Fast Implementation of KLT-Based Speech Enhancement Using Vector Quantization

Yoshifumi Nagata, Kenji Mitsubori, Takahiko Kagi, Toyota Fujioka, and Masato Abe, *Member, IEEE*

***Abstract*—We propose a new method for implementing Karhunen–Loeve transform (KLT)-based speech enhancement to exploit vector quantization (VQ). The method is suitable for real-time processing. The proposed method consists of a VQ learning stage and a filtering stage. In the VQ learning stage, the autocorrelation vectors comprising the first K elements of the autocorrelation function are extracted from learning data. The autocorrelation vectors are used as codewords in the VQ codebook. Next, the KLT bases that correspond to all the codeword vectors are estimated through eigendecomposition (ED) of the empirical Toeplitz covariance matrices constructed from the codeword vectors. In the filtering stage, the autocorrelation vectors that are estimated from the input signal are compared to the codewords. The nearest one is chosen in each frame. The precomputed KLT bases corresponding to the chosen codeword are used for filtering instead of performing ED, which is computationally intensive. Speech quality evaluation using objective measures shows that the proposed method is comparable to a conventional KLT-based method that performs ED in the filtering process. Results of subjective tests also support this result. In addition, processing time is reduced to about 1/66 that of the conventional method in the case where a frame length of 120 points is used. This complexity reduction is attained after the computational cost in the learning stage and a corresponding increase in the associated memory requirement. Nevertheless, these results demonstrate that the proposed method reduces computational complexity while maintaining the speech quality of the KLT-based speech enhancement.**

***Index Terms*—Complexity, Karhunen–Loeve transform (KLT), speech enhancement, subspace, vector quantization.**

I. Introduction

SPECTRAL SUBTRACTION (SS) based on discrete Fourier transform (DFT) [1] is a widely used technique for speech enhancement. Nevertheless, SS suffers from self-generated residual noise, called musical noise, when the signal-to-noise ratio (SNR) is low.

On the other hand, Ephraim *et al.* proposed Karhunen–Loeve transform (KLT)-based speech enhancement [2]. This method is promising because it generates little musical noise. Nevertheless, it has two important drawbacks. One is the large computational complexity involved in KLT base estimation, which is usually achieved by eigendecomposition or singular value decomposition. This disadvantage renders this method unsuitable for real-time processing in most cases. The second problem is performance in colored-noise conditions. Noise whitening proposed in this method requires the inverse of the autocorrelation matrix, which can be numerically unstable in narrowband noise conditions.

Regarding the colored-noise problem, Ephraim's method is extended and an explicit solution has been given [3]. This method utilizes a noise whitening approach, as used in its original method. The present paper includes the assertion that instability in computing the inverse matrix, which is required for whitening, can be avoided through modification of the autocorrelation matrix. However, the computational cost for obtaining the inverse matrix and the filtering to accomplish both whitening and its inverse cannot be disregarded. In contrast, Mittal and Phamdo [4] and Rezayee and Gazor [5] proposed methods that do not require noise-whitening. In those methods, the spectral component power along each KLT axis for the current frame is estimated from a noise signal that has been preserved from a past noise period. This processing corresponds to noise spectrum estimation that is commonly performed in the DFT-based SS, whereas the noise spectrum should be calculated in every speech frame because KLT-based SS uses time-varying bases for signal decomposition.

Regarding the problem of computational complexity, Rezayee *et al.* [5] proposed an adaptive estimation of KLT bases. However, the KLT bases estimation in this method requires at least K^2 multiplication iterations in each frame, where K is the frame sample length. This requirement implies that K^3 multiplications, which are proportional to the ED complexity, are necessary to produce a K-point output signal because a frame shift of 1 point is recommended in this method. A larger frame shift might provide larger complexity reduction in this method. However, speech quality can degrade as the shift increases. This tradeoff is not addressed in this paper.

By contrast to the above methods with accurate KLT bases estimation, utilization of approximated KLT bases has been proposed for reducing the computational cost [6]. This method utilizes discrete cosine transform (DCT) and reduces the complexity of the KLT bases estimation to $O(N^2)$. This approximation is valid only for the AR(p) process. Therefore, an improved method exploiting wavelet transform has also been proposed [7]. This method requires a wavelet packet search of only $O(N \log N)$ for KLT bases estimation. However, the total ef-

Manuscript received February 17, 2004; revised September 20, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kuldip K. Paliwal.

Y. Nagata, T. Fujioka, and M. Abe are with the Department of Computer and Information Sciences, Iwate University, Morioka 020-8550, Japan (e-mail: nagata@cis.iwate-u.ac.jp).

K. Mitsubori was with the Department of Computer and Information Sciences, Iwate University, Morioka 020-8550, Japan. He is now with Sendai Technology Center, Chuo Electronics Company, Ltd., Sendai 983-0852, Japan.

T. Kagi was with the Department of Computer and Information Sciences, Iwate University, Morioka 020-8550, Japan. He is now with NTT East (Nippon Telegraph and Telephone East Corporation) Akita Group, Akita 010-0816, Japan.

Digital Object Identifier 10.1109/TASL.2006.872622

ficiency of reducing computational cost and speech quality remains unknown.

Next we propose an alternative method for approximating KLT bases by applying vector quantization (VQ). From the fact that ordinary KLT bases are obtained by performing eigendecomposition (ED) of the Toeplitz covariance matrix constructed from the autocorrelation vector, which is the vector comprising of the first K elements of the autocorrelation function, we chose the autocorrelation vector as a codeword vector of the VQ codebook. The VQ codebook is designed from a set of autocorrelation vectors that are computed from learning data by performing clustering. Following the VQ codebook acquisition, the KLT bases corresponding to all clusters are estimated through ED. The above processes are done in the learning stage preliminary to filtering for speech enhancement.

In the filtering stage for speech enhancement, the autocorrelation vector estimated from the input signal is compared to codeword vectors in the codebook. Then the nearest one is chosen in each frame. The above precomputed KLT bases corresponding to the chosen cluster are used as the approximated KLT bases for filtering instead of performing ED in each frame.

Aside from KLT bases estimation, a large calculation cost is required for estimation of the spectral component power along each KLT axis. That cost is necessary to deal with colored-noise in the method using Toeplitz covariance matrix. Regarding this problem, we apply the method of fast power estimation introduced in [6], which is allowed by precomputing the part of the power estimation using known bases of the VQ clusters. The total calculation cost is reduced drastically by avoiding ED and the power estimation required by conventional methods.

A speech-signal power spectrum is well known to be important information for a speech recognition system (e.g., [8]). The autocorrelation function and the power spectrum have one to one correspondence. Therefore, careful acquisition of the VQ cluster using an autocorrelation vector can also cover speech-signal variation, as in a speech recognition system.

This paper is organized as follows: Section II provides an overview of KLT-based speech enhancement. We show that an assumption to deal with colored noise proposed by Rezayee *et al.* is applicable to our proposed method. Section III describes the proposed method based on VQ. Section IV describes evaluation results of the proposed method in terms of speech quality and processing speed. Speech quality evaluation contains objective measure evaluation and subjective listening tests. Section V describes investigation and discussion of the error induced by the VQ method. Section VI concludes this paper.

II. KLT-Based Speech Enhancement for Colored Noise Conditions

Two varieties of solution to the estimation problem posed by Ephraim and Van Trees exist: one is based on a time-domain constraint and one is based on a spectral-domain constraint. See [2] for further details. The time-domain constrained estimator (TDCE) was extended for the colored-noise condition by Rezayee and Gazor [5] using the assumption that the noise covariance matrix is a diagonal rather than an identity matrix. This assumption is an approximation and is not always accurate for a real noise condition, but this assumption makes the problem easier to solve. Evaluation of the drawbacks engendered by disregarding the off-diagonal element of the noise covariance remains.

On the other hand, Lev-Ari and Ephraim [3] presented an accurate solution to the general colored noise case utilizing a whitening approach without an expedient assumption. However, whitening also converts the autocorrelation function, which is an important function for the VQ-based method we propose. This VQ-based method can become more complicated by this conversion. For this reason, we do not employ a noise-whitening approach, but instead apply the assumption of the noise diagonal matrix to the spectral domain constrained estimator (SDCE), which we adopted for VQ-based method because Ephraim *et al.* pointed out that SDCE offers performance and implementation advantages. In this section, we recall the fundamental of SDCE and discuss the calculation cost of the covariance matrix and the spectral component power along the KLT axis.

