ACADEMIC
PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Sound and Vibration 268 (2003) 717–729

JOURNAL OF
SOUND AND
VIBRATION

www.elsevier.com/locate/jsvi

An approach based on simplified KLT and wavelet transform for enhancing speech degraded by non-stationary wideband noise

Hong Wei Lou*, Guang Rui Hu

Department of Electronic Engineering, Shanghai Jiaotong University, 200030 Shanghai, China.

Received 9 May 2002; accepted 27 November 2002

Abstract

It is well known that the non-stationary wideband noise is the most difficult to be removed in speech enhancement. In this paper a novel speech enhancement algorithm based on the dyadic wavelet transform and the simplified Karhunen–Loeve transform (KLT) is proposed to suppress the non-stationary wideband noise. The noisy speech is decomposed into components by the wavelet space and KLT-based vector space, and the components are processed and reconstructed, respectively, by distinguishing between voiced speech and unvoiced speech. There are no requirements of noise whitening and SNR pre-calculating. In order to evaluate the performance of this algorithm in more detail, a three-dimensional spectral distortion measure is introduced. Experiments and comparison between different speech enhancement systems by means of the distortion measure show that the proposed method has no drawbacks existing in the previous methods and performs better shaping and suppressing of the non-stationary wideband noise for speech enhancement.
© 2003 Published by Elsevier Science Ltd.

1. Introduction

When one communicates with another in speech communication, there is a great deal of interferential noise existing in the surrounding environment, transmitting media, electronic communication device and other speakers' sound, etc. The existing noise results from the received speech, not being the clean original speech signal, but a contaminative copy. For this reason, speech enhancement of noisy speech is of great importance. The aim is to extract the original speech from the noisy signal. The purpose of the speech recognition system is to improve the performance and reduce the recognition error rate. In speech communication the purpose is to

*Corresponding author.
E-mail address: louhw@sina.com.cn (H.W. Lou).

0022-460X/03/$ - see front matter © 2003 Published by Elsevier Science Ltd.
doi:10.1016/S0022-460X(02)01556-0

elevate the objective quality of speech signal and the intelligibility of noisy speech in order to reduce the listener fatigue. Speech enhancement is related not only to the theory of digital speech signal processing, but also to the auditory perception and phonetics. Since its origin is of diverse sources, and in different situations, the various noises have different properties. Even in lab simulations, it is difficult to attain a universal speech enhancement algorithm, which fits in with all kinds of noisy environments. So the different measures are taken in accordance with the different noises. The most frequently studied is the additive noises, which include the periodical noise, pulse noise, wideband noise, other speaker's noise, etc. Furthermore the wideband noise can be stationary or non-stationary. Because the shape of wideband noise and speech signal in time and frequency domain overlap each other, it is the most difficult, to remove this type of noise. As for the stationary wideband noise, it is generally considered to be a Gaussian noise and can be whitened beforehand, but it is sophisticated to suppress the non-stationary wideband noise.

Spectral subtraction is a traditional method of speech enhancement [1]. The major drawback of this method is the remaining musical noise resulting from the real-time changeable filter. Additionally a drawback of speech enhancement methods is the distortion of the useful signal. The resolution is the compromise between signal distortion and residual noise. Though this problem is well known, the study results indicate that both of these cannot be minimized simultaneously. Minimum mean square error (MMSE) [2] estimates on speech spectrum have been proposed. And Ephraim and Van Trees proposed a signal-subspace-based spectral domain algorithm [3], which controls the energy of residual noise in a certain threshold while minimizing the signal distortion. Hence the probability of noise perception can be minimized. The drawback of this method is that it deals only with white noise. For non-white noise, the measure of noise whitening should be taken by constraining the spectral domain, and so it does not ensure that the residual noise has spectral shape similar to that of the clean speech. This method requires inverting the noise covariance matrix; therefore, it cannot be used for narrowband noise.

Mittal and Phamdo [4] propose an approach which fits in with the colored and stationary noise and does not require noise whitening and matrix inversion, but it deals only with the low-frequency narrowband noise. The approach requires evaluation of noise spectrum power by long-term averaging of the noise-only (silence) frames and estimations of SNR. Based on the a priori knowledge of SNR, the noisy speech frames are classified into two categories; namely, speech-dominated frames and noise-dominated frames. They are enhanced differently, which leads to the large computation, and the resulting impact are poor for the non-stationary wideband noise.

Wavelet theory is the newly developed time–frequency analysis technology and is especially of interest in non-stationary signals such as speech, sonar, seismic signal, etc [5]. One important application of the wavelet transform is to clear up noisy speech. In recent years, there have been many methods, based on the wavelet transform. Mallat [6] has shown that effective suppression of noise may be achieved by transforming the noisy signal into the wavelet domain and removing the noise components by defining appropriate threshold value. Though the speech signal is reconstructed and noise is suppressed to some extent. The requirement for the robustness and precision is difficult to be guaranteed. In the complex background environment, the performance of speech enhancement needs to be improved. In addition, the interfering components resulting from the wavelet transform will impact the speech enhancement.

