. This solution possess one important disadvantage. Namely, there is no possibility to track the noise statistics during a voice activity. On the other hand, a computation of the whitening filter for each processed frame is computationally intensive and unpractical. In fact a whitening step can be interpreted as the noise subspace filtering

$$\mathbf{R}_{\mathbf{n}}^{-0.5}(t)\mathbf{x}(t) = \mathbf{U}_{\mathbf{n}}(t)\mathbf{\Lambda}_{\mathbf{n}}^{-0.5}(t)\mathbf{U}_{\mathbf{n}}^T(t)\mathbf{x}(t). \quad (10)$$

Thus we do neither require direct estimation of the noise covariance matrix nor finding the square root of its inverse. Our goal is to estimate only the eigenvectors and eigenvalues of the noise signal. As we will show, it can be done adaptively using the subspace tracking algorithm.

3. NOISE SUBSPACE TRACKING

We can split the problem of noise estimation into two tasks, the noise eigenvalues tracking and eigenvectors tracking. It can be easily observed that the KLT bases of noisy speech and noise coincide at low SNRs, i.e. $\mathbf{U}_{\mathbf{x}}(t) \approx \mathbf{U}_{\mathbf{n}}(t)$. Since the noise is usually more stationary than the speech signal it is reasonable to assume that the noise subspace does not vary rapidly during the speech activity. On the other hand if the energy fluctuations in direction of the particular eigenvectors are not tracked quickly enough, musical tones may be generated. It was verified empirically that the problem of noise KLT basis tracking is not so crucial for the speech enhancement as the estimation of the noise energies. Therefore we conclude that the noise eigenvectors can be estimated in the non-speech activity periods using the VAD, but the corresponding eigenvalues should be tracked all the time, even during the speech activity. A block diagram of the proposed method is depicted in Fig. 1. In the following sections we describe the details of the noise tracking scheme.

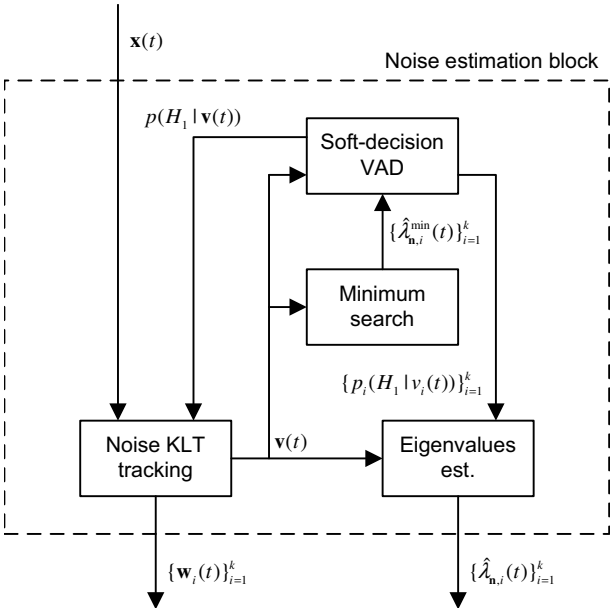

Figure 1: Block diagram of the noise subspace tracking method

3.1 Adaptive KLT tracking

Suppose a set of orthogonal vectors $\{\mathbf{w}_i(t),\ i = 1,...,k\}$ as an approximation of the noise KLT basis. Let's denote the noisy speech vector in the transformed domain as

$$\mathbf{v}(t) = [v_1(t), v_2(t), \ldots, v_k(t)]^T, \quad (11)$$

where the scalars

$$v_i(t) = \mathbf{w}_i^T(t)[\mathbf{y}(t) + \mathbf{n}(t)], \quad i = 1,...,k, \quad (12)$$

are interpreted as the KLT coefficients. An average power of the noisy speech in the direction of the vector $\mathbf{w}_i(t)$ can be defined as follows

$$\lambda_{\mathbf{x},i}^{\mathbf{w}}(t) = E\left\{|v_i(t)|^2\right\} = \lambda_{\mathbf{y},i}^{\mathbf{w}}(t) + \lambda_{\mathbf{n},i}^{\mathbf{w}}(t), \quad (13)$$