A. Signal Model

Let $\boldsymbol{y}$, $\boldsymbol{w}$, and $\boldsymbol{z}(=\boldsymbol{y}+\boldsymbol{w})$, respectively, represent the K-dimensional vectors of the clean speech, the additive noise, and the noisy speech signals. Additive noise is assumed to be uncorrelated to the speech signal. Consider linear filtering with $\boldsymbol{H}$, which is a $K \times K$-dimensional matrix, to estimate the speech signal as

$$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{z} = \boldsymbol{H}\boldsymbol{y} + \boldsymbol{H}\boldsymbol{w} \tag{1}$$

where $\hat{\boldsymbol{y}}$ is the estimate of the speech signal vector. Then, the residual error signal is given as

$$\boldsymbol{r} = \hat{\boldsymbol{y}} - \boldsymbol{y} = (\boldsymbol{H} - \boldsymbol{I})\boldsymbol{y} + \boldsymbol{H}\boldsymbol{w} = \boldsymbol{r}_y + \boldsymbol{r}_w \tag{2}$$

where $\boldsymbol{r}_y = (\boldsymbol{H} - \boldsymbol{I})\boldsymbol{y}$ represents signal distortion, and $\boldsymbol{r}_w = \boldsymbol{H}\boldsymbol{w}$ represents the residual noise [2]. Furthermore, let $\boldsymbol{R}_y$, $\boldsymbol{R}_w$, and $\boldsymbol{R}_z$, respectively, denote covariance matrices of $\boldsymbol{y}$, $\boldsymbol{w}$, and $\boldsymbol{z}$. Because the noise signal is uncorrelated to the speech and additive, $\boldsymbol{R}_z = \boldsymbol{R}_y + \boldsymbol{R}_w$ holds.

The eigendecomposition of $\boldsymbol{R}_y$ is given as

$$\boldsymbol{R}_y = \boldsymbol{U}\boldsymbol{\Lambda}_y\boldsymbol{U}^T \tag{3}$$

$$\boldsymbol{U} = [\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{K-1}] \tag{4}$$

$$\boldsymbol{\Lambda}_y = \mathrm{diag}[\lambda_{y,0}, \lambda_{y,1}, \ldots, \lambda_{y,K-1}] \tag{5}$$

where $(\cdot)^T$ denotes the transpose operation, and $\boldsymbol{\Lambda}_y$ denotes a diagonal $K \times K$ matrix that contains clean speech covariance matrix eigenvalues. $\boldsymbol{U}$ contains its eigenvectors.

B. Spectral Domain Constrained Estimator for Colored Noise

The problem to solve for obtaining SDCE is given as

$$\boldsymbol{H}_{\mathrm{opt}} = \arg\left[\min_{H} \overline{\epsilon_y^2}\right] \tag{6}$$

subject to

$$\overline{\epsilon_{w,k}^2} \leq \alpha_k \sigma_k^2, \quad k = 0, 1, \ldots, K-1 \tag{7}$$

where $\boldsymbol{H}_{\text{opt}}$ is the optimum filter to provide SDCE, and $\epsilon_{w,k}^2 = |\boldsymbol{u}_k^T \boldsymbol{H} \boldsymbol{w}|^2$ is the kth spectral component power of residual noise. $\epsilon_y^2 = \text{tr}(\boldsymbol{r}_y \boldsymbol{r}_y^T)$ is the power of the signal distortion vector $\boldsymbol{r}_y$, $\sigma_k^2 = \boldsymbol{u}_k^T \boldsymbol{R}_w \boldsymbol{u}_k$ is the kth spectral component power of noise in the input signal, and $\alpha_k (0 \leq \alpha_k \leq 1)$ is the constant to restrict the upper limit of the residual noise power along axis k. To solve the above problem, let the Lagrangian be

$$L(\boldsymbol{H}, \mu_0, \ldots, \mu_{K-1}) = \overline{\epsilon_y^2} + \sum_{k=0}^{K-1} \mu_k \left(\overline{\epsilon_{w,k}^2} - \alpha_k \sigma_k^2 \right). \tag{8}$$

From $\nabla_H L = 0$, it can be shown that $\boldsymbol{H}_{\text{opt}}$ must satisfy the following matrix equation:

$$(\boldsymbol{U}^T \boldsymbol{H} \boldsymbol{U} - \boldsymbol{I}) \boldsymbol{\Lambda}_y + \boldsymbol{\Lambda}_\mu \boldsymbol{U}^T \boldsymbol{H} \boldsymbol{R}_w \boldsymbol{U} = 0 \tag{9}$$

where $\boldsymbol{\Lambda}_\mu = \text{diag}[\mu_0, \mu_1, \ldots, \mu_{K-1}]$. The Kuhn–Tucker conditions for constrained minimization

$$\mu_k \geq 0 \tag{10}$$

are also necessary to guarantee the solution to (9) as optimal.

Now, we introduce the assumption on colored noise covariance proposed by Rezayee *et al.* as

$$\begin{aligned} \boldsymbol{U}^T \boldsymbol{R}_w \boldsymbol{U} &\simeq \boldsymbol{\Lambda}_w \\ &= \text{diag}\left[\sigma_0^2, \ldots, \sigma_k^2, \ldots, \sigma_{K-1}^2\right]. \end{aligned} \tag{11}$$

Rezayee *et al.* used $\boldsymbol{\Lambda}_w$ as the approximation of $\boldsymbol{U}^T \boldsymbol{R}_w \boldsymbol{U}$ to adopt TDCE for colored-noise conditions. This assumption is expedient for fast implementation, as stated previously. From (11), the eigendecomposition of the noisy speech covariance is given as

$$\boldsymbol{R}_z = \boldsymbol{R}_y + \boldsymbol{R}_w \simeq \boldsymbol{U}(\boldsymbol{\Lambda}_y + \boldsymbol{\Lambda}_w)\boldsymbol{U}^T. \tag{12}$$

This equation indicates that we treat the eigenvectors of the noise covariance matrix as identical to those of the speech covariance matrix. See [5] for more details.

Next, we apply this assumption to SDCE for implementing our VQ-based method. By substituting $\boldsymbol{U}^T \boldsymbol{H} \boldsymbol{U}$ with $\boldsymbol{Q}$ and using the assumption $\boldsymbol{R}_w \simeq \boldsymbol{U} \boldsymbol{\Lambda}_w \boldsymbol{U}^T$, (9) becomes

$$\boldsymbol{Q} \boldsymbol{\Lambda}_y + \boldsymbol{\Lambda}_\mu \boldsymbol{Q} \boldsymbol{\Lambda}_w = \boldsymbol{\Lambda}_y. \tag{13}$$

From comparison between the elements of this matrix equation

$$q_{kl} \left(\lambda_{y,l} + \mu_k \sigma_l^2 \right) = 0 \quad (k \neq l) \tag{14}$$

$$q_{kl} \left(\lambda_{y,l} + \mu_k \sigma_l^2 \right) = \lambda_{y,l} \quad (k = l) \tag{15}$$

are necessary. Because $\mu_k = -\lambda_{y,l}/\sigma_l^2 \leq 0$ violates (10), $q_{kl} = 0 (k \neq l)$ is forced. Consequently, we obtain a possible solution to (13) when $\boldsymbol{Q}$ is diagonal, as

$$\boldsymbol{Q} = \boldsymbol{\Lambda}_y (\boldsymbol{\Lambda}_y + \boldsymbol{\Lambda}_\mu \boldsymbol{\Lambda}_w)^{-1}. \tag{16}$$

On the other hand, from (7) we obtain

$$\boldsymbol{u}_k^T \boldsymbol{H} \boldsymbol{R}_w \boldsymbol{H}^T \boldsymbol{u}_k \leq \alpha_k \boldsymbol{u}_k^T \boldsymbol{R}_w \boldsymbol{u}_k \quad (k = 0, 1, \ldots, K-1). \tag{17}$$

Using $\boldsymbol{R}_w \simeq \boldsymbol{U} \boldsymbol{\Lambda}_w \boldsymbol{U}^T$, the kth diagonal element of $\boldsymbol{Q}$ satisfies $q_{kk} \leq \alpha_k^{1/2}$. Furthermore, as in the manner shown in [4], we obtain

$$\begin{aligned} \overline{\epsilon_y^2} &= \text{tr}\left[E\left(\boldsymbol{r}_y \boldsymbol{r}_y^T\right)\right] = \text{tr}\left[\boldsymbol{U}(\boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{\Lambda}_y (\boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{U}^T\right] \\ &= \sum_{k=0}^{K-1} \lambda_{y,k} (q_{kk} - 1)^2. \end{aligned} \tag{18}$$

Because $\alpha_k \leq 1$, we obtain

$$q_{kk} = \alpha_k^{\frac{1}{2}}. \tag{19}$$

From (16), (19), and $\alpha_k \leq 1$, the Kuhn–Tucker conditions are satisfied as

$$\mu_k = \frac{\lambda_{y,k}}{\sigma_k^2} \left(\frac{1}{\alpha_k^{\frac{1}{2}}} - 1 \right) \geq 0. \tag{20}$$

This result is identical to that of the case in which noise is assumed as white [2]. Therefore, we can obtain a filter using

$$\alpha_k = \exp\left(-\frac{\nu \sigma_k^2}{\lambda_{y,k}} \right) \tag{21}$$

with respect to the choice of weighting function as used in [2], [4], where ν is the constant to control the degree of noise suppression.

