A SINGLE-CHANNEL NOISE CANCELATION FILTER IN THE SHORT-TIME-FOURIER-TRANSFORM DOMAIN

Xianghui Wang[1], *Jacob Benesty*[2], *and Jingdong Chen*[1]

[1] CIAIC and School of Marine Science and Technology
Northwestern Polytechnical University
Xi'an, Shaanxi 710072, China

[2]INRS-EMT, University of Quebec
800 de la Gauchetiere Ouest, Suite 6900
QC H5A 1K6, Canada

ABSTRACT

This paper develops a single-channel noise cancelation filter in the short-time Fourier transform (STFT) domain by combining the subspace method and the optimal filtering technique via joint diagonalization of the desired clean speech and noise signal correlation matrices. This filter is shown to be flexible in controlling the compromise between the output signal-to-noise ratio (oSNR) and the amount of speech distortion. Simulations are performed to justify the property of this filter.

***Index Terms*—**Noise cancelation, noise reduction, speech enhancement, STFT domain.

1. INTRODUCTION

Noise is ubiquitous. In real-world applications of speech processing such as hands-free voice communication, teleconferencing, hearing aids, smart phones, and VoIP, the existence of noise can cause significant degradation of speech quality and impairment of speech intelligibility if the signal-to-noise ratio (SNR) is low. To mitigate the noise effect, a widely used method is to pass the noisy speech signal through a filter that can dramatically suppress the unwanted noise while keeping the desired speech relatively unchanged. This filtering process is called noise reduction, which has been an important research topic in the field of speech processing [1–4]. Many efforts have been devoted to this area over the last few decades [1–18]; however, noise reduction remains an open problem primarily due to its extreme difficulty.

This paper is concerned with noise reduction using a single microphone. It develops a single-channel noise cancelation filter in the STFT domain. This approach combines the subspace method and the optimal filtering technique through the use of joint diagonalization of the clean speech and noise signal correlation matrices and is basically an extension of the method in [6, 7].

2. SIGNAL MODEL AND PROBLEM FORMULATION

The noise reduction problem considered in this paper is one of recovering the signal of interest (or clean speech) $x(t)$, t being the discrete-time index, from the noisy observation (microphone signal) [2], [8]:

$$y(t) = x(t) + v(t), \tag{1}$$

where $v(t)$ is the unwanted additive noise, which is assumed to be uncorrelated with $x(t)$. All signals are considered to be real, zero mean, and broadband.

In the STFT domain, the signal model in (1) can be rewritten as

$$Y(k,n) = X(k,n) + V(k,n), \tag{2}$$

This work was supported in part by the NSFC "Distinguished Young Scientists Fund" under grant No. 61425005.

where the zero-mean complex random variables $Y(k,n)$, $X(k,n)$, and $V(k,n)$ are the STFTs of $y(t)$, $x(t)$, and $v(t)$, respectively, at frequency bin $k \in \{0, 1, \ldots, K-1\}$ and time frame n. The variance of $Y(k,n)$ is

$$\phi_Y(k,n) = E\left[|Y(k,n)|^2\right] = \phi_X(k,n) + \phi_V(k,n), \tag{3}$$

where $\phi_X(k,n)$ and $\phi_V(k,n)$ are the variances of $X(k,n)$ and $V(k,n)$, respectively, which are defined in a similar way to $\phi_Y(k,n)$.

By considering the N most recent successive time frames of the observations as described in [9], we can rewrite (2) as

$$\begin{aligned}\mathbf{y}(k,n) &= [\, Y(k,n) \;\; Y(k,n-1) \;\; \cdots \;\; Y(k,n-N+1) \,]^T \\ &= \mathbf{x}(k,n) + \mathbf{v}(k,n),\end{aligned} \tag{4}$$

where $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$ are also vectors containing the N most recent successive time frames of the speech and noise signal, respectively. We deduce the correlation matrix of $\mathbf{y}(k,n)$ as

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k,n) &= E\left[\mathbf{y}(k,n)\mathbf{y}^H(k,n)\right] \\ &= \mathbf{\Phi}_{\mathbf{x}}(k,n) + \mathbf{\Phi}_{\mathbf{v}}(k,n),\end{aligned} \tag{5}$$

where the superscript ^{H} is the conjugate-transpose operator, and $\mathbf{\Phi}_{\mathbf{x}}(k,n) = E\left[\mathbf{x}(k,n)\mathbf{x}^H(k,n)\right]$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n) = E\left[\mathbf{v}(k,n)\mathbf{v}^H(k,n)\right]$ are the correlation matrices of $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$, respectively. The $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ matrix is assumed to be full rank. Then, the objective of noise reduction is to estimate $X(k,n)$ from $\mathbf{y}(k,n)$.