where the components

$$\lambda_{\mathbf{y},i}^{\mathbf{w}}(t) = \mathbf{w}_i^T(t)\mathbf{R}_{\mathbf{y}}(t)\mathbf{w}_i(t), \quad (14)$$

$$\lambda_{\mathbf{n},i}^{\mathbf{w}}(t) = \mathbf{w}_i^T(t)\mathbf{R}_{\mathbf{n}}(t)\mathbf{w}_i(t), \quad (15)$$

denote an average power of the clean speech and noise respectively. It can be easily observed that as the vector $\mathbf{w}_i(t)$ coincides with the *i*-th noise eigenvector, the corresponding eigenvalue is approximated by the component $\lambda_{\mathbf{n},i}^{\mathbf{w}}(t)$, i.e.

$$\mathbf{w}_i(t) \approx \mathbf{u}_{\mathbf{n},i}(t) \Rightarrow \lambda_{\mathbf{n},i}^{\mathbf{w}}(t) \approx \lambda_{\mathbf{n},i}(t). \quad (16)$$

The vectors $\{\mathbf{w}_i(t),\ i=1,...,k\}$ can be calculated using the modified PASTd algorithm (see Table 1). Our proposition is to use a time-varying smoothing parameter $\tilde{\beta}(t)$ instead of fixed one $\beta \in (0;1)$. This parameter can be driven by the output of the soft-decision VAD in a similar way as in [2]. Let's denote the conditional speech presence probability at time instant t as $p(H_1|\mathbf{v}(t))$, then

$$\tilde{\beta}(t) = \beta + (1-\beta)p(H_1|\mathbf{v}(t-1)). \tag{17}$$

Note that if $p(H_1|\mathbf{v}(t-1)) = 1$, a subspace tracking is stopped, but if $p(H_1|\mathbf{v}(t-1)) = 0$, fast convergence is provided. The parameters $d_i(t)$ are the exponentially weighted estimates of the noise eigenvalues, however we can not use them, since fast adaptation of the eigenvalues during the speech activity is considered. Therefore, we propose to track the noise eigenvalues independently in each KLT subband. Let's denote the conditional speech presence probability in the i-th subband at time instant t as $p_i(H_1|v_i(t))$. Thus the noise eigenvalues can be estimated as follows

$$\hat{\lambda}_{\mathbf{n},i}(t) = \beta_i(t)\hat{\lambda}_{\mathbf{n},i}(t-1) + (1-\beta_i(t))\,|v_i(t)|^2. \tag{18}$$

Similarly, the time-varying parameters $\{\beta_i(t),\ i=1,...,k\}$ are adjusted by the corresponding subband speech presence probabilities

$$\beta_i(t) = \beta + (1-\beta)p_i(H_1|v_i(t-1)). \tag{19}$$

In the above, we use the same smoothing parameter β as for the PASTd algorithm. Therefore, in the non-speech activity periods, the estimates $\hat{\lambda}_{\mathbf{n},i}(t)$ and $d_i(t)$ do not differ from each other in substance.

3.2 Noise power minima tracking

A key point of the proposed approach is a robust implementation of the soft-decision VAD. As we will show, the speech presence probabilities can be estimated using the noise power minima. Suppose, an exponentially weighted estimate of noisy speech power along the vector $\mathbf{w}_i(t)$

$$\hat{\lambda}^{\mathbf{w}}_{\mathbf{x},i}(t) = \alpha\hat{\lambda}^{\mathbf{w}}_{\mathbf{x},i}(t-1) + (1-\alpha)\,|v_i(t)|^2, \tag{20}$$

where $\alpha \in (0;\beta)$ is a smoothing parameter. It follows easily from (13) that if the speech is absent, then $\lambda^{\mathbf{w}}_{\mathbf{x},i}(t) = \hat{\lambda}_{\mathbf{n},i}(t)$. Thus, the noise power in the i-th subband can be roughly estimated as follows