C. Calculation Cost of KLT-Based Speech Enhancement

KLT-based speech enhancement requires a large calculation cost, which is mainly engendered by KLT bases estimation, covariance matrix estimation, and spectral power estimation along the KLT axis. We discuss the implementation and calculation cost of the last two issues here. KLT bases estimation is replaced

with codebook searching in our method and is described in the next section.

1) Covariance Matrix Estimation: The empirical Toeplitz covariance matrix constructed from the autocorrelation function (e.g., [9]) is often used for speech processing because of its advantages of calculation cost and performance. In the estimation process, the input signal is divided into frames with K samples. The autocorrelation function, truncated by K point length, is estimated from a series of these frames, which contain T frames located before and after the current frame. Therefore, the number of samples for estimating the autocorrelation function is given as $(2T+1)K$. In the literature, $(2T+1)K$ becomes 360 or 440 because $T = 4$ or 5 and $K = 40$ are used at an 8-kHz sampling rate.

The autocorrelation function can be estimated efficiently from this signal via fast Fourier transform (FFT) with 512-point length. Alternatively, we can reduce computation by exploiting duplication of the samples needed for estimation in adjacent frames.

In contrast to the case of empirical Toeplitz covariance estimation, Rezayee *et al.* averaged the outer product $\boldsymbol{z}_n\boldsymbol{z}_n^T$ with the forgetting factor β to estimate the covariance matrix as

$$\boldsymbol{R}_z(n) = \sum_{i=1}^{n} \beta^{(n-i)} \boldsymbol{z}_i \boldsymbol{z}_i^T. \tag{22}$$

They used the frame shift of one sample. Therefore, more than K^3 operations of multiplication and addition arise from this computation to produce an output signal of K samples. Apparently, they avoided this large cost by choosing small K of 20 samples. Nevertheless, this covariance estimation can be a disadvantage when using larger K because larger K tends to provide better performance, as shown in a later experiment.

2) Estimation of Spectral Component Power: Calculation of the spectral component power is required to obtain noise power σ_k^2 in (21) as follows:

$$\sigma_k^2 = \boldsymbol{u}_k^T \boldsymbol{R}_w \boldsymbol{u}_k. \tag{23}$$

The speech signal power can also be obtained using a similar expression. Huang *et al.* proposed a method of fast computation for DCT bases to reduce the computational cost for the same type of expression above [6]. This method is applicable to the proposed method because bases in the VQ clusters are known prior to filtering, as in the DCT-based method. Using this method, the complexity of calculating $\boldsymbol{R}_w\boldsymbol{u}_k$ of $O(K \log K)$ when $\boldsymbol{R}_w$ is Toeplitz is reduced to one inner product of a K-dimensional vector. Therefore, K^2 multiplication operations are required for K vectors.

In the method based on adaptive KLT, Rezayee *et al.* used time averaging with forgetting factor β for this calculation as

$$\sigma_k^2(n) = \sigma^2(n-1)\beta + \left|\boldsymbol{u}_k(n)^T \boldsymbol{w}_i(n)\right|^2. \tag{24}$$

This computation requires more than K^2 multiplication operations for K vectors in each frame, which amounts to more than K^3 multiplication operations to process K signal samples if a frame shift of a one sample is chosen.

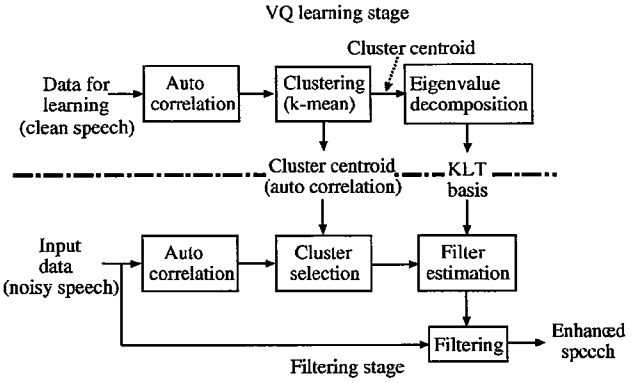

Fig. 1. Block diagram of the proposed method.

As discussed above, a method using an autocorrelation function offers clear advantages in calculation costs of covariance matrix estimation and spectral component power estimation.

III. Fast Implementation of the KLT-Based Method

As presented in Section II-C, conventional KLT-based speech enhancement requires a heavy computational load to perform the following three procedures:

1) KLT bases estimation;
2) spectral component power calculation;
3) autocorrelation estimation.

In addition, autocorrelation-based processing presents some advantages to reduce the computational cost of 2) and 3). Next, we present the proposed system, which comprises a faster version of the above each procedures: KLT-base approximation is accomplished without direct estimation of KLT bases from the signal; spectral component power calculation exploits a precomputed table; autocorrelation estimation based on division of the autocorrelation function into fractional vectors is undertaken and the reuse of fractional vectors among successive frames.

A. System Overview

Fig. 1 shows a block diagram of the proposed speech enhancement system. The upper half of the figure shows the VQ learning stage, in which the codeword vectors of autocorrelation and KLT bases are estimated from learning data of clean speech: the lower half shows the filtering stage, which performs speech enhancement for noisy speech input using the filter calculated from codeword vectors and corresponding KLT bases. In each frame of the filtering stage, the K-dimensional autocorrelation vector is estimated from the input signal and is compared to codeword vectors in the codebook to select the one codeword that is nearest to the input autocorrelation vector. The KLT bases that belong to the same cluster as the chosen vector are used as the approximation of the current KLT bases for calculating the filter. Details of the stages are described below.

1) VQ Learning Stage: Initially, a set of K-dimensional vectors comprising the first K elements of the autocorrelation function is estimated from clean-speech learning data. Clustering for the set of the vectors then determines the codewords that are

the centroid of the cluster. Next, empirical Toeplitz covariance matrices are constructed from codeword vectors. Then, the KLT bases are computed via ED for all clusters. We consulted [8] and [10] for designing VQ cluster and used the K-means algorithm with the initial clusters generated by the maximum distance algorithm [11].

2) Filtering Stage (1) (Cluster Selection): In the filtering stage, the sample autocorrelation vector of input signal $\boldsymbol{\gamma}_z = \{\gamma_z(0), \gamma_z(1), \ldots, \gamma_z(K-1)\}^T$ is estimated from the input signal. That of the speech signal is estimated as

$$\hat{\boldsymbol{\gamma}}_y = \boldsymbol{\gamma}_z - \boldsymbol{\gamma}_w \tag{25}$$

where $\boldsymbol{\gamma}_w$ is the K-dimensional noise autocorrelation vector. This subtraction is intended to validate the codeword vector in the codebook for various noise conditions. Comparison of subtracted vector $\hat{\boldsymbol{\gamma}}_y$ to the codeword vectors in the VQ codebook is followed. Thereby, the nearest one is chosen. We used the squared Euclidean distance for measuring the distance between vectors. Thereby, the chosen cluster index is given as

$$N_c = \arg\left(\min_i |\boldsymbol{c}_i - \hat{\boldsymbol{\gamma}}_y|^2\right) \tag{26}$$

where $\boldsymbol{c}_i$ is the codeword vector belonging to the ith cluster. Eigenvectors corresponding to this number $\boldsymbol{U}_{N_c}$ are utilized to obtain the filter.

Comparison among all clusters can engender a large calculation cost. Therefore, only a fraction of the codebook is used for searching assuming that the autocorrelation function varies slowly from frame to frame. The number of clusters to be sought and compared is denoted as N_{srch} and is limited to $0 \leq N_{\text{srch}} \leq N_{\text{all}}$, where N_{all} is the total number of clusters in the VQ codebook. For this purpose, the indices of nearest N_{srch} clusters are searched for all clusters prior to filtering. Using them, an index table is obtained.