In this paper, aiming to suppress the complex non-stationary wideband noise, we propose a novel speech enhancement algorithm based on the simplified Karhunen–Loeve transform (KLT) and the Dyadic wavelet transform ($\mathrm{D_yWT}$). There are no requirements of the matrix inversion, but of the processing of decision about the noisy voiced or unvoiced, and the non-stationary wideband noise is easily controlled.

2. DyWT- and KLT-based speech enhancement

2.1. Noise suppression based on D_yWT

If the admissibility condition is satisfied by the wavelet function $\Psi(x) \in L^2$, i.e.

$$\int \Psi(x)\,\mathrm{d}x = 0, \tag{1}$$

then the $\mathrm{D_yWT}$ of signal with the finite energy ($x \in L^2$) is defined by the following formula [5–7]:

$$D_y WT_x(n, 2^j) = \frac{1}{2^{j/2}} \int_{-\infty}^{\infty} x(t)\Psi^*(2^{-j}t - n)\,\mathrm{d}t = x(t) \otimes \Psi^*_{2^j}(t), \tag{2}$$

where 2^j is scale factor, by changing the value of j, contractions and dilations can be achieved. $\Psi^*(t)$ is the complex conjugate of a wavelet function $\Psi(t)$. From signal processing point of view, the $\mathrm{D_yWT}$ can be considered as the output of a bank of constant-Q and octave band-pass filters whose impulse responses are $(1/2^j)\Psi(t/2^j)$. The bandwidth and the center frequency of each such filter are proportional to $1/2^j$.

$\mathrm{D_yWT}$ has the property of time shift invariant and is very useful for several applications such as breakdown point detection. In the practical speech analysis, the signal has many components including pinnacles and breakdown points. The traditional Fourier analysis has a serious drawback. In transforming to the frequency domain, time information is lost. When looking at a Fourier transform of a signal, it is impossible to tell when a particular event took place. Since the wavelet transform has the ability of time-scale analysis at the same time and performing local analysis, the suppression could be realized. During the course of the analysis of speech signal, it is modelled as a linear aggregation of sine wave with various time shift and damp. It is this property that makes wavelet transform of great importance in the analysis of non-stationary signal.

The original signal can be reconstructed by the following equation:

$$x(t) = \sum_{j \in Z} \frac{1}{2^{j/2}} \int D_y WT_x(n, 2^j)\Psi(2^{-j}t - n)\,\mathrm{d}n. \tag{3}$$

Let a noisy signal with one dimension be the form of

$$\mathbf{x}(i) = \mathbf{s}(i) + \mathbf{n}(i), \quad i = 0, \ldots, K - 1, \tag{4}$$

where $\mathbf{s}(i)$ is clean speech signal and $\mathbf{n}(i)$ is noise. In the following section, the noise is limited in the frequency range of non-stationary wideband. The clearing up procedure is detailed as follows:

(a) the wavelet basis and its decomposing level N are adopted and speech signal is decomposed;
(b) processing high-frequency parameters at all levels by appropriate choice of soft threshold;
(c) wavelet reconstruction by merging the Nth approximation coefficients and the detail coefficients from level 1 to N.

The soft threshold function is defined by [8]

$$Th(\omega_{n,j}, \gamma_a) = \begin{cases} 0, & |\omega_{n,j}| \leqslant \gamma_a, \\ sgn(\omega_{n,j})(|\omega_{n,j} - \gamma_a|), & |\omega_{n,j}| \succ \gamma_a, \end{cases} \tag{5}$$

where $\omega_{n,j} = D_y WT_x(n, 2^j)$ is the wavelet decomposition coefficients, γ_a is the dynamic threshold. The key to suppression of noise is how to choose the value of threshold and digitization of it. In this paper we obtain the value of γ_a depicted in Ref. [9].

$$\gamma_a = \sigma\sqrt{2\log(K)}, \tag{6}$$

with

$$\sigma = MAD/0.6745, \tag{7}$$

where median absolute deviation (MAD) is estimated in the first scale.

The biorthogonal wavelet basis is adopted, which has the property of linear phase and widely used in the reconstruction of signal [5].

Fig. 1 shows an example which depicts the method using the wavelet transform to clear up the non-stationary noise. In (a), the noisy signal consists of the original signal and noise. The former is synthesized by several functions, and the latter is non-stationary noise with different SNR in different time sequence. The curve in (b) represents the adaptively de-noised signal using the best soft threshold. We can conclude that the de-noising effect is remarkable in the situation of different SNR by modifying wavelet coefficients of different scales.

2.2. KLT-based noise suppression

Excellent speech enhancement could not be reached by only utilizing the wavelet transform. It is necessary to take an ulterior step to deal with the consequence resulting from the method above.