$$\lambda^{\min}_{\mathbf{n},i}(t) \stackrel{\Delta}{=} \min\left\{\hat{\lambda}^{\mathbf{w}}_{\mathbf{x},i}(t-j), \quad j=0,1,\ldots,T\right\}, \tag{21}$$

where T is a size of a minimum search window. It should be noted that using a search window is memory consuming and one can use a simplified method [2] to track the noise minima. The estimator $\lambda^{\min}_{\mathbf{n},i}(t)$ is biased, thus we recommend to perform a bias compensation

$$\hat{\lambda}^{\min}_{\mathbf{n},i}(t) \stackrel{\Delta}{=} B_{\min}\lambda^{\min}_{\mathbf{n},i}(t). \tag{22}$$

The factor $B_{\min}$ generally depends on the values of T and α. In our experiments it was simply set to fixed value. Due to energy fluctuations of the noisy speech, the estimator $\hat{\lambda}^{\min}_{\mathbf{n},i}(t)$ has relatively large variance. However it is sufficiently precise for computation of the speech presence probabilities.

Table 1: PASTd algorithm with a time-varying smoothing parameter.

$\mathbf{x}_1(t) = \mathbf{x}(t)$

FOR $\quad i = 1,2,...,k \quad$ DO

$v_i(t) = \mathbf{w}_i^T(t-1)\mathbf{x}_i(t)$

$d_i(t) = \tilde{\beta}(t)d_i(t-1) + \left(1-\tilde{\beta}(t)\right)|v_i(t)|^2$

$\mathbf{E}_i(t) = \mathbf{x}_i(t) - \mathbf{w}_i(t-1)v_i(t)$

$\mathbf{w}_i(t) = \mathbf{w}_i(t-1) + \left(1-\tilde{\beta}(t)\right)\mathbf{E}_i(t)v_i(t)/d_i(t)$

$\mathbf{x}_{i+1}(t) = \mathbf{x}_i(t) - \mathbf{w}_i(t)v_i(t)$

END

3.3 Speech presence probabilities

The global speech presence probability $p(H_1|\mathbf{v}(t))$ as well as the subband probabilities $\{p_i(H_1|v_i(t)),\ i=1,...,k\}$ are computed at once using the same Gaussian-Laplacian mixture model [10]. In each subband, we have to evaluate two statistical hypotheses, $H_0(i,t)$ and $H_1(i,t)$, which indicate, respectively, speech absence and presence at the time instant t, i.e.

$$\begin{aligned} H_0(i,t) &: v_i(t) = \mathbf{w}_i^T\mathbf{n}(t), \\ H_1(i,t) &: v_i(t) = \mathbf{w}_i^T\left[\mathbf{y}(t)+\mathbf{n}(t)\right]. \end{aligned} \tag{23}$$

We assume that the speech components $\mathbf{w}_i^T\mathbf{y}(t)$ in the KLT domain have zero-mean Laplacian distribution and the noise components $\mathbf{w}_i^T\mathbf{n}(t)$ are Gaussian. The probability density functions (PDFs) of $v_i(t)$ can be derived as follows

$$\begin{aligned} f_i(v_i(t)|H_0) &= \frac{1}{\sqrt{2\pi\lambda^{\mathbf{w}}_{\mathbf{n},i}(t)}}\exp\left(-\frac{v_i^2(t)}{2\lambda^{\mathbf{w}}_{\mathbf{n},i}(t)}\right), \\ f_i(v_i(t)|H_1) &= \frac{1}{4a_i(t)}\exp\left(\frac{\lambda^{\mathbf{w}}_{\mathbf{n},i}(t)}{2a_i^2(t)}\right)\left[A^{(+)}+A^{(-)}\right]. \end{aligned} \tag{24}$$

For convenience, the following substitution is used

$$A^{(\pm)} = \exp\left(\pm\frac{v_i(t)}{a_i(t)}\right)\operatorname{erfc}\left(\frac{\lambda^{\mathbf{w}}_{\mathbf{n},i}(t) \pm a_i(t)v_i(t)}{a_i(t)\sqrt{2\lambda^{\mathbf{w}}_{\mathbf{n},i}(t)}}\right). \tag{25}$$