3) Filtering Stage (2) (Filtering): In obtaining the filter, α_k is first calculated using (21). σ_k^2 and $\lambda_{y,k}$ are required for calculating α_k. They are obtained as

$$\sigma_k^2 = \boldsymbol{u}_{N_c,k}^T \boldsymbol{R}_w \boldsymbol{u}_{N_c,k} \tag{27}$$

$$\lambda_{y,k} = \boldsymbol{u}_{N_c,k}^T \hat{\boldsymbol{R}}_y \boldsymbol{u}_{N_c,k} \tag{28}$$

where $\boldsymbol{u}_{N_c,k}$ is the kth eigenvector in $\boldsymbol{U}_{N_c}$, and $\hat{\boldsymbol{R}}_y$ is the Toeplitz speech covariance matrix constructed from $\hat{\boldsymbol{\gamma}}_y$. Calculations of (27) and (28) are performed efficiently using the method presented in the next section.

Next, the filter for speech enhancement $\boldsymbol{H} = \boldsymbol{U}_{N_c} \boldsymbol{Q} \boldsymbol{U}_{N_c}^T$ is obtainable using $\boldsymbol{Q} = \text{diag}(\alpha_k^{1/2})$, the enhanced signal is estimated by

$$\hat{\boldsymbol{y}} = \boldsymbol{U}_{N_c}\left(\boldsymbol{Q}\left(\boldsymbol{U}_{N_c}^T \boldsymbol{z}\right)\right). \tag{29}$$

The final output signal is obtainable using a standard overlap–add synthesis.

Noise covariance required in (27) should be updated to adapt the system to noise condition changes. We used the maximum spectral component power to determine noise-only frames for this update, which is the maximum one among $\lambda_{y,0}, \ldots, \lambda_{y,K-1}$ according to the measure for determining the noise/speech frame as proposed by Rezayee *et al.* When the frame is determined as the noise frame and the number of consecutive noise frames is larger than $2T+1$, both the noise covariance update and the above filtering are performed, whereas filtering alone is performed otherwise. Noise covariance is updated using the autocorrelation vector obtained from the consecutive noise signal.

B. Faster Computation of the Spectral Component Power

The calculation cost of the spectral component power emerging in (27) and (28) is reduced by exploiting the method proposed in [6]. This method is applicable because the empirical Toeplitz matrix is symmetrical and all KLT bases are known prior to filtering in the proposed method. Using this method, (27) and (28) are performed using

$$\sigma_k^2 = \boldsymbol{b}_{N_c} \boldsymbol{\gamma}_w \tag{30}$$

$$\lambda_{y,k} = \boldsymbol{b}_{N_c} \hat{\boldsymbol{\gamma}}_y \tag{31}$$

where $\boldsymbol{b}_{N_c}$ is the precalculated K-dimensional vector with the jth component

$$b_{N_c,j} = \begin{cases} \sum_{i=0}^{K-1} u_{N_c,i} u_{N_c,i}, & j = 0 \\ 2\sum_{i=0}^{K-1-j} u_{N_c,i} u_{N_c,i+j}, & j = 1, \ldots, K-1. \end{cases} \tag{32}$$

This vector $\boldsymbol{b}_{N_c}$ should be calculated for all KLT base vectors in all clusters before filtering.

C. Reduced Computation of Autocorrelation Vector

This computational cost cannot be disregarded because the autocorrelation vector is estimated from $(2T+1)K$ samples of the input signal. Typically, FFT is used for this estimation. However, calculation using only FFT is not absolutely efficient because this method computes the autocorrelation function of FFT length, whereas its first K components are needed here. Therefore, we propose a method to further reduce the computational cost by exploiting the fact that the two series of signals for estimating autocorrelation vectors corresponding to two adjacent frames have duplication of $2TK + K/2$ samples, whereas the difference is only $K/2$ samples when the frame shift is chosen as $K/2$.

Fig. 2 shows data segments for estimating autocorrelation vectors. Let $\boldsymbol{z}_n$ denote the K-dimensional signal vector for the current frame n, $z(m)$ denote the sample of beginning point of $\boldsymbol{z}_n$, and g_n denote the signal segment required for estimating the autocorrelation of nth frame, where m is the sample number. In addition, s_{n+j}, $(j = \ldots, -1, 0, 1, \ldots)$ denotes those signal segments that have length of frame shift of L. Furthermore, we assume that frame length K is an integer multiple of frame shift

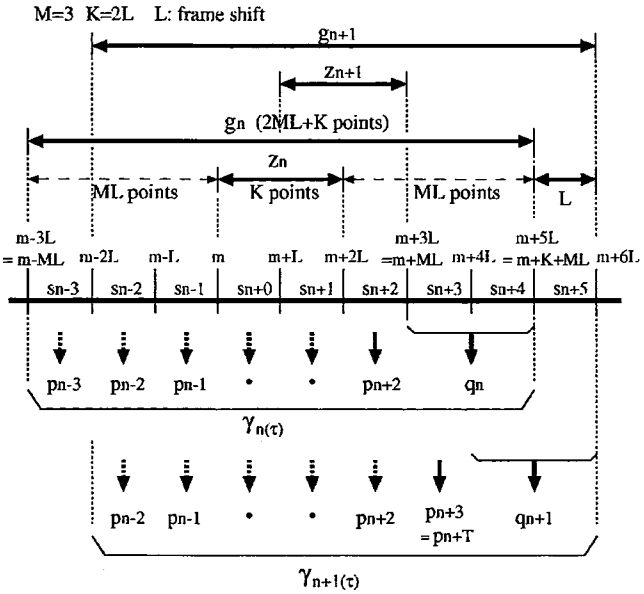

Fig. 2. Data segments for estimating the autocorrelation function with reduced computation.

L and that g_n has ML signal samples before and after the vector $\boldsymbol{z}_n$, which implies that the length of g_n amounts to $2ML+K$. It is noteworthy that $T = ML/K$, as demonstrated using the notations herein. Fig. 2 shows the case where $M = 3$ and $K = 2L$.

Next, we compute parts of the autocorrelation vectors $p_{n+j}(\tau)$, $(j = -M, \ldots, M-1)$ corresponding to the signal vector s_{n+j}, $(j = -M, \ldots, M-1)$ as follows:

$$p_{n+j}(\tau) = \frac{1}{L}\sum_{i=0}^{L-1} z(m+jL+i)z(m+jL+i+\tau) \quad (\tau = 0, 1, \ldots, K-1). \tag{33}$$

The above calculation requires $K+L$ samples from $z(m+jL)$ to $z(m+jL+K+L-1)$. In addition, $q_n(\tau)$ is another part of the autocorrelation vector corresponding to a segment of g_n's end consisting of K samples. Examples are the connected segments of s_{n+3} and s_{n+4}, as shown in Fig. 2. Thereupon, $q_n(\tau)$ is given as

$$q_n(\tau) = \frac{1}{K}\sum_{i=0}^{K-1-\tau} z(m+ML+i)z(m+ML+i+\tau) \quad (\tau = 0, 1, \ldots, K-1). \tag{34}$$

From (33) and (34), the autocorrelation vector estimated from g_n is given as

$$\gamma_n(\tau) = \frac{1}{N}\sum_{i=0}^{N-1-\tau} g_n(i)g_n(i+\tau) = q_n(\tau) + \sum_{j=-M}^{M-1} p_{n+j}(\tau). \tag{35}$$

In the same manner, $\gamma_{n+1}(\tau)$ is expressed as

$$\gamma_{n+1}(\tau) = q_{n+1}(\tau) + \sum_{j=-M}^{M-1} p_{n+1+j}(\tau). \tag{36}$$

From comparison of (35) with (36), we find that $p_{n+j}(j = -M+1, \ldots, M-1)$ emerges both in $\gamma_n(\tau)$ and $\gamma_{n+1}(\tau)$. This fact implies that the essential computation in each frame is reduced to $p_{n+M}(\tau)$ and $q_{n+1}(\tau)$. These terms require KL and $K(K+1)/2$ multiplication operations of real numbers, which amounts to $K^2 + K/2$ when $K = 2L$.

The FFT length N_F should be larger than $2LM + 2K$ if FFT is used for estimating autocorrelation. The last K samples of the length correspond to the zero pad required to avoid contamination caused by convolution via FFT. For example, real valued FFT and IFFT with $N_F = 512$ can be used in cases where $K = 40$, $L = 20$, and $M = 6(T = 3)$ are used. That fact implies that the method using FFT requires $2(N_F/2)\log_2(N_F/2) = 4096$ multiplication operations involving complex numbers, whereas the proposed method requires $K^2 + K/2 = 1620$ real-number operations. Accurate evaluation results and advantages of the proposed method are shown in the next section.

IV. Evaluation

We conducted experiments for evaluating both the speech quality and processing speed of the proposed method. Performance of the proposed method was compared with the conventional KLT-based method, which employs eigendecomposition. We call this conventional method "KLT-ED" and the proposed method "KLT-VQ" hereafter.