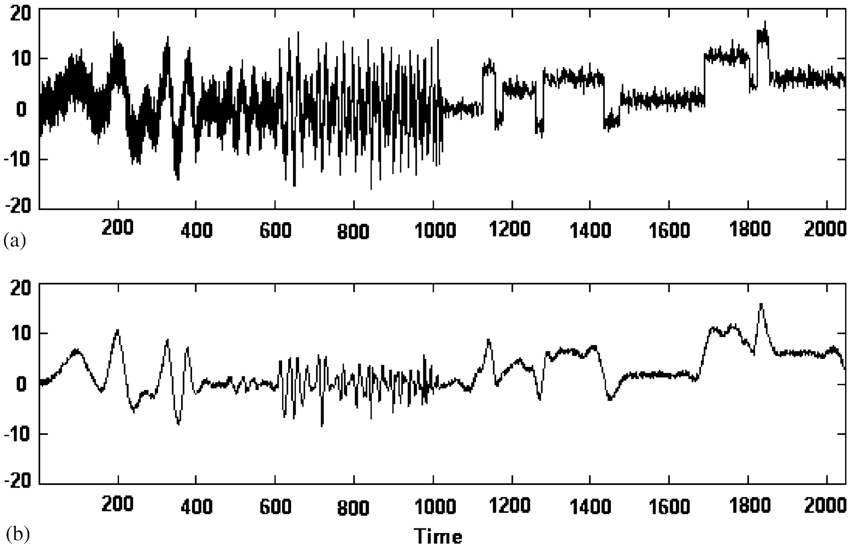

Fig. 1. (a) Signal added with non-stationary noise; (b) de-noised signal by using wavelet method.

Mittal and Phamdo propose an approach which fits in with the colored and stationary noise and does not require noise whitening and matrix inversion, but it deals only with the low-frequency narrowband noise. The approach requires evaluation of noise spectrum power by long-term averaging of the noise-only (silence) frames and estimation of SNR. Based on the prior knowledge of SNR, the noisy speech frames are classified into two categories, namely speech-dominated frames and noise-dominated frames. They are enhanced differently and the computation is large [4]. Now in the subsequent section, a model based on KLT is proposed.

We assume that a $K \times K$ matrix $\mathbf{R_x}$ represents the covariance matrix of the vector $\mathbf{x}$. Let $\hat{\mathbf{x}}$ be an estimate of $\mathbf{x}$ and $\hat{\mathbf{R}}_\mathbf{x}$ be an estimate of $\mathbf{R_x}$. For a given matrix $\mathbf{A}$, $\mathbf{A}'$ is the transpose of A and $tr(\mathbf{A})$ is the trace of $\mathbf{A}$ (the sum of the diagonal elements of $\mathbf{A}$). Let

$$\mathbf{R_x} = \mathbf{U_x \Lambda_x U'_x}, \tag{8}$$

be the decomposition of $\mathbf{R_x}$, $\lambda_x(k)$ be the kth diagonal element of $\mathbf{\Lambda_x}$ and $\mathbf{u_{xk}}$ be the corresponding eigenvector. Let $\mathbf{s}$, $\mathbf{n}$ and $\mathbf{x} = \mathbf{s} + \mathbf{n}$ be K-dimensional vectors of the clean speech, the additive noise and the noisy speech signal, respectively. Since speech and noise are independent, then $\mathbf{R_x} = \mathbf{R_s} + \mathbf{R_n}$. As in Eq. (8), let

$$\mathbf{R_s} = \mathbf{U_s \Lambda_s U'_s}, \tag{9}$$

be the eigenvalue decomposition of $\mathbf{R_s}$. Thus the subspace-based approach is referred as the KLT-based approach. Let $\mathbf{U_s} = \begin{bmatrix} \mathbf{U_1} & \mathbf{U_2} \end{bmatrix}$, where $\mathbf{U_1}$ denotes the $K \times M$ matrix of eigenvectors of $\mathbf{R_s}$ with M positive eigenvalues and $M \leqslant K$. Note that the span of $\mathbf{U_1}$ is the signal subspace and the span of $\mathbf{U_2}$ is orthogonal to the signal space. The subspace-based estimates are based on the assumption that the span of $\mathbf{R_s}$ can be accurately estimated from the data. It is also named the sampled subspace. Note that $\mathbf{U_1 U'_1}$ and $\mathbf{U_2 U'_2}$ are orthogonal projections on the signal subspace and the orthogonal subspace, respectively. The energy of the signal part in the orthogonal projection $\mathbf{U_2 U'_2 x}$ is equal to zero [3]. Hence this part does not provide any information about the signal $\mathbf{s}$. In order to minimize the signal distortion and omit the noise-whitening procedure in speech enhancement, some measures are taken as follows [4].

Assume that speech distortion is $\mathbf{r_s}$ and residual noise is $\mathbf{r_n}$, here $\mathbf{r_s} = (\mathbf{H} - \mathbf{I})\mathbf{s}$ and $\mathbf{I}$ is a unit diagonal matrix. In order to minimize $\mathbf{r_s}$, let ε_s be the trace of the covariance matrix $\mathbf{r_s}$. Now there is an optimization problem given by

$$\min_{\mathbf{H}} \ \varepsilon_s, \tag{10}$$

where $\mathbf{H}$ is the speech enhancement filter based on KLT. The following Eq. (11) is constraints for Eq. (10):

$$E(|\mathbf{u'_{sk} r_n}|^2) \leqslant \alpha_k \sigma_n^2(k), \quad k = 1, 2, \ldots, K, \tag{11}$$