In the above the parameter $a_i(t)$ is a Laplacian factor [10] and $\operatorname{erfc}(.)$ denotes error function. It is known that the variance of a Laplace distributed zero-mean random variable is equal to

$$2a_i^2(t) = E\left\{\left|\mathbf{w}_i^T(t)\mathbf{y}(t)\right|^2\right\} = \lambda^{\mathbf{w}}_{\mathbf{x},i}(t) - \lambda^{\mathbf{w}}_{\mathbf{n},i}(t). \tag{26}$$

We propose the following minima-controlled estimators

$$\lambda^{\mathbf{w}}_{\mathbf{n},i}(t) \approx \hat{\lambda}^{\min}_{\mathbf{n},i}(t), \tag{27}$$

$$a_i(t) \approx \sqrt{0.5\left(\hat{\lambda}^{\mathbf{w}}_{\mathbf{x},i}(t) - \hat{\lambda}^{\min}_{\mathbf{n},i}(t)\right)}. \tag{28}$$

Thus the likelihood ratios, given by

$$L_i(t) = \frac{f_i(v_i(t)|H_1)}{f_i(v_i(t)|H_0)}, \quad i = 1,...,k, \tag{29}$$

can be easily computed using the estimates (27) and (28). The simplest way to improve speech probability estimation is to take into account the strong correlation of the samples in the consecutive frames. Therefore we employ temporally smoothed log-likelihood ratio [11]

$$\hat{L}_i(t) = \alpha_L \hat{L}_i(t-1) + (1-\alpha_L)\log L_i(t), \tag{30}$$

where $\alpha_L \in (0;1)$ is a smoothing parameter. The subband speech presence probabilities are calculated using Bayes rule

$$p_i(H_1|v_i(t)) = \frac{\exp\left(\hat{L}_i(t)\right)}{\exp\left(\hat{L}_i(t)\right) + [1-p_i(H_1)]/p_i(H_1)}, \tag{31}$$

where $p_i(H_1)$ is *a priori* speech presence probability in the i-th subband. It can be simply set to fixed value, however better solution is to use binary Markov model for prediction as was suggested in [10]

$$p_i(H_1) = \Pi_{01} + (\Pi_{01} + \Pi_{11})\, p_i(H_1|v_i(t-1)), \tag{32}$$

where Π_{ij} denotes the probability of the transition from the state H_i to H_j. In our experiments we set $\Pi_{01} = 0.05$ and $\Pi_{11} = 0.9$. In order to compute global speech presence probability $p(H_1|\mathbf{v}(t))$, we require multivariate probability density functions $f(\mathbf{v}(t)|H_1)$ and $f(\mathbf{v}(t)|H_0)$, however if we assume that the noisy speech components $\{v_i(t),\ i = 1,...,k\}$ are uncorrelated, the likelihood ratio can be approximated by

$$L(t) = \frac{f(\mathbf{v}(t)|H_1)}{f(\mathbf{v}(t)|H_0)} \approx \prod_{i=1}^{k} \frac{f_i(v_i(t)|H_1)}{f_i(v_i(t)|H_0)} = \prod_{i=1}^{k} L_i(t). \tag{33}$$

The corresponding temporally smoothed log-likelihood ratio is given by

$$\hat{L}(t) = \alpha_L \hat{L}(t-1) + (1-\alpha_L)\log(L(t)) = \sum_{i=1}^{k} \hat{L}_i(t). \tag{34}$$

Since the computation of the global speech presence probability $p(H_1|\mathbf{v}(t))$ is identical to (31), we omit an explicit summary of this derivation.

We found that the proposed estimator of global speech presence probability is not sensitive to the KLT-basis fluctuations. As can be seen in Fig. 2, even if the noise KLT is replaced by an arbitrary (randomly generated) orthogonal basis, the estimate of the global speech probability is precise enough to discriminate the speech activity and silence regions. This implies that the resulting eigenvectors converge to the true noise KLT basis. Therefore the subband speech presence probabilities as well as the corresponding eigenvalues are also correctly estimated.