Using the well-known joint diagonalization technique [10], the two Hermitian matrices $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$ can be jointly diagonalized as follows:

$$\mathbf{B}^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{B}(k,n) = \mathbf{\Lambda}(k,n), \tag{6}$$

$$\mathbf{B}^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{B}(k,n) = \mathbf{I}_N, \tag{7}$$

where

$$\mathbf{\Lambda}(k,n) = \mathrm{diag}\left[\lambda_1(k,n), \lambda_2(k,n), \ldots, \lambda_N(k,n)\right] \tag{8}$$

and

$$\mathbf{B}(k,n) = [\, \mathbf{b}_1(k,n) \;\; \mathbf{b}_2(k,n) \;\; \cdots \;\; \mathbf{b}_N(k,n) \,] \tag{9}$$

are the eigenvalue and eigenvector matrices of $\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)$, and $\mathbf{I}_N$ is the identity matrix of size $N \times N$. In this paper, we assume that the eigenvalues $\lambda_1(k,n), \lambda_2(k,n), \ldots, \lambda_N(k,n)$ are ordered in such way that $0 \leq \lambda_1(k,n) \leq \lambda_2(k,n) \leq \cdots \leq \lambda_N(k,n)$. So, $\lambda_1(k,n)$ and $\lambda_N(k,n)$ are the smallest and largest eigenvalues of $\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)$, respectively. Note that the eigenvector matrix $\mathbf{B}(k,n)$ is full rank but not necessarily orthogonal.

3. LINEAR ESTIMATION AND PERFORMANCE MEASURES

First, let us estimate the noise at the STFT subband k by passing the noisy signal at that subband through a complex linear filter, i.e.,

$$Z(k,n) = \mathbf{h'}^H(k,n)\mathbf{y}(k,n) = X_{\rm fd}(k,n) + V_{\rm fn}(k,n), \quad (10)$$

where $Z(k,n)$ is an estimate of $V(k,n)$, $\mathbf{h}'(k,n)$ is a filter of length N, which is called the noise estimation filter, $X_{\rm fd}(k,n) = \mathbf{h'}^H(k,n)\mathbf{x}(k,n)$ is the filtered desired signal, and $V_{\rm fn}(k,n) = \mathbf{h'}^H(k,n)\mathbf{v}(k,n)$ is the filtered noise signal. It follows then that the variance of $Z(k,n)$ is

$$\phi_Z(k,n) = \phi_{X_{\rm fd}}(k,n) + \phi_{V_{\rm fn}}(k,n), \quad (11)$$

where $\phi_{X_{\rm fd}}(k,n) = \mathbf{h'}^H(k,n)\mathbf{\Phi_x}(k,n)\mathbf{h}'(k,n)$ and $\phi_{V_{\rm fn}}(k,n) = \mathbf{h'}^H(k,n)\mathbf{\Phi_v}(k,n)\mathbf{h}'(k,n)$.

Given the noise estimate, $Z(k,n)$, the estimate of the desired speech, $X(k,n)$, is obtained as

$$\widehat{X}(k,n) = Y(k,n) - Z(k,n) = \mathbf{h}^H(k,n)\mathbf{y}(k,n), \quad (12)$$

where

$$\mathbf{h}(k,n) = \mathbf{i} - \mathbf{h}'(k,n) \quad (13)$$

is the noise cancelation filter for the estimation of $X(k,n)$, with $\mathbf{i}$ being the first column of $\mathbf{I}_N$.