where α_k is the kth constraint constant, $\sigma_n^2(k)$ is the kth diagonal element $\mathbf{U'_s R_n U_s}$. An additional assumption is that

$$\mathbf{H} = \mathbf{U_s Q U'_s}, \tag{12}$$

where $\mathbf{Q}$ is a diagonal matrix. Note that if $\alpha_k = \lambda_s(k)/\sigma_n^2(k)$, the spectrum of residual noise is similar to the spectrum of clean speech provided that equality in Eq. (11) is achieved. Let q_{kk} be the kth diagonal element of $\mathbf{Q}$, then the condition of equality is

$$q_{kk} \leqslant \alpha_k^{1/2} \tag{13}$$

and

$$\varepsilon_x = tr(E\{\mathbf{r}_\mathbf{s}\mathbf{r}'_\mathbf{s}\}) = tr(\mathbf{U}_\mathbf{s}(\mathbf{Q}-\mathbf{I})\mathbf{\Lambda}_\mathbf{s}(\mathbf{Q}-\mathbf{I})\mathbf{U}'_\mathbf{s}) = \sum_{k=1}^{K} \lambda_s(k)(1-q_{kk})^2. \tag{14}$$

Eq. (14) is reasoned out in Appendix A. From Eq. (14), we can attain

$$q_{kk} = \min(1, \alpha_k^{1/2}), \tag{15}$$

which is the necessary condition of Eq. (13).

Following the analysis above, a simplified KLT-based speech enhancement matrix $\mathbf{H}$ is put forward:

$$\hat{\mathbf{s}} = \mathbf{H}\mathbf{x}, \tag{16}$$

which makes the speech enhancement with no requirement of noise pre-whitening. The algorithm is as follows:

(a) First a frame of noisy speech $\mathbf{x}$ is obtained and noise $\mathbf{n}$ in silence is attained by long-term averaging. The covariance matrix $\hat{\mathbf{R}}_\mathbf{x}$ and $\mathbf{R}_\mathbf{n}$ is estimated, respectively. Their calculation is introduced in the next part.

(b) Then the covariance is calculated in terms of the equation $\hat{\mathbf{R}}_\mathbf{s} = \hat{\mathbf{R}}_\mathbf{x} - \mathbf{R}_\mathbf{n}$, based on which, the eigenvalues and eigenvectors of $\hat{\mathbf{R}}_\mathbf{s}$ are computed. The eigenvector is represented in the form of $\mathbf{U}_\mathbf{s} = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}$, where $\mathbf{U}_1$ denotes the $K \times M$ matrix of eigenvectors of $\mathbf{R}_\mathbf{s}$ with M positive eigenvalues and $M \leqslant K$, i.e. $\mathbf{U}_1 = \{\mathbf{u}_{\mathbf{s}k} : \lambda_s(k) > 0\}$.

(c) Let

$$\mathbf{\Lambda} = diag(\alpha_k^{1/2}), \tag{17}$$

$$\alpha_k = \exp\left(-\frac{v\sigma_{n_T}^2(k)}{\lambda_s(k)}\right), \quad k = 1, 2, \ldots, M, \tag{18}$$

where v is an experimentally chosen constant, $\sigma_{n_T}^2(k)$ is the kth diagonal element of $\mathbf{R}_{\mathbf{n}_\mathbf{T}}$, and $\mathbf{n}_\mathbf{T} = \mathbf{U}'_\mathbf{s}\mathbf{n}, \mathbf{R}_{\mathbf{n}_\mathbf{T}} = \mathbf{U}'_\mathbf{s}\mathbf{R}_\mathbf{n}\mathbf{U}_\mathbf{s}$. Assuming that $\mathbf{Q} = \begin{bmatrix}\mathbf{\Lambda} & 0 \\ 0 & 0\end{bmatrix}$, it can be proved that

$$\mathbf{H} = \mathbf{U}_1\mathbf{\Lambda}\mathbf{U}'_1. \tag{19}$$

Eq. (19) is reasoned out in Appendix B. As we can see, the estimation of $\mathbf{U}_2$ can be omitted and the calculating procedure of KLT is simplified.

The calculations of $\hat{\mathbf{R}}_\mathbf{x}$ and $\mathbf{R}_\mathbf{n}$ are as follows.

Let x_t denote the sampling value of noisy speech signal $\mathbf{x}$ at time instant t. The kth autocorrelation coefficient at this sampling instant is evaluated from $K(2T-1)$ samples, which include $K(T-1)$ past samples and KT future samples. Thus $\hat{\mathbf{R}}_\mathbf{x}$ is evaluated from the $K(2T-1)$-*dimensional* vector $[x_{t-K(T-1)+1}, \ldots, x_t, \ldots, x_{t+KT}]$. The kth autocorrelation coefficient is now calculated as

$$R_x(k) = \frac{1}{K(2T-1)} \sum_{i=1}^{K(2T-1)-k} (x_{t-K(T-1)+i})(x_{t-K(T-1)+i+k}), \quad (k = 0, \ldots, K-1). \tag{20}$$

Then $\mathbf{R_x}(k)$ is arranged into $(K \times K)$ matrix $\hat{\mathbf{R}}_\mathbf{x}$ called empirical Toeplitz covariance matrix [3].

$$\hat{\mathbf{R}}_\mathbf{x} = \begin{bmatrix} R_x(0) & R_x(1) & R_x(2) & \ldots & R_x(K-1) \\ R_x(1) & R_x(0) & R_x(1) & \ldots & R_x(K-2) \\ R_x(2) & R_x(1) & R_x(0) & \ldots & R_x(K-3) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ R_x(K-1) & R_x(K-2) & R_x(K-3) & \ldots & R_x(0) \end{bmatrix}. \tag{21}$$

This matrix is found to be quite useful in speech enhancement applications; likewise, this calculating method is used in the estimation of $\mathbf{R_n}$ of noise $\mathbf{n}$.