A. Experimental Conditions

Clean speech signals were recorded for the VQ learning stage and as test data for filtering evaluation. We used 492 phoneme-balanced Japanese words [12] uttered by five females for the learning stage and used 100 Japanese city names [13] uttered by five females as the test signals for filtering evaluation. The averaged results obtained using these five female speeches are presented in this section. We additionally used test signals uttered by five males in Section IV-D. No common utterer exists for the learning data set and test data set. Speech signals for evaluation were added by noise recorded in a car moving at 100 km/h to obtain desired segmental SNRs (Seg-SNRs). The Seg-SNRs were computed during the speech periods. Because the noise contained high-level low-frequency components, the noise signal was high-pass filtered with a 150-Hz cutoff frequency. The sampling frequency was 11 kHz. Speech quality was evaluated using Seg-SNR and a subjective listening test.

B. Preliminary Setting for VQ

First, to ensure the effect of the number of clusters for the codebook search, which is denoted by N_{srch}, the KLT-VQ processed signal was evaluated in terms of Seg-SNR changing N_{srch}. The constant to control noise reduction ν, frame shift L, and number T, which relates the number of samples for autocorrelation estimation, were chosen respectively as $\nu = 3.0$,

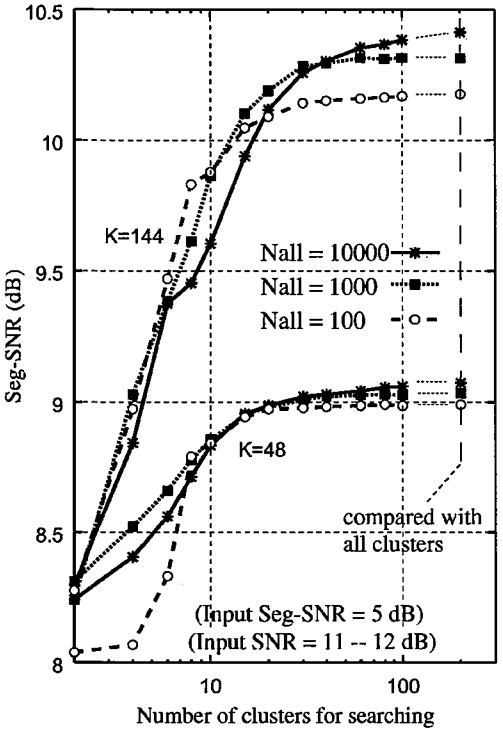

Fig. 3. Seg-SNR versus number of clusters for searching $N_{\rm srch}$ (averaged result using five female speech, $T = 3, \nu = 3.0, L = K/2, N_{\rm all} \simeq 10\,000$, 1000, 100).

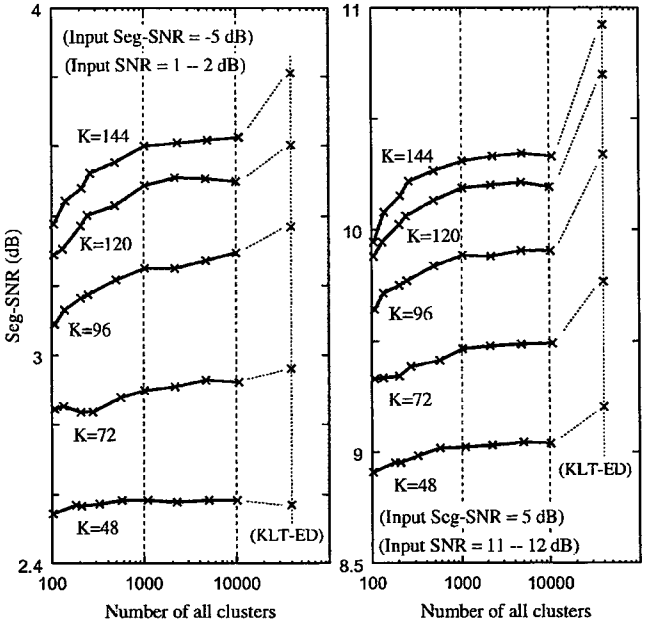

Fig. 4. Seg-SNR versus number of all clusters $N_{\rm all}$ evaluated in the proposed method ($T = 3, \nu = 3.0, L = K/2, N_{\rm srch} = 100$).

$L = K/2$, and $T = 3$. These were determined in the preliminary experiment and were used unchanged thereafter. Fig. 3 shows results obtained for $K = 48$ and 144 at input Seg-SNR of 5 dB.

The numbers of clusters $N_{\rm all}$ were arranged to be about 10 000, 1000, and 100. However, they actually became 10 550 for $K = 48$ and 10217 for $K = 144$ when $N_{\rm all} \simeq 10\,000$, for example, because the number of clusters cannot be determined explicitly using the clustering algorithm that we selected. The actual number of clusters $N_{\rm act}$ was adjusted to be $N_{\rm all} \leq N_{\rm act} \leq 1.1 \cdot N_{\rm all}$ for each $N_{\rm all}$.

As shown in Fig. 3, the KLT-VQ performance increases as $N_{\rm srch}$ increases. In the case of $N_{\rm srch} = 100$, it reaches almost the same Seg-SNR as those obtained in cases where all clusters were used for comparison. Moreover, smaller $N_{\rm srch}$ appears to be possible when $N_{\rm all}$ and/or K are small. Similar results were obtained when other values of K and input Seg-SNR were used. Therefore, searching only a fraction of the codebook is sufficient because it achieves nearly equal performance to that of a full search.

Next, a KLT-VQ processed signal was evaluated by changing $N_{\rm all}$ to investigate the effect of numbers of all clusters. Based on the results described above, $N_{\rm srch}$ was set to 100 in this experiment. Fig. 4 shows results obtained when $K = 48$, 72, 96, 120, and 144. The left figure is for input Seg-SNR $= -5$ dB; the right figure is for input Seg-SNR $= 5$ dB. Results of KLT-ED are also plotted in those figures.

As shown in Fig. 4, larger $N_{\rm all}$ and larger K provide higher performance. Results of KLT-ED were better than those obtained using KLT-VQ, particularly when large K was used at higher SNR. When used with $K = 144$, the difference between KLT-ED and KLT-VQ was 0.2 dB at input Seg-SNR $= -5$ dB; it was 0.6 dB at SNR $= 5$ dB. We observe that the curve gradient of the performance decreases and becomes almost constant as $N_{\rm all}$ gets larger. From these figures, $300 \lesssim N_{\rm all} \lesssim 1000$ appears to be appropriate.

This method consumes memory, mainly for storing eigenvectors of VQ clusters and a vector table for fast power calculation (32). Because these two sizes are the same, the summed size is given by

$$M_{\rm size} = 8 \cdot K^2 N_{\rm all} \quad \text{bytes} \tag{37}$$

when single precision floating point data are assumed. For example, this size is 34.6 MB when $N_{\rm all} = 300$ with $K = 120$; 165.9 MB when $N_{\rm all} = 1000$ with $K = 144$. These are realistic sizes for contemporary personal computers.

Regarding the problem of learning data size for the VQ cluster, we varied the amount of data to use. We used speech signals uttered respectively by 1, 5, and 20 females to learn the VQ cluster. As a result, the performance in the case when one female's speech signal was used for VQ cluster learning was obviously lower than that in the case when five speakers' speech signals were used, whereas the performance in case when 20 speakers' speech were used was almost the same as that of five speakers. From these results, we conducted evaluations using the VQ cluster obtained from speech signals uttered by the five females hereafter.