4. EXPERIMENTS

The performance of the presented method was evaluated and compared to idealized VAD-based approach. To differentiate these methods we called our algorithm noise subspace tracking (NST). The both noise estimation algorithms were combined with the KLT-based speech enhancement system. For

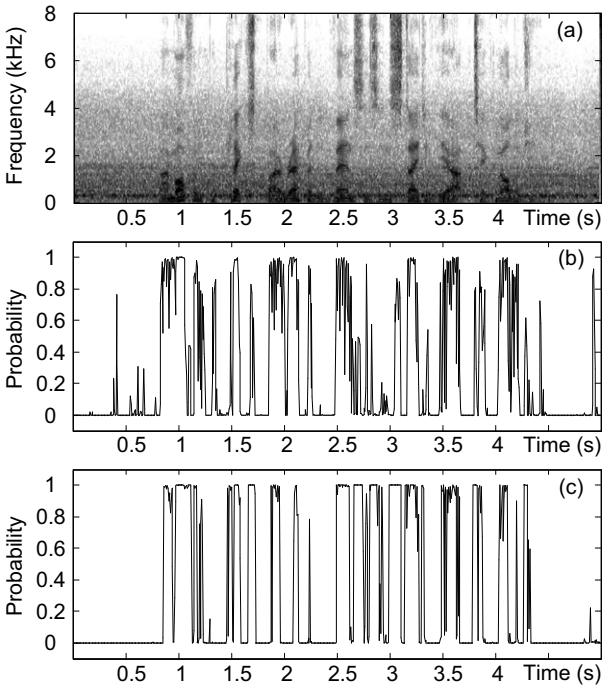

Figure 2: Global speech presence probability. (a) Spectrogram of the noisy speech (SNR at 0 dB). (b) Output of soft-decision VAD for arbitrary orthogonal basis. (c) Output of soft-decision VAD for the noise KLT basis.

our experiments we selected time domain constrained (TDC) approach [4]. For the VAD-based version of this method the matrix $\mathbf{R_n}(t)$ was simply estimated using sample correlation matrix during non-speech activity periods. In order to simulate idealized VAD, speech activity regions were marked manually. Both methods have been implemented assuming a frequency sampling at 16 kHz and signal space dimension $k = 32$. The following smoothing parameters have been chosen $\beta = 0.99$, $\alpha = 0.995$ and $\alpha_L = 0.95$. The length of the minima search window was about 0.5 s. In our implementation of the TDC estimator, we use a empirically fixed Lagrange multiplier $\mu = 5$ for simplicity, however optimal procedure [6] is also known.

For our experiments a set of 8 sentences was used. As a degradation signal we selected slowly varying vehicular noise. It was added to the clean speech signals such that the segmental SNR was between 0 dB and 15 dB. The amount of noise reduction was measured using noise attenuation factor defined as the mean ratio between the input noise power and output noise power. Speech distortions were measured using segmental SNR, where the noise was interpreted as a difference between original and enhanced speech. The higher the value of this factor, the weaker the speech distortions.

The spectrograms and informal listening tests show that both methods provide similar speech distortions and noise attenuation. This observation has been confirmed with objective measurement (see Table 2). The performance of most practical VADs strongly depends on SNR and it is far from the theoretical limit that we have here. From this reason idealized VAD-based solution gives slightly better results, especially at low SNRs. On the other hand, VAD-based approach suffers from unnatural sharpness at transitions be-

Table 2: Objective measurement.

Input SegSNR	Speech distortions		Noise attenuation	
(dB)	VAD	NST	VAD	NST
0	4.65	3.45	18.46	18.16
5	7.02	6.05	13.67	13.55
10	9.68	9.27	9.17	9.08
15	12.30	13.16	5.39	5.28

tween speech and silence regions which is uncomfortable for a listener. If these transitions are incorrectly detected the low-power speech components that are usually present in these regions can be lost. The proposed method provides the smooth transitions which improves speech intelligibility and overall listener comfort.