To assess the goodness of the noise cancelation filter, $\mathbf{h}(k,n)$, we adopt two performance metrics: the SNR and the speech distortion index. The subband input SNR at frequency bin k is defined as

$$\text{iSNR}(k,n) = \frac{\phi_X(k,n)}{\phi_V(k,n)}, \quad (14)$$

while the subband output SNR at frequency bin k is given by

$$\text{oSNR}\left[\mathbf{h}(k,n)\right] = \frac{\mathbf{h}^H(k,n)\mathbf{\Phi_x}(k,n)\mathbf{h}(k,n)}{\mathbf{h}^H(k,n)\mathbf{\Phi_v}(k,n)\mathbf{h}(k,n)}. \quad (15)$$

The subband speech-distortion index is defined as [1], [11]

$$\upsilon_{\rm sd}\left[\mathbf{h}(k,n)\right] = \frac{J_{\rm d}\left[\mathbf{h}(k,n)\right]}{\phi_X(k,n)}, \quad (16)$$

where

$$J_{\rm d}\left[\mathbf{h}(k,n)\right] = E\left[\left|X(k,n) - \mathbf{h}^H(k,n)\mathbf{x}(k,n)\right|^2\right]. \quad (17)$$

4. OPTIMAL NOISE ESTIMATION AND CANCELATION FILTERS

From the linear filtering model given in (10), we define the output SNR associated with the noise estimation filter $\mathbf{h}'(k,n)$ as

$$\begin{aligned}\text{oSNR}_Z\left[\mathbf{h}'(k,n)\right] &= \frac{\phi_{X_{\rm fd}}(k,n)}{\phi_{V_{\rm fn}}(k,n)} \\ &= \frac{\mathbf{h'}^H(k,n)\mathbf{\Phi_x}(k,n)\mathbf{h}'(k,n)}{\mathbf{h'}^H(k,n)\mathbf{\Phi_v}(k,n)\mathbf{h}'(k,n)}. \end{aligned} \quad (18)$$

It can be checked that

$$\lambda_1(k,n) \leq \text{oSNR}_Z\left[\mathbf{h}'(k,n)\right] \leq \lambda_N(k,n),\ \forall \mathbf{h}'(k,n). \quad (19)$$

The best way to estimate the noise signal is by minimizing $\text{oSNR}_Z\left[\mathbf{h}'(k,n)\right]$, since by doing so, the speech signal of interest left in $Z(k,n)$ is minimal.

Let the noise estimation filter, $\mathbf{h}'(k,n)$, be of the form:

$$\mathbf{h}'(k,n) = \sum_{p=1}^{P} \beta_p(k,n)\mathbf{b}_p(k,n) = \mathbf{B}_P(k,n)\boldsymbol{\beta}(k,n), \quad (20)$$

where

$$\mathbf{B}_P(k,n) = \left[\ \mathbf{b}_1(k,n)\ \ \mathbf{b}_2(k,n)\ \ \cdots\ \ \mathbf{b}_P(k,n)\ \right] \quad (21)$$

is a matrix of size $N \times P$ containing the eigenvectors corresponding to the eigenvalues $\lambda_p(k,n),\ p = 1, 2, \ldots, P$, and

$$\boldsymbol{\beta}(k,n) = \left[\ \beta_1(k,n)\ \ \beta_2(k,n)\ \ \cdots\ \ \beta_P(k,n)\ \right]^T \neq \mathbf{0} \quad (22)$$

is an arbitrary complex-valued vector of length P.

If $\lambda_1(k,n)$ is of multiplicity P[1], i.e., $\lambda_1(k,n) = \lambda_2(k,n) = \cdots = \lambda_P(k,n)$, it can be checked that the filter given in (20) minimizes $\text{oSNR}_Z\left[\mathbf{h}'(k,n)\right]$. As a matter of fact, we have

$$\begin{aligned}\text{oSNR}_Z\left[\mathbf{h}'(k,n)\right] &= \frac{\sum_{p=1}^{P}\lambda_p(k,n)\left|\beta_p(k,n)\right|^2}{\sum_{p=1}^{P}\left|\beta_p(k,n)\right|^2} \\ &= \lambda_1(k,n). \end{aligned} \quad (23)$$