C. Efficiency of the Covariance Assumption

We examined the assumptions of the noise covariance matrix (11). For this purpose, we calculated the relative average mag-

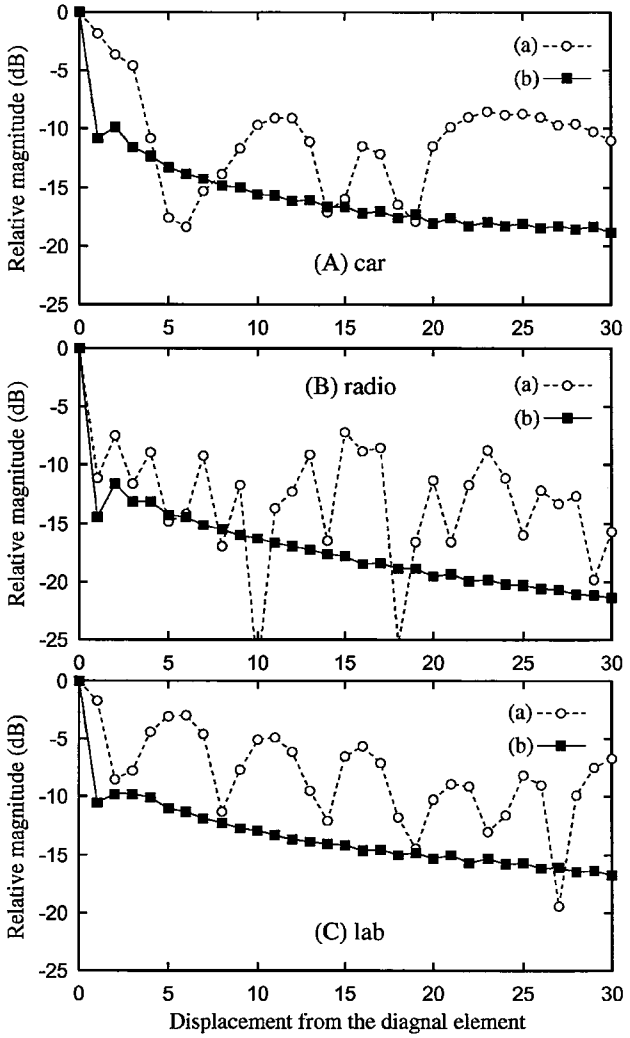

Fig. 5. Relative magnitudes of the off-diagonal elements: curve (a) is obtained from noise covariance matrix $\boldsymbol{R}_w$, and curve (b) is obtained from the transformed matrix $\boldsymbol{\Lambda}_w = \boldsymbol{U}^T\boldsymbol{R}_w\boldsymbol{U}$.

nitude of the kth off-diagonal elements of the matrix $\boldsymbol{A} = (a_{i,j})$ as

$$M(\boldsymbol{A})_k = 10\log\left\{\frac{\frac{1}{K-k}\sum_{i=0}^{i<K-k}|a_{i,i+k}|}{\frac{1}{K}\sum_{i=0}^{i<K}|a_{i,i}|}\right\}, \quad k = 0,1,\ldots,K-1. \quad (38)$$

We present the relative magnitude corresponding to $\boldsymbol{R}_w$ and $\boldsymbol{U}^T\boldsymbol{R}_w\boldsymbol{U}$, as shown in Fig. 5. This figure consists of three panels, which correspond to three noise conditions. The noises are denoted as (A) "car," (B) "radio," and (C) "lab." Noise (A) is the same as that in the previous section. Noise (B) contains speech from radio in the car moving at 15 km/h. Noise (C) contains mainly computer fan noise in a laboratory room. The noise covariance matrix $\boldsymbol{R}_w$ is estimated from each noise signal, and $\boldsymbol{R}_w$ is transformed by $\boldsymbol{U}^T\boldsymbol{R}_w\boldsymbol{U}$, where $\boldsymbol{U}$ is the eigenvector in the VQ cluster. The VQ cluster is learned with parameters of $T = 3$, $K = 48$, $L = K/2$, and $N_{\text{all}} \simeq 1000$. The transformation $\boldsymbol{U}^T\boldsymbol{R}_w\boldsymbol{U}$ is performed for all clusters. The resultant $M(\boldsymbol{U}^T\boldsymbol{R}_w\boldsymbol{U})_k$ are averaged for all clusters. In this figure, curve (a) indicates $M(\boldsymbol{R}_w)_k$, and curve (b) indicates the averaged $M(\boldsymbol{U}^T\boldsymbol{R}_w\boldsymbol{U})_k$.

We observe that the average amplitudes of the off-diagonal elements $M(\boldsymbol{U}^T\boldsymbol{R}_w\boldsymbol{U})_k$, $(k > 0)$ are mostly lower than those of $M(\boldsymbol{R}_w)_k$. Those amplitudes are lower than -10 dB in all the noise cases. We consider that this result supports the assumption (11) to a certain extent, but we cannot confirm that they are sufficiently small to be negligible. Therefore, the effect attributable to this error should be examined, but such an investigation is beyond the scope of this paper. We provided a simple solution that is implied by this assumption in our method.

D. Speech Quality Evaluation in Seg-SNR

We evaluated the speech quality provided by the proposed method KLT-VQ in terms of Seg-SNR. Cluster parameters inferred from the discussion in Section IV-B were used. Fig. 6 shows the obtained results when $N_{\text{all}} \simeq 1000$, 200, 100, and $N_{\text{srch}} = 100$ at input Seg-SNR $= -5$, 0, 5 dB. Frame length K was varied from 36 to 162 with a 12-sample step. This figure portrays two panels, (A) and (B), which show results obtained respectively using female and male speech.

As shown in Fig. 6(a), which is obtained in the case of female speech, as K yields larger, performance of all the methods increases and the curve gradient decreases. Performance of KLT-VQ decreases slightly as K increases when input Seg-SNR = 5 dB in the case of male speech, as illustrated in Fig. 6(b). In contrast to the female case, the curve gradient of KLT-ED is much smaller than that of female speech. Degradation of KLT-VQ compared to KLT-ED for male speech is attributable to the fact that no male speech signal is used for VQ cluster learning. A shortage of learning data when using large K should also be expected because learning with larger dimensions generally requires more numerous data. We note that performance of KLT-ED between female and male speech varying K differs greatly. This difference is an interesting problem that should be investigated in future work.

Results of female speech imply that, whereas the larger K is desirable for better performance, the efficiency of the calculation cost to performance decreases because of the larger K. From an efficiency perspective, $K \simeq 120$ appears to be appropriate from Fig. 6(a). Although $K \simeq 120$ represents a much larger value than those used in the literature, the processing speed is not affected markedly in KLT-VQ, as shown in the next section. We consider that the memory requirement for KLT-VQ is not a serious problem when $K \simeq 120$ and $N_{\text{all}} \leq 1000$ are chosen as described in Section IV-B.

Regarding the performance for female speech, the performance of KLT-VQ is lower than that of KLT-ED as K becomes large at high SNR. However, the difference between the two methods is, for example, 0.12 dB when $K = 120$ at input Seg-SNR $= -5$ dB, with no marked degradation. Furthermore, the difference is 0.52 dB (10.70−10.18) when $K = 120$ at input Seg-SNR = 5 dB; moreover, the difference between 10.7 and 10.18 dB is not an onerous problem in practice. The audible difference is to be tested by subjective listening test in the next section.

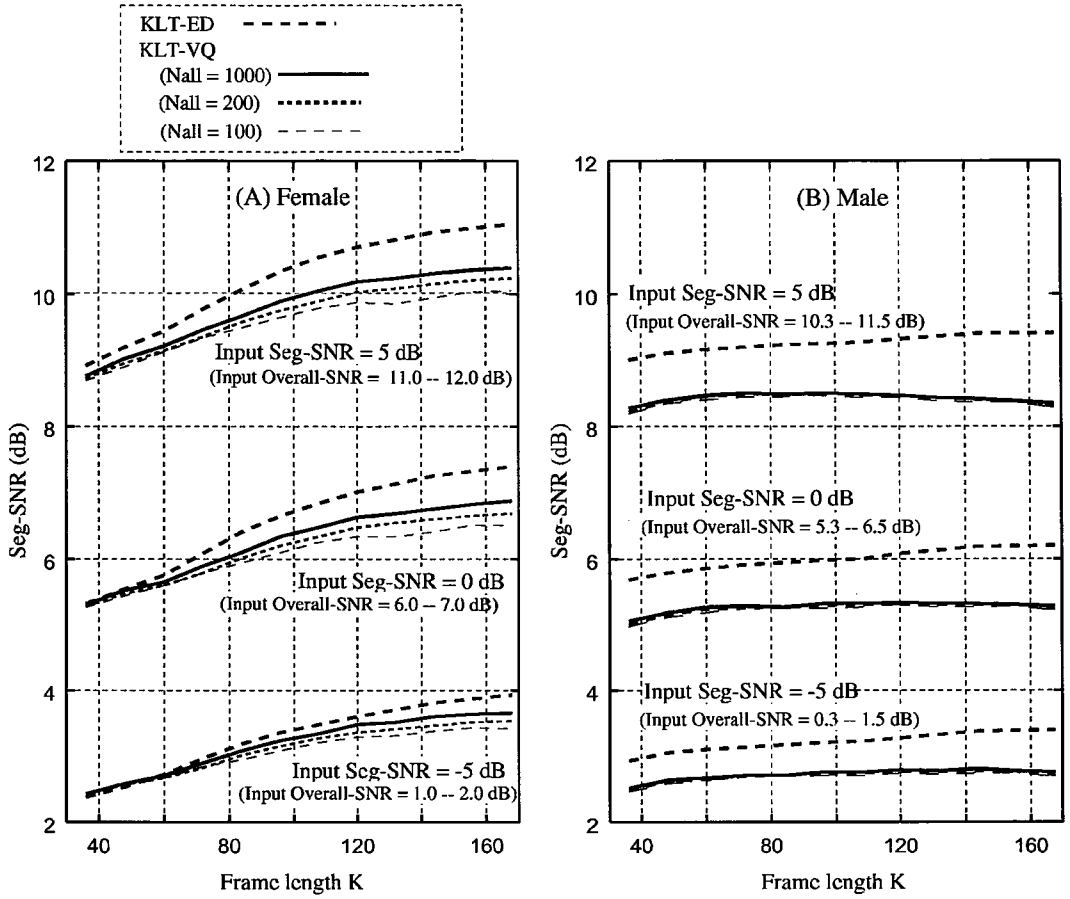

Fig. 6. Performance of KLT-ED and KLT-VQ versus frame length K in terms of Seg-SNR ($T = 3, \nu = 3.0, L = K/2, N_{\mathrm{srch}} = 100$), (VQ cluster is designed using only female speech signals).