Fig. 3 shows the spectrograms of an example sentence used in the tests. In order to simulate the substantial change in the noise statistics, a pre-filtered Gaussian noise was added to the original noise signal with the delay of about 2 s. It can be seen that the adaptation period of NST algorithm is proportional to the size of assumed minimum search window. It can be also seen that the proposed method performs as well as the idealized VAD-based approach or even better.

5. CONCLUSIONS

A novel noise estimation algorithm for KLT-based speech enhancement has been presented. Instead of estimating noise covariance matrix, a recursive averaging technique is used to track the noise eigenvectors as well as the corresponding eigenvalues. The major adventage of the proposed method is that the tracking is also performed during the speech activity periods. This process is controlled by the noise power minima. The resulting estimates are especially useful if application considers whitening transformation. The proposed noise estimation method has been evaluated and compared with idealized VAD-based solution in the KLT-based speech enhancement system. Experiments and informal ilstening tests show that the both methods offer similar performance.

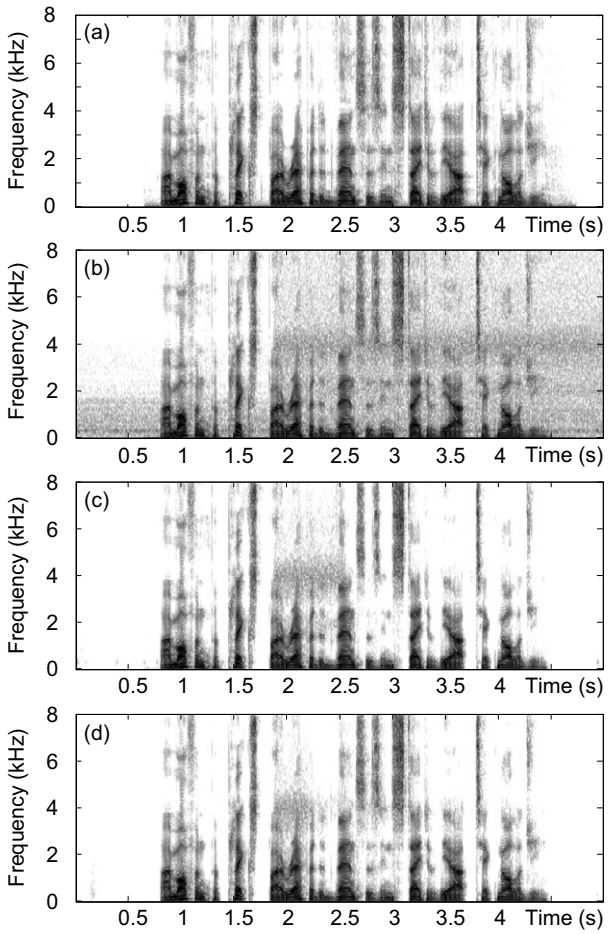

Figure 3: Speech spectrograms. (a) Original speech signal. (b) Noisy speech signal (vehicular noise with increase at 2 s). (c) Speech enhanced with the NST-based method. (d) Speech enhanced with the method based on the idealized VAD.

REFERENCES

[1] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[2] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Lett.*, vol. 9, no. 1, pp. 12–15, 2002.

[3] 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, 1995.

[4] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace enhancement to colored noise," *IEEE Signal Processing Lett.*, vol. 10, no. 4, pp. 104–106, 2003.

[5] Y. Hu and P.C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," in *Proc. ICASSP*, May 2003, vol. 1, pp. 573–576.

[6] A. Borowicz and A. Petrovsky, "Perceptually constrained subspace method for enhancing speech degraded by colored noise," in *Proc. AES 118th*, May 2005, 12 p.

[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, 2001.

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

[9] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Processing*, vol. 43, no. 1, pp. 95–107, 1995.

[10] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 498–504, 2003.

[11] Y.D. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Proc. ICASSP*, May 2001, vol. 2, pp. 737–740.