So, the output SNR, $\text{oSNR}_Z\left[\mathbf{h}'(k,n)\right]$, is always minimized regardless of the value of the $\boldsymbol{\beta}(k,n)$ vector involved. However, the choice of $\boldsymbol{\beta}(k,n)$ plays an important role on the noise estimate, and consequently, the speech estimate. In order to find the optimal $\boldsymbol{\beta}(k,n)$ vector, let us define the MSE criterion between the desired and estimated signals:

$$J\left[\mathbf{h}(k,n)\right] = E\left[\left|X(k,n) - \mathbf{h}^H(k,n)\mathbf{y}(k,n)\right|^2\right]. \quad (24)$$

Using (17) and defining

$$J_{\rm r}\left[\mathbf{h}(k,n)\right] = E\left[\left|\mathbf{h}^H(k,n)\mathbf{v}(k,n)\right|^2\right], \quad (25)$$

we can rewrite (24) as

$$\begin{aligned}J\left[\mathbf{h}(k,n)\right] &= J_{\rm d}\left[\mathbf{h}(k,n)\right] + J_{\rm r}\left[\mathbf{h}(k,n)\right], \\ &= \phi_V(k,n) - \boldsymbol{\beta}^H(k,n)\mathbf{B}_P^H(k,n)\mathbf{\Phi_v}(k,n)\mathbf{i} - \\ &\quad \mathbf{i}^T\mathbf{\Phi_v}(k,n)\mathbf{B}_P(k,n)\boldsymbol{\beta}(k,n) + \boldsymbol{\beta}^H(k,n)\boldsymbol{\beta}(k,n) + \\ &\quad \boldsymbol{\beta}^H(k,n)\mathbf{\Lambda}_P(k,n)\boldsymbol{\beta}(k,n), \end{aligned} \quad (26)$$

where

$$\begin{aligned}\mathbf{\Lambda}_P(k,n) &= \text{diag}\left[\lambda_1(k,n), \lambda_2(k,n), \ldots, \lambda_P(k,n)\right] \\ &= \lambda_1(k,n)\mathbf{I}_P \end{aligned} \quad (27)$$

and $\mathbf{I}_P$ is the $P \times P$ identity matrix.

Now, the optimal $\boldsymbol{\beta}(k,n)$ vector can be found by minimizing $J\left[\mathbf{h}(k,n)\right]$, $J_{\rm d}\left[\mathbf{h}(k,n)\right]$, or $J_{\rm r}\left[\mathbf{h}(k,n)\right]$. For instance, minimizing $J_{\rm r}\left[\mathbf{h}(k,n)\right]$, we can find the optimal $\boldsymbol{\beta}(k,n)$ as

$$\begin{aligned}\boldsymbol{\beta}_{\rm o}(k,n) &= \mathbf{B}_P^H(k,n)\mathbf{\Phi_v}(k,n)\mathbf{i} \\ &= \mathbf{\Lambda}_P^{-1}(k,n)\mathbf{B}_P^H(k,n)\mathbf{\Phi_x}(k,n)\mathbf{i}, \end{aligned} \quad (28)$$

[1] In practice, we may consider the P smallest eigenvalues of $\mathbf{\Phi_v}^{-1}(k,n)\mathbf{\Phi_x}(k,n)$.

Substituting (28) into (20) and (13), we obtain the optimal noise estimation and cancelation filters, respectively,

$$\begin{aligned}\mathbf{h}'_{\mathrm{o}}(k,n) &= \mathbf{B}_P(k,n)\mathbf{B}_P^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{i} \\ &= \mathbf{B}_P(k,n)\mathbf{\Lambda}_P^{-1}(k,n)\mathbf{B}_P^H(k,n)\mathbf{\Phi}_{\mathbf{x}}(k,n)\mathbf{i},\end{aligned} \tag{29}$$

and

$$\mathbf{h}_{\mathrm{o}}(k,n) = \mathbf{i} - \mathbf{B}_P(k,n)\mathbf{B}_P^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{i}. \tag{30}$$

Note that if $P = N$, we have $\mathbf{h}'_{\mathrm{o}}(k,n) = \mathbf{i}$. In this case, $\mathbf{h}_{\mathrm{o}}(k,n) = \mathbf{0}$. So, the noise is completely nulled out.