E. Subjective Listening Test

We show results of a subjective listening test for speech quality evaluation. We used a preference test algorithm similar to those used in [2] and [5] and performed an additional confidence test to infer the similarity between the signals. Ten subjects participated. All were engineering students whose ages were 21–23. We used speech material comprising three consecutive words among the 100 Japanese city names; they were spoken by three female speakers. Two series of three words were used: "Hachinohe, Kesennuma, Yukuhashi" and "Sapporo, Kitami, Eniwa." Consequently, six speech signals were prepared for each test stage. These six speech signals were added by noises with Seg-SNRs of −5, 0, and 5 dB. We used computer-generated white Gaussian noise and real noise samples recorded in the laboratory room, along with the previously described car noise. The laboratory noise mainly contains computer fan noise; the car noise is the same as that used in the previous section. These three noises are denoted as "white," "lab," and "car." Subsequently, they were processed using the KTL-ED and KLT-VQ methods. Subjects were presented two pairs of signals through headphones. One of the pairs consisted of a KLT-VQ enhanced signal and a KLT-ED enhanced signal; the other consisted of a KLT-VQ enhanced signal and a nonprocessed signal.

First, at each test stage, subjects were asked to choose one of the two signals. Subsequently, they were asked to choose one of the next three choices about the previous choice of preference:

a) totally uncertain;
b) almost uncertain;
c) other.

We informed the subjects to choose a) when the two presented signals are indistinguishable. We intended to infer the similarity of the signals from the degree of confidence in their preference because the more similar signals might provide lower confidence at their choice.

Table I presents a summary of the results. In this table, the value in the "confidence" column shows the ratio of the number of choice c) (other) to the total number of inquiries. As this table shows, in all noise cases, the preference score of the proposed method, KLT-VQ, compared to KLT-ED is from 40% to 60%. This result indicates that KLT-VQ is equally preferred to KLT-ED, but does not exactly mean that both signals are closely resemblant. For instance, a similar preference value is observed in comparison to a nonprocessed signal in the cases of lab noise and car noise, but the signals are markedly different in these cases. We note that the confidence score is extremely high ($\geq$88%) in the comparison to a nonprocessed signal, whereas those obtained in the comparison to KLT-ED are low. The latter score varies from 25% to 50%, indicating that many trials of

TABLE I
PREFERENCE TEST RESULTS

Noise	Seg-SNR	Compared with non-processed(%)		Compared with KLT-ED(%)	
		preferred	confidence	preferred	confidence
white	5 dB	71	98	54	40
	0 dB	79	96	40	33
	-5 dB	85	96	46	25
lab	5 dB	58	92	56	33
	0 dB	52	96	56	44
	-5 dB	60	88	60	50
car	5 dB	62	96	46	48
	0 dB	46	94	44	44
	-5 dB	46	94	50	50

TABLE II
TRANSLATED PREFERENCE TEST RESULTS USING CONFIDENCE

Noise	Seg-SNR	Compared with non-processed(%)			Compared with KLT-ED(%)		
		KLT-VQ	equal	non-process.	KLT-VQ	equal	KLT-ED
white	5 dB	69	2	29	19	60	21
	0 dB	79	4	17	19	67	14
	-5 dB	83	4	13	15	75	10
lab	5 dB	56	8	36	21	67	12
	0 dB	50	4	46	31	56	13
	-5 dB	54	12	33	40	50	10
car	5 dB	60	4	36	25	52	23
	0 dB	42	6	52	25	56	19
	-5 dB	44	6	50	21	50	29

comparison between KLT-VQ and KLT-ED were done with less confidence.

To clarify the difference between the comparison to a nonprocessed signal and that to KLT-ED, we converted the result of Table I using the confidence score. We regard choices a) and b) of the confidence inquiry as indicating that the preference was chosen with insufficient confidence and that the two signals are similar. We consider the possibility of the low confidence when the two signals for comparison differ greatly because totally different impressions of signals can also make the comparison difficult. However, because KLT-VQ and KLT-ED processed signals are nearly identical, we ignored this possibility in this conversion. Thereby, we translated the sum of the numbers of choices a) and b) to the third choice "equally heard." The "preferred" data are decreased according to the emergence of the third choice. Consequently, Table II is obtained through this translation. We note that the translated result only using choice a) as "equally heard" is almost identical to that shown in Table II because the number of choices b) was very small.

Table II shows that KLT-VQ is preferred to the nonprocessed signal in cases of white noise and lab noise, whereas both signals are equally preferred in the case of car noise. The number of selections of "equal" is very small in the comparison to the nonprocessed signal, which is a reasonable result because the nonprocessed signal and KLT-VQ processed signals differ distinctly and the nonprocessed signal contains much noise. In contrast, comparison with KLT-ED shows clearly that both processed signal is equally preferred, but that the number of choices of the "preferred" is small whereas those of "equal" is much larger. Consequently, we can confirm that the two processed signals closely resemble one another.

TABLE III
IMPLEMENTATION OF PROCESSING

function	implementation	
	KLT-ED	KLT-VQ
(a) autocorrelation estimation	FFT	method(s).III-C
(b) KLT bases estimation	eigen-decomposition	cluster selection
(c) spectral component power estimation	Eq. (27,28)	Eq. (30,31)
(d) filtering	filtering (Eq. 29)	

F. Processing Speed Performance

We measured the processing time for speech enhancement to evaluate the calculation cost of KLT-VQ compared to KLT-ED. A personal computer with a Pentium-III CPU of 1.2-GHz clock frequency was used. The programs of KLT-ED and KLT-VQ were compiled using a gcc compiler with "-O2 -pg" options to measure the processing time. Duration of the speech signal used for evaluation was 140 s. We varied K from 48 to 144 with a 24-sample step and set the number of clusters N_{all} to 10 000. Although $N_{\mathrm{all}} \leq 1000$ appears to be sufficient from the results as indicated in Section IV-B, we used $N_{\mathrm{all}} = 10\,000$ to confirm that no disclosed computation arises from the use of a large memory size.

We employed an autocorrelation estimation with 512-points real-valued FFT when $K = 48$ and 72; we used 1024 point real-valued FFT when $K = 96$, 120, and 144 for KLT-ED. In addition, eigendecomposition was implemented for KLT-ED using the Householder transform and "Tridiagonal QL Implicit," which are given respectively as C programs "tred2()" and "tqli()" in [14]. The KLT bases are treated as single precision floating-point data. Table III lists the implemented function procedures of KLT-ED and KLT-VQ.

Fig. 7 shows resultant processing times measured using the gcc profile. The left figure corresponds to KLT-ED and the right one corresponds to KLT-VQ. Plotted values represent average values of 10 measurements. We observe that KLT-VQ reduced processing times remarkably compared to KLT-ED, particularly in KLT bases estimation and spectral component power calculation. Respective ratios of the processing times corresponding to those procedures were 597:1 and 43:1 when $K = 120$; that of the total processing time was 66:1 when $K = 120$.

The left figure shows that the processing time of the autocorrelation estimation in KLT-ED varies irregularly. It does so because the FFT length varies according to K and the number of FFT required to process the entire signal decreases as K

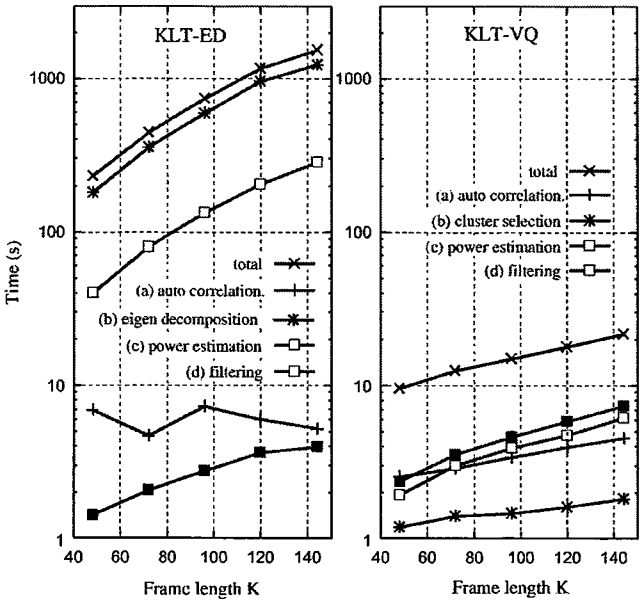

Fig. 7. Results of processing time measurement.