2.3. D_yWT- and KLT-based speech enhancement

We see that the clear-up method based on wavelet transform can preserve the useful pinnacles and breakdown points in speech signal. Wavelet transform has finer and coarser resolutions, especially in high-frequency region. There are its characteristic properties in suppression of non-stationary noise, which the traditional Fourier analysis could not possess. But it must be noted that it is important not to harm the unvoiced sound in speech signal when applying the soft threshold method. Since the unvoiced sound contains lots of noise like high intelligibility in reconstructed signal, it is necessary to separate unvoiced region from the noisy speech at first. Hence the transform coefficients of voiced and unvoiced speech should be processed differently. Based on these facts, the wavelet transform merged with KLT-based approach without noise pre-whitening can derive a better speech enhancement algorithm in suppression of non-stationary noise.

The steps concerning the proposed method are as follows:

(a) First a frame of noisy speech $\mathbf{x}(k)$ is separated. This frame is classified into the voiced or unvoiced sound by using the method referred to in Ref. [6].

(b) This frame is decomposed into wavelet coefficients with N levels. If this frame belongs to voiced speech, the detail coefficients from level 1 to level N are processed by means of soft threshold. If this frame belongs to unvoiced speech, the detail coefficients only at level N are processed.

(c) Speech signal is reconstructed by merging the approximation coefficients and all the detail coefficients from level 1 to level N, which have been processed in step (b).

(d) The covariance matrix $\hat{\mathbf{R}}_\mathbf{x}$ is estimated by applying Eqs. (20) and (21). In the same way, the covariance matrix $\mathbf{R_n}$ is estimated.

(e) Hence the matrix $\mathbf{U_s}$ is derived from step (d). Moreover $\mathbf{\Lambda}$ and $\mathbf{U_1}$ are calculated using Eqs. (17) and (18).

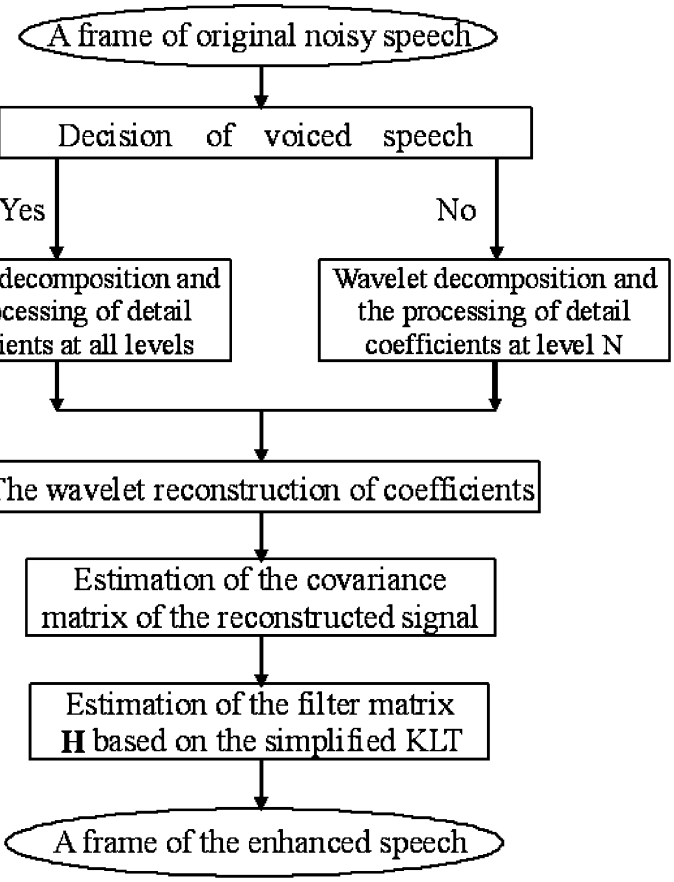
A frame of original noisy speech
Decision of voiced speech
Yes
No
Wavelet decomposition and the processing of detail coefficients at all levels
Wavelet decomposition and the processing of detail coefficients at level N
The wavelet reconstruction of coefficients
Estimation of the covariance matrix of the reconstructed signal
Estimation of the filter matrix **H** based on the simplified KLT
A frame of the enhanced speech

Fig. 2. The flowchart of the proposed method.

(f) The speech enhancement filter matrix **H** is attained according to Eqs. (B.1)–(B.3) in Appendix B. It is estimated using the simplified KLT method. Finally, the enhanced speech $\hat{\mathbf{s}}$ is derived by $\hat{\mathbf{s}} = \mathbf{Hx}$.