5. PERFORMANCE ANALYSIS

It can be checked that

$$\left[\begin{array}{cc}\mathbf{B}_P(k,n) & \mathbf{B}_{N-P}(k,n)\end{array}\right]\left[\begin{array}{c}\mathbf{B}_P^H(k,n) \\ \mathbf{B}_{N-P}^H(k,n)\end{array}\right] = \mathbf{\Phi}_{\mathbf{v}}^{-1}(k,n), \tag{31}$$

where

$$\mathbf{B}_{N-P}(k,n) = \left[\begin{array}{cccc}\mathbf{b}_{P+1}(k,n) & \mathbf{b}_{P+2}(k,n) & \cdots & \mathbf{b}_N(k,n)\end{array}\right]. \tag{32}$$

Applying (31) to (30), one can obtain

$$\mathbf{h}_{\mathrm{o}}(k,n) = \mathbf{B}_{N-P}(k,n)\mathbf{B}_{N-P}^H(k,n)\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{i}. \tag{33}$$

Then, substituting (33) into (15), we can get the output SNR:

$$\mathrm{oSNR}\left[\mathbf{h}_{\mathrm{o}}(k,n)\right] = \frac{\sum_{p=P+1}^{N}\left|\mathbf{i}^H\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{b}_p(k,n)\right|^2\lambda_p(k,n)}{\sum_{p=P+1}^{N}\left|\mathbf{i}^H\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{b}_p(k,n)\right|^2}. \tag{34}$$

It is observed from (34) that the output SNR increases with P. When $P = N-1$, we obtain the maximum output SNR, which is $\lambda_N(k,n)$.

Substituting (30) into (16), one can write the subband speech-distortion index as

$$\upsilon_{\mathrm{sd}}\left[\mathbf{h}_{\mathrm{o}}(k,n)\right] = \frac{\sum_{p=1}^{P}\lambda_p(k,n)\left|\mathbf{i}^H\mathbf{\Phi}_{\mathbf{v}}(k,n)\mathbf{b}_p(k,n)\right|^2}{\phi_X(k,n)}. \tag{35}$$

It is clearly seen that the speech-distortion index is a monotonic non-decreasing function of P. Now, if the rank of the $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ matrix is equal to $N-P$, we have $\lambda_p(k,n) = 0, p = 1,2,\ldots,P$. So, we have $\upsilon_{\mathrm{sd}}\left[\mathbf{h}_{\mathrm{o}}(k,n)\right] = 0$. Consequently, the optimal filter $\mathbf{h}_{\mathrm{o}}(k,n)$ is also a distortionless filter in this case.

6. SIMULATIONS

In this section, we examine the performance of the optimal noise cancelation filter in (30) through simulations. The clean speech is taken from the TIMIT database [12]. Note that we only take all the speech signals from one male (MSVS0) and one female (FKSR0) speakers from that database. The sampling rate considered in this simulation is 8 kHz. So, the clean signals from the TIMIT database are downsampled from 16 kHz to 8 kHz. Noise is then added into the speech signal to control the input SNR. We consider two types of noise: white Gaussian and babble signal recorded in a New York Stock Exchange (NYSE) room.

To implement the noise cancelation filter, we divide the noisy signal into short-time frames with a frame length of 128 samples (16 ms). The overlapping between neighboring frames is 75%. A