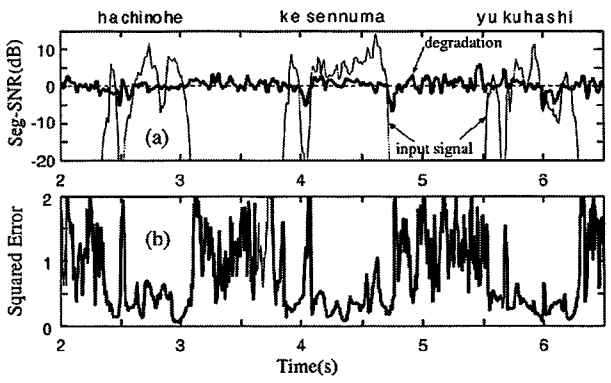

Fig. 8. (a) Degradation in Seg-SNR caused by KLT-VQ compared to KLT-ED denoted by the thick curve and Seg-SNR of test signal denoted by the thin curve. (b) Temporal change in matching error.

increases. The processing time of autocorrelation estimation employed in KLT-VQ was less than half of that employed in KLT-ED when $K = 48$; they were comparable when $K = 144$. In cases where K becomes much larger, e.g., in cases where large sampling frequency is used, estimation using FFT can be efficient again. However, because FFT with small length is applicable to estimate the divided autocorrelation vector $q_{n+M}(\tau)$ (34) and $q_n(\tau)$ (36), the proposed method for autocorrelation estimation does not lose efficiency compared to the case using $(2T+1)K$ points FFT, which directly estimates the autocorrelation vector.

V. VQ Matching Error

In this section, we describe the relation between performance and the error induced by vector quantization in the proposed method. The error mainly arises from matching error between an input autocorrelation vector and the chosen codeword vector for cluster selection. Therefore, the relation between the matching error and Seg-SNR of output signal is mentioned. For that purpose, we calculated the temporal change of the squared distance as calculated by

$$E_{rr}(n) = \left|\boldsymbol{c}_{N_c}(n) - \boldsymbol{\gamma}_y(n)\right|^2 \tag{39}$$

where n is the frame number, and N_c is the cluster number given by (26). We used the test signal with Seg-SNR = 0 dB, which was used in Section IV-D. Fig. 8(b) shows the calculated error.

We additionally present the temporal change of difference in Seg-SNR between the KLT-ED processed signal and the KLT-VQ processed signal, as shown in Fig. 8(a). The positive difference indicates that more degradation results from the use of KLT-VQ than from use of KLT-ED. Fig. 8(a) also includes the temporal change in Seg-SNR of the input test signal. The same period of the test signal is used for the results shown in the two panels; the positions of the uttered words are shown at the top of panel (a). In calculation of Seg-SNR in Fig. 8(a), −30-dB white noise is added to both clear speech and the signal for which Seg-SNR is obtained, thereby avoiding a zero division.

As shown in Fig. 8(b), the error appears to be higher in the noise period than that in the speech period. In the noise period, the subtracted autocorrelation vector given by (25) can become an unpredictable random vector. Because VQ clusters have been learned using only speech signals, none of the clusters can cope with this vector. For that reason, the matching error can become large in the noise period. Nevertheless, as shown in Fig. 8(a), dependency of the degradation of KLT-VQ on the matching error is not obvious. We consider that because this large matching error occurs mainly in the noise period, the quality of the speech period is not affected seriously.

VI. Conclusion

This paper presented a method of fast implementation for KLT-based speech enhancement. The proposed method exploits VQ for approximating KLT bases. It also introduces fast calculation of spectral component power. Furthermore, it uses divided autocorrelation functions among successive frames. Experimental results for evaluating the calculation cost showed that the proposed method reduced total processing time to 1/66 when $K = 120$ compared to the conventional KLT-based method using eigendecomposition. Evaluation of speech quality showed that degradation in terms of Seg-SNR was sufficiently small as to be deemed negligible in practical conditions. Evaluation results of subjective listening tests also supported the results described above. Taken together, these results demonstrate the effectiveness of the proposed method.

Whereas this method resolved the problem of complexity, some problems, e.g., performance for male speech and effects of matching error in the noise period are not addressed. Furthermore, VQ learning data containing male speech and efficient clustering algorithm in place of the K-means algorithm can be explored for improving this method. These subjects are promising as goals of future investigations.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 113–120, Apr. 1979.

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

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

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

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

[6] J. Huang and Y. Zhao, "A DCT-based fast signal subspace technique for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 747–751, Nov. 2000.

[7] C. H. Yang and J. F. Wang, "Noise suppression based on approximate KLT with wavelet packet expansion," in *Proc. ICASSP*, 2002, pp. I-565–I-568.

[8] L. R. Labiner and B. H. Juang, Eds., *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[9] S. L. Marple, Jr., Ed., *Digital Spectral Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[10] J. G. Wilpon and L. R. Labiner, "A modified k-means clustering algorithm for use in isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 6, pp. 587–590, Jun. 1985.

[11] M. Nagao, Ed., *Methods of Image Pattern Recognition* (in Japanese). Tokyo, Japan: Corona, 1983.

[12] K. Tanaka and S. Hayamizu, "ETL speech databases for research (in Japanese)," *J. Acoust. Soc. Japan.*, vol. 48, pp. 883–887, Dec. 1992.

[13] S. Itabashi, "A noise database and Japanese common speech data corpus (in Japanese)," *J. Acoust. Soc. Japan*, vol. 47, pp. 951–953, Dec. 1991.

[14] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, Eds., *Numerical Recipes in C (Japanese Edition)*. Cambridge, MA: Cambridge Univ. Press (Japanese Edition: Gijutsu Hyoron Sha), 1988.

Yoshifumi Nagata received the B.E. degree in electronics in 1984 and the M.E. and Dr. Eng. degrees in information science in 1987 and 1990, respectively, all from Tohoku University, Sendai, Japan.

In 1990, he joined the Research and Development Center, Toshiba Corporation, where he has been engaged in research and development of speech processing systems. Since 1997, he has been an Associate Professor at Iwate University, Morioka, Japan. His interests include multimedia human interface and speech signal processing.

Dr. Nagata is a member of Acoustical Society of Japan and Information Processing Society of Japan.

Kenji Mitsubori received the B.E. and M.E. degrees in information science in 2001 and 2003, respectively, from Iwate University, Morioka, Japan.

Since 2003, he has been an Engineer with the Sendai Technology Center, Chuo Electronics Company, Ltd., Sendai, Japan.

Takahiko Kagi received the B.E. and M.E. degrees in information science from Iwate University, Morioka, Japan, in 2003 and 2005, respectively.

Since 2005, he has been an Engineer with NTT East (Nippon Telegraph and Telephone East Corporation) Akita Group, Akita, Japan.

Toyota Fujioka was born in Akita, Japan, on August 21, 1969. He received the B.E. and M.E. degrees in electrical and electronic engineering from the Mining College, Akita University, in 1992 and 1994, respectively, and the Ph.D. degree in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1997.

He is currently a Research Associate in the Department of Computer and Information Science Faculty of Engineering, Iwate University, Morioka, Japan. His research interests include parallel computer and data compression.

Dr. Fujioka is a member of the Information Processing Society of Japan.

Masato Abe (M'85) received the B.E, M.E., and Ph.D. degrees in electrical engineering from Tohoku University, Sendai, Japan, in 1976, 1978, and 1981, respectively.

From 1981 to 1989, he was a Research Associate with the Research Center for Applied Information Sciences, Tohoku University. From 1989 to 1996, he was an Associate Professor in the Department of Information Science, Iwate University, Morioka, Japan. His research interests include digital signal processing for acoustics and computer architecture.

Dr. Abe is a member of the Acoustical Society of America, Acoustical Society of Japan, the Institute of Noise Control Engineering of Japan, Information Processing Society of Japan, the Institute of Electronics, Information, and Communication Engineers, the Association for Computing Machinery, and the Japan Society of Mechanical Engineers.