The flowchart of this novel algorithm is shown in Fig. 2.

3. Performance evaluation

The proposed algorithm in this paper aims to enhance the noisy speech degraded by the non-stationary wideband noise. At first, a subjective evaluation of it is performed by experiment. A Chinese transcription of female speech is added with noise, which includes AR (1) noise with parameter 0.85 and the Gaussian noise with different SNR in different sample time section. The average SNR is 4 dB. The magnitude spectrum of AR (1) noise with parameter 0.85 is depicted in Fig. 3 and the experimental result is in Fig. 4. As we can see, though the complex and abominable

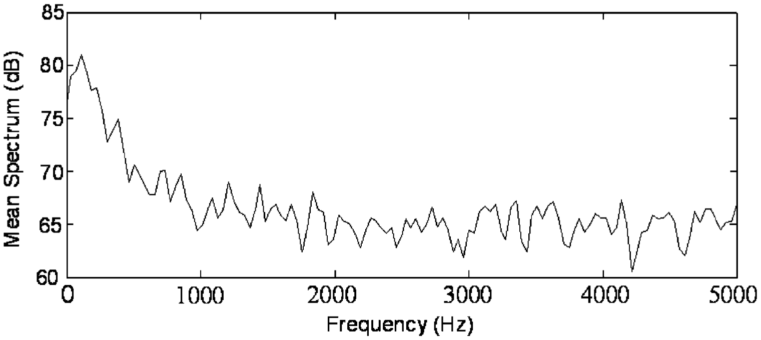

Fig. 3. The magnitude spectrum of noise including AR (1) with parameter 0.85.

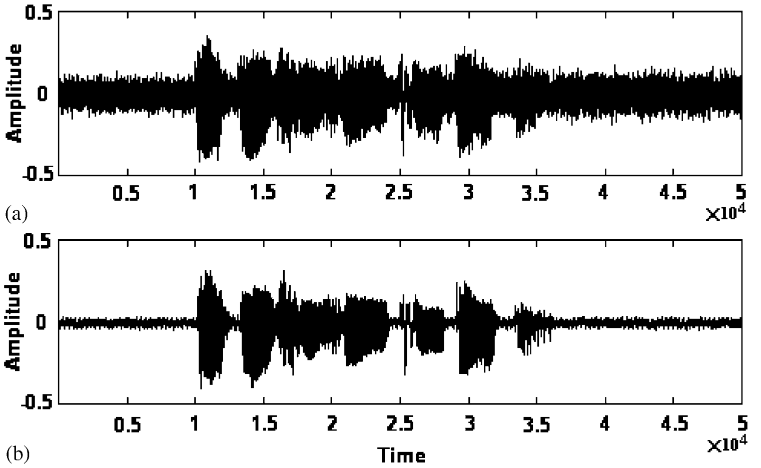

Fig. 4. (a) Speech degraded by non-stationary wideband noise; (b) speech enhanced by the proposed method.

background noise possesses the different SNR and statistical properties, the impact of speech enhancement is remarkable. In the same way other transcriptions added with the same noise are enhanced. We replay the enhanced speech in computer and no musical residual noise occurred.

In order to evaluate the performance of this algorithm in more detail, we make the objective measure. Intelligibility of speech enhancement is an objective measure. One important evaluating tool is speech distortion measure. Here a three-parameter spectral distortion measure is proposed [4]. Let two vectors $\mathbf{s}$ and $\mathbf{x}$ be a clean speech signal and the noise-contaminated speech signal, respectively. The additive noise has the average SNR named *snr*. Assume that the enhanced speech be $\hat{\mathbf{s}}$. The length of $\mathbf{s}$ is N. In order to obtain the speech distortion measure between $\mathbf{s}$ and $\hat{\mathbf{s}}$, two steps are taken beforehand. First, $\mathbf{s}$ and $\hat{\mathbf{s}}$ are normalized to have unit energy (0 dB). Second, a white noise vector $\mathbf{n_0}$ with -30 dB energy is added to them in order to prevent computation of log(0) in a logarithmic distortion measure. The estimation of distortion measure is defined by

$$DM(snr, \hat{\mathbf{s}}, \mathbf{s}) = \sum_{p=1}^{P} \sum_{k=0}^{255} 20|\log|\tilde{\hat{S}}_p(k)| - \log|\tilde{S}_p(k)|, \tag{22}$$

$$\begin{aligned} \tilde{\hat{\mathbf{s}}} &= \hat{\mathbf{s}}/|\hat{\mathbf{s}}| + \mathbf{n_0}, \\ \tilde{\mathbf{s}} &= \mathbf{s}/|\mathbf{s}| + \mathbf{n_0}, \end{aligned} \tag{23}$$