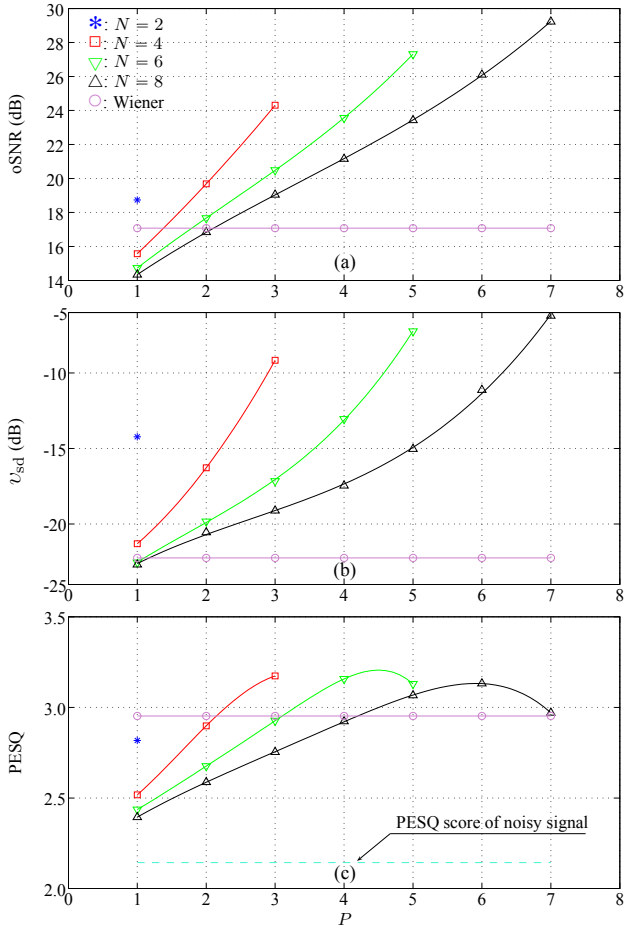

Fig. 1. Performance of $\mathbf{h}_{\mathrm{o}}(k,n)$ and the traditional Wiener filter in white noise as a function of P and N: (a) oSNR, (b) υ_{sd}, and (c) PESQ score. The input SNR is 10 dB.

128-point FFT is then used to transform every frame from time domain to STFT domain. In each STFT subband, the noise cancelation filter given in (30) is computed and applied to the noisy STFT coefficients. The overlap-add technique is finally used to transform the noise reduced STFT coefficients into the time domain.

To compute the optimal noise cancelation filter in (30), we need to know the two matrices $\mathbf{\Phi}_{\mathbf{x}}(k,n)$ and $\mathbf{\Phi}_{\mathbf{v}}(k,n)$. In practice, one has to apply a noise estimation algorithm to obtain the noise statistic characteristic. However, in order to avoid the estimate error of the noise statistic characteristic and focus on the noise reduction performance of the proposed filter, we assume the noise signal has been known in our simulations, and compute the correlation matrices directly from the noisy and noise signals using the following recursions:

$$\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n) = \alpha_y\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n-1) + (1-\alpha_y)\mathbf{y}(k,n)\mathbf{y}^H(k,n), \tag{36}$$

$$\widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,n) = \alpha_v\widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,n-1) + (1-\alpha_v)\mathbf{v}(k,n)\mathbf{v}^H(k,n), \tag{37}$$

where α_y and α_v are forgetting factors. We take $\alpha_y = \alpha_v$ for simplicity in our experiments. Then the clean speech correlation matrix $\widehat{\mathbf{\Phi}}_{\mathbf{x}}(k,n)$ can be obtained as $\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,n) - \widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,n)$. To make sure $\widehat{\mathbf{\Phi}}_{\mathbf{x}}(k,n)$ is positive semidefinite, we apply the eigenvalue decomposition to $\widehat{\mathbf{\Phi}}_{\mathbf{x}}(k,n)$ and force all the very small eigenvalues to zeros.