where $\tilde{\hat{S}}_p(k)$ and $\tilde{S}_p(k)$ are the kth frequency components of the pth frame of $\tilde{\hat{\mathbf{s}}}$ and $\tilde{\mathbf{s}}$, respectively. Eq. (22) means the speech distortion between $\mathbf{s}$ and $\hat{\mathbf{s}}$. Using the distortion measure, we compare the proposed algorithm with the speech enhancement system introduced by Ephraim and Van Trees [3]. The speech material consists of 10 sentences, which are spoken by three male and three female speakers. The total transcriptions sum up to 60. Every transcription is added by noise with different SNR (such as 0, 5, 10 and 15 dB). The adopted noise consists of two types. One of them is the Gaussian noise with different SNR in different time sequence added with the AR (1) noise generated by passing artificially generated Gaussian noise through a first order all-pole filter. Fig. 3 shows the spectrum of it. The other is the non-stationary wideband emphasizing particularly on moderate and high frequency, whose spectrum is depicted in Fig. 5. The comparison results are shown in Fig. 6. Every dot means the average distortion measure $DM(snr, \hat{\mathbf{s}}, \mathbf{s})$ of 60 transcriptions. As we can see from Fig. 6, the effects of speech enhancement for

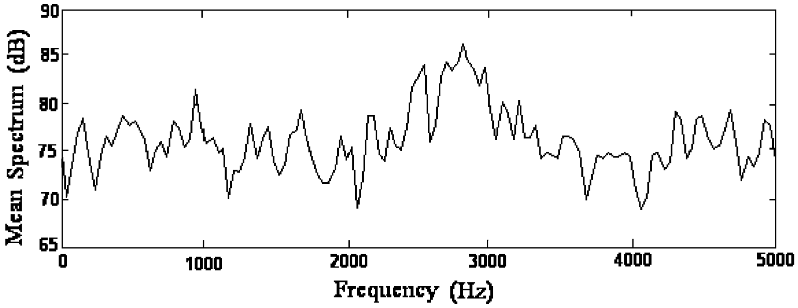

Fig. 5. The magnitude spectrum of the special non-stationary noise.

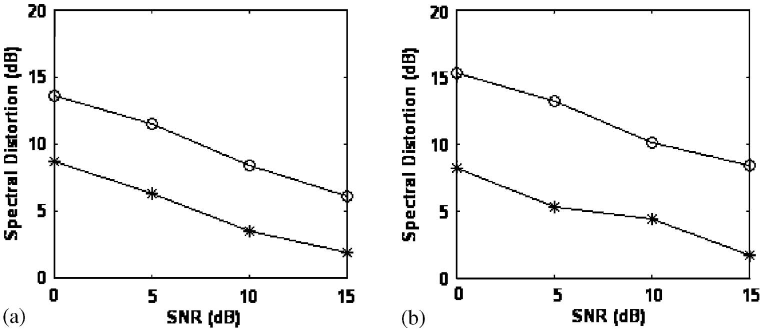

Fig. 6. The proposed method (*) and signal subspace-based method (o): (a) degraded by the noise shown in Fig. 3; (b) degraded by the noise shown in Fig. 5.

the two types of noise are both more remarkable than the signal subspace-based method. The experiment also shows that the impact of speech enhancement is excellent in the region of high frequency. It is well known that the spectrum of the unvoiced sound is overlapped with that of the interfering noise; hence, the traditional speech enhancement systems are difficult to discriminate them and have serious distortion and low intelligibility. In the proposed method, the voiced and unvoiced sound is segmented beforehand and their wavelet transform coefficients are processed dissimilarly. So the drawbacks in traditional methods are gotten over and the non-stationary wideband noise can be suppressed.

4. Conclusions

In this paper, a novel speech enhancement system for non-stationary wideband noise based on the wavelet transform and the simplified KLT transform is proposed. In order to evaluate the performance of this algorithm in more detail, a three-dimensional spectral distortion measure is introduced. By comparison, the proposed approach is very useful at speech enhancement in the situation of non-stationary wideband noise. The noisy speech is enhanced with no musical residual noise.

Appendix A

The derivation of Eq. (14):

$$\varepsilon_x = tr(E\{\mathbf{r_s r_s'}\}) = tr(\mathbf{U_s(Q-I)\Lambda_s(Q-I)U_s'}) = \sum_{k=1}^{K} \lambda_s(k)(1-q_{kk})^2$$

Proof of Eq. (14): Because $\mathbf{r_s} = \mathbf{(H-I)s}$ and $\mathbf{H} = \mathbf{U_sQU_s'}$, the expected value

$$\begin{aligned} E\{\mathbf{r_s r}_s'\} &= E\{\mathbf{(H-I)ss'(H-I)'}\} = \mathbf{(H-I)R_s(H-I)'} \\ &= \mathbf{(U_sQU_s'-I)R_s(U_sQU_s'-I)'} \\ &= \mathbf{(U_sQU_s'-I)U_s\Lambda_sU_s'(U_sQ'U_s'-I)'} \\ &= \mathbf{(U_sQU_s'U_s-U_s)\Lambda_s(U_sQ'U_s'U_s-U_s)'}. \end{aligned} \tag{A.1}$$