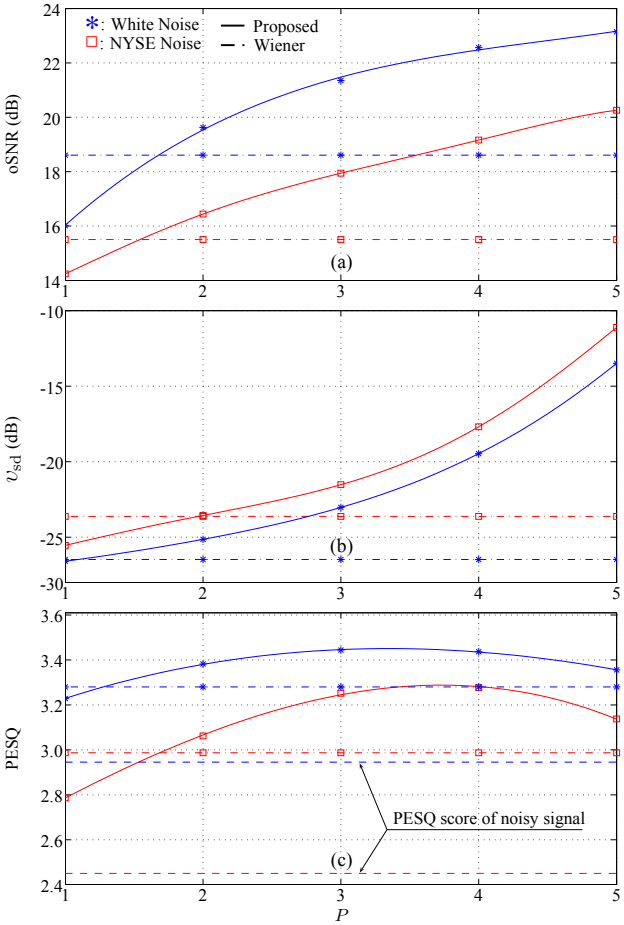

Fig. 2. Performance of $\mathbf{h}_{\rm o}(k,n)$ and the traditional Wiener filter in white and NYSE noises as a function of P: (a) oSNR, (b) $\upsilon_{\rm sd}$, and (c) PESQ score. The input SNR is 10 dB and $N = 6$.

For ease of performance presentation, we use the long-time fullband output SNR and speech-distortion index as the performance measures, which are defined as

$$\text{oSNR} = \frac{E[x_{\rm fd}^2(t)]}{E[v_{\rm rn}^2(t)]}, \tag{38}$$

$$\upsilon_{\rm sd} = \frac{E\left\{[x(t) - x_{\rm fd}(t)]^2\right\}}{E[x^2(t)]}, \tag{39}$$

where $x_{\rm fd}(t)$ and $v_{\rm rn}(t)$ are the time-domain counterparts of the enhanced speech and residual noise, respectively.

In the first simulation, we consider the case with white Gaussian noise. The performance as a function of the parameter P with different values of N is plotted in Fig. 1. For the purpose of comparison, we also plotted the results of the traditional Wiener gain, $H_{\rm W}(k,n) = \phi_X(k,n)/\phi_Y(k,n)$ [11], where both $\phi_X(k,n)$ and $\phi_Y(k,n)$ are computed in the same way as described previously.

It can be seen that both the output SNR and speech-distortion index with a specified value of N increase with P. This coincides with the theoretical analysis in Section 5. In contrast, the PESQ score does not bear a monotonic relationship with P when the value of N is large. For example, with $N = 8$, the PESQ score increases with P first and then decreases. It is also seen that the maximum PESQ score for $N = 8$ is smaller than that for $N = 6$. This indicates that the filter length N should not be too large. The underlying reason is that there is not much correlation between STFT coefficients from far-distant frames. Besides, the estimate error of the correlation matrices may increase with N, which eventually translates into performance degradation. For $N = 4$, 6 and 8, the noise cancelation filter $\mathbf{h}_{\rm o}(k,n)$ can achieve a PESQ score higher than that of the traditional Wiener filter if the value of P is properly chosen.

The second simulation is performed in a real reverberant office room. The reverberation time T_{60} of this room is approximately 0.24 s. A loudspeaker is placed in the room to play back some speech signals from the TIMIT database. A microphone is used to record the signal at a sampling rate of 8 kHz. To make the simulation repeatable and also for the ease of performance evaluation, the impulse response from the loudspeaker to the microphone is measured first. Convolution between the measured impulse response and the signals taken from the TIMIT database (again, from the speakers MSVS0 and FKSR0) is performed. This convolved speech is used as the "clean" speech in our simulation. Either white Gaussian or NYSE noise is then added to control the input SNR to be 10 dB. Based on the previous simulation, we set $N = 6$ and investigate the impact of P on the performance. Figure 2 plots the output SNR, speech-distortion index, and PESQ score, all as functions of P, of both the noise cancelation filter $\mathbf{h}_{\rm o}(k,n)$ and the traditional noise reduction Wiener gain. As seen, both the output SNR and the speech-distortion index of the noise cancelation filter $\mathbf{h}_{\rm o}(k,n)$ increase with P, which, again, coincides with the theoretical analysis. The PESQ score of the noise cancelation filter $\mathbf{h}_{\rm o}(k,n)$ depends on the value of P. When the value of P is properly chosen, the noise cancelation filter has a better PESQ score than the traditional Wiener gain.