From the property of the unit orthogonal matrix $\mathbf{U_s} : \mathbf{U_s'U_s} = \mathbf{I}$, Eq. (A.1) is reasoned into

$$\begin{aligned} E\{\mathbf{r_s r_s'}\} &= \mathbf{(U_sQ-U_s)\Lambda_s(U_sQ'-U_s)'} \\ &= \mathbf{U_s(Q-I)\Lambda_s[U_s(Q-I)]'} \quad (\ \mathbf{Q}=\mathbf{Q'}, \quad \mathbf{I}=\mathbf{I'}\ ) \\ &= \mathbf{U_s(Q-I)\Lambda_s(Q-I)U_s'} \end{aligned} \tag{A.2}$$

Therefore,

$$\begin{aligned} \varepsilon_x &= tr(E\{\mathbf{r_s r_s'}\}), \\ &= tr(\mathbf{U_s(Q-I)\Lambda_s(Q-I)U_s'}). \end{aligned} \tag{A.3}$$

Let the unit orthogonal matrix be $\mathbf{U_s} = \{u_{k,m}\}$, then $\sum_{k=1}^{K} u_{km}^2 = 1$. More assumption

$$\begin{aligned}(\mathbf{Q}-\mathbf{I})\mathbf{\Lambda_s}(\mathbf{Q}-\mathbf{I}) &= diag(\eta_k)\\ &= diag(\lambda_k(1-q_{kk})^2).\end{aligned} \tag{A.4}$$

Then, the diagonal element of $\mathbf{U_s}(\mathbf{Q}-\mathbf{I})\mathbf{\Lambda_s}(\mathbf{Q}-\mathbf{I})\mathbf{U}'_\mathbf{s}$ is equal to $\sum_{m=1}^{K} u_{km}^2 \eta_m$. Hence the sum of the diagonal element is equal to

$$\begin{aligned}\varepsilon_x &= \sum_{k=1}^{K}\sum_{m=1}^{K} u_{km}^2\eta_m = \sum_{m=1}^{K}\eta_m \sum_{k=1}^{K} u_{km}^2\\ &= \sum_{m=1}^{K}\eta_m = \sum_{k=1}^{K}\lambda_s(k)(1-q_{kk})^2.\end{aligned} \tag{A.5}$$

Appendix B

The derivation of Eq. (19):

$$\mathbf{H} = \mathbf{U_s}\mathbf{Q}\mathbf{U}'_\mathbf{s} = \mathbf{U_1}\mathbf{\Lambda}\mathbf{U}'_\mathbf{1},$$

where $\mathbf{Q} = \begin{bmatrix}\mathbf{\Lambda} & 0\\ 0 & 0\end{bmatrix}$ and $\mathbf{\Lambda}$ is the $M \times M$ diagonal matrix.

Proof of Eq. (19): First let $\mathbf{U_s} = [\,\mathbf{U_1} \quad \mathbf{U_2}\,]$, where $\mathbf{U_1}$ is $K \times M$ matrix and $\mathbf{U_2}$ is $K \times (K-M)$ matrix. So

$$\begin{aligned}\mathbf{H} = \mathbf{U_s}\mathbf{Q}\mathbf{U}'_\mathbf{s} &= [\,\mathbf{U_1} \quad \mathbf{U_2}\,]\begin{bmatrix}\mathbf{\Lambda} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}[\,\mathbf{U_1} \quad \mathbf{U_2}\,]'\\ &= [\,\mathbf{U_1}\mathbf{\Lambda} \quad \mathbf{0}\,]\begin{bmatrix}\mathbf{U}'_\mathbf{1}\\ \mathbf{U}'_\mathbf{2}\end{bmatrix} = \mathbf{U_1}\mathbf{\Lambda}\mathbf{U}'_\mathbf{1},\end{aligned} \tag{B.1}$$

where

$$\mathbf{Q} = \begin{bmatrix}\mathbf{\Lambda} & 0\\ 0 & 0\end{bmatrix}, \tag{B.2}$$

$$\mathbf{U_s} = [\,\mathbf{U_1} \quad \mathbf{U_2}\,]. \tag{B.3}$$

References

[1] S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Transactions on Acoustics, Speech and Signal Processing 27 (1979) 113–120.

[2] J.H.L. Hansen, M.A. Clements, Constrained iterative speech enhancement with application to speech recognition, IEEE Transaction on Signal Processing 39 (1991) 795–805.

[3] Y. Ephraim, H.L. Van Trees, A signal subspace approach for speech enhancement, IEEE Transactions on Speech and Audio Processing 3 (4) (1995) 251–266.

[4] U. Mittal, N. Phamdo, Signal/noise KLT based approach for enhancing speech degraded by colored noise, IEEE Transactions on Speech and Audio Processing 8 (3) (2000) 159–167.

[5] F. Yang, Wavelet Transform for Engineering Analysis and Application, Science Press, Beijing, 2000.

[6] S. Mallat, A theory of multiresolution signal decomposition: the wavelet transform, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7) (1989) 674–693.

[7] S. Mallat, Multiresolution approximation and wavelet orthonormal bases of L2, Transaction of the American Mathematic Society 315 (1989) 69–87.

[8] L. Qin, G. Hu, C. Li, A new speech enhancement method, Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, May 2001, pp. 92–94.

[9] D.L. Donoho, De-noising by soft-thresholding, IEEE Transaction on Information Theory 41 (3) (1995) 613–627.