7. CONCLUSIONS

This paper developed a noise cancelation approach to single-channel noise reduction in the STFT domain. It first obtains an estimate of the noise in each STFT subband and then subtracts the noise estimate from the noisy STFT coefficients to achieve noise reduction. To obtain a good estimate of the noise, an optimal noise estimation filter was developed, which basically combines the well-known subspace method and the optimal filtering technique via joint diagonalization of the clean speech and noise signal correlation matrices. The optimal noise cancelation filter is then constructed from the optimal noise estimation filter. Some theoretical analysis was provided to show the impact of the order of the noise subspace on the noise reduction performance. Simulations were performed in both an ideal and a real room environments. The results demonstrated that the optimal noise cancelation filter can yield larger output SNR and better PESQ score than the popularly used Wiener gain if the subspace parameter is properly chosen.

8. RELATION TO PRIOR WORK

In this paper, we developed a single-channel noise cancelation in the STFT domain, which is basically an extension of the method in [6,7] with more comprehensive theoretical analysis and more flexibility in controlling the tradeoff between the output SNR and speech distortion for better speech quality improvement. This method combines the subspace method [15, 16] and the optimal filtering technique [1, 2, 8] through the use of joint diagonalization [10, 17, 18] of the clean speech and noise correlation matrices. Because of the use of the joint diagonalization, this algorithm can deal with both white and colored noise. It can also achieve a distortionless estimate of the clean speech if the speech correlation matrix is rank-deficient. Simulation results showed that the developed algorithm can achieve a higher PESQ score than the widely used Wiener filter in most cases.

9. REFERENCES

[1] J. Benesty, S. Makino, J. Chen, Eds., *Speech Enhancement*. Berlin, Germany: Springer-Verlag, 2005.

[2] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[4] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[5] J. Chen, J. Benesty, Y. Huang, and E. J. Diethorn, "Fundamentals of Noise Reduction," in *Springer Handbook on Speech Processing and Speech Communication*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., Berlin: Springer-Verlag, 2007.

[6] S. M. Nørholm, J. Benesty, J. R. Jensen, and M. G. Christensen, "Single-channel noise reduction using unified joint diagonalization and optimal filtering," *EURASIP J. Advances Signal Process.*, 2014, 2014:37.

[7] S. M. Nørholm, J. Benesty, J. R. Jensen, and M. G. Christensen, "Noise reduction in the time domain using joint diagonalization," in *Proc. IEEE ICASSP*, 2014, pp. 7058–7062.

[8] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Berlin, Germany: Springer-Verlag, 2009.

[9] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 1256–1269, May 2012.

[10] J. N. Franklin, *Matrix Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1968.

[11] J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*. Berlin, Germany: Springer Briefs in Electrical and Computer Engineering, 2011.

[12] J. Garofolo, *DARPA TIMIT acoustic-phonetic continuous speech corpus*. Gaithersburg, MD, USA: Nat. Inst. of Standards Technol., 1993.

[13] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, and R. C. Goodlin, "Adaptive noise canceling: principles and applications," *Proc. IEEE*, vol. 63, pp. 1692–1716, Dec. 1975.

[14] J. Chen, Y. Huang, and J. Benesty, "Filtering techniques for noise reduction and speech enhancement," in *Adaptive Signal Processing: Application to Real-World Problems*, J. Benesty and Y. Huang, Eds., Berlin: Springer-Verlag, 2003.

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

[16] P. C. Hansen and S. H. Jensen, "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: survey and analysis," *EURASIP J. Advances Signal Process.*, vol. 2007, p. 24, 2007.

[17] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, Jul. 2003.

[18] J. Benesty and J. Chen, *Optimal Time-domain Noise Reduction Filters-A Theoretical Study*. Berlin, Germany: Springer Briefs in Electrical and Computer Engineering, 2011.