SINGLE-CHANNEL NOISE REDUCTION IN THE STFT DOMAIN BASED ON THE BIFREQUENCY SPECTRUM

Jingdong Chen[1] *and Jacob Benesty*[2]

[1]: Northwestern Polytechnical University
127 Youyi West Road
Xi'an, Shaanxi 710072, China

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

ABSTRACT

This paper studies the problem of noise reduction in the short-time Fourier transform (STFT) domain. Traditionally, the STFT coefficients in different frequency bands are assumed to be independent. This assumption holds when the signals are stationary and the fast Fourier transform (FFT) length is sufficiently large. In practice, however, speech is nonstationary and also the FFT length cannot be very large due to practical reasons. So, there always exists some correlation between STFT coefficients from neighboring frequency bands. An important question then arises: how the interband correlation can be used to optimize noise reduction performance? This paper addresses this issue. We discuss two solutions in the framework of the bifrequency spectrum. One considers the cross-correlation between all the frequency bands and the other takes into account only the cross-correlation between neighboring bands. While the former is optimal from a theoretical perspective, the latter is more practical as it is more immune to the error in correlation matrix estimation.

Index Terms— Single-channel noise reduction, bifrequency spectrum, interband correlation, Wiener filter, tradeoff filter.

1. INTRODUCTION

Noise reduction is an important problem and has a wide range of applications in areas such as cellular phones, hands-free communication, teleconferencing, hearing aids, human-machine interfaces, etc. Mathematically, this problem can be described as one of recovering the desired signal (or clean speech) $x(t)$, t being the time index, of zero mean from the noisy observation (microphone signal)

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

where $v(t)$ is the unwanted zero-mean additive noise. The three signals $x(t)$, $v(t)$, and $y(t)$ are real and generally assumed to be broadband. The noise process $v(t)$ can be either white or colored but it is assumed to be uncorrelated with $x(t)$.

Using the short-time Fourier transform (STFT), (1) can be rewritten in the frequency domain as

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

where the zero-mean complex random variables $Y(k,m)$, $X(k,m)$, and $V(k,m)$ 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 m. With the signal model given in (2), the objective of noise reduction is to estimate the desired signal, $X(k,m)$, from the noisy observation, $Y(k,m)$. Traditionally, the STFT components from different frequency bins are assumed to be uncorrelated. So, the estimation of $X(k,m)$ is achieved by applying a complex gain in each subband to the noisy spectrum $Y(k,m)$. This assumption is true when the signals are stationary and the FFT length is sufficiently large. However, it is well known that speech is nonstationary. Also, the FFT length cannot be set too large in reality due to practical reasons. For example, in telecommunications, the maximum allowed delay for each processor is limited, and for noise reduction, the delay should not exceed 20 ms in most cases. With nonstationary speech and small FFT lengths, the STFT coefficients from neighboring frequency bins exhibit some correlation. To illustrate this, we recorded a 1-second long speech signal in a quiet office room with a sampling rate of 8 kHz, as shown in Fig. 1. We divided this signal into overlap frames with a frame length of 16 ms and 75% overlap. Each frame is transformed into the STFT domain using a 128-point FFT. We then computed the normalized cross-correlation coefficients [1] between different frequency bins. Figure 2 plots the results for the 4th, 8th, and 16th bins. It is clearly seen that there is a strong correlation between frequency bins that are next to each other. A legitimate question one would ask: how do we use the interband correlation information to improve noise reduction? This question will be answered in the following sections.

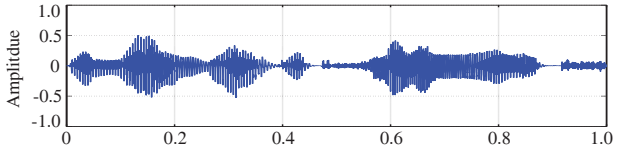

Fig. 1. A clean speech signal sampled at 8 kHz.

2. THE BIFREQUENCY SPECTRUM

Before discussing how to use the interband correlation, we first introduce the term bifrequency spectrum. Let $a(t)$ be a zero-mean real random variable for which its frequency-domain representation is $A(k,m)$. We define the bifrequency spectrum as [2], [3]

$$\phi_A(k_1,k_2,m) = E\left[A(k_1,m)A^*(k_2,m)\right], \tag{3}$$

where k_1 and k_2 are possibly two different frequency bins. Basically, the bifrequency spectrum is a measure of the correlation between two different frequencies of the same signal. If $a(t)$ is a wide-sense stationary signal and if we use a long FFT length to compute $A(k,m)$, the bifrequency spectrum reduces to

$$\phi_A(k_1,k_2,m) = \begin{cases} \phi_A(k,m), & k_1 = k_2 = k \\ 0, & k_1 \neq k_2 \end{cases}, \tag{4}$$

where $\phi_A(k,m) = \phi_A(k,k,m)$. Thus, for a stationary random process, the Fourier coefficients from two different bands are uncorrelated. However, for a nonstationary random process like speech, the bifrequency spectrum will exhibit non-zero correlations along the so-called support curves other than the main diagonal $k_1 = k_2$ as we showed in the previous section. It seems then appropriate when deriving noise reduction algorithms in the STFT domain to take into account the spectral correlation that may not be negligible in this context.

Effort of the first author is partially supported by the Anhui Science and Technology Project (11010202191).

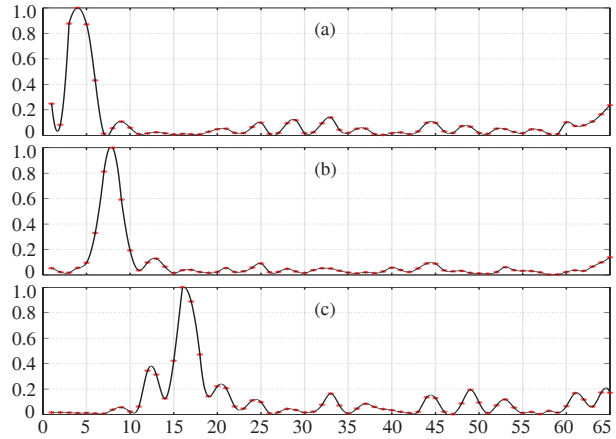

Fig. 2. The magnitude of the cross-correlation coefficients between: (a) the 4th and other frequency bins, (b) the 8th and other bins, (c) the 16th and other bins. The sampling rate is 8 kHz, the frame length is 16 ms (128 points), the FFT length is 128, and the overlapp is 75%.

3. THE OPTIMAL APPROACH

Let us first consider to use the correlation among all the frequency bins. To do so, we concatenate the desired signal at all frequency bins in a vector of length K:

$$\mathbf{x}(m) = [X(0,m) \quad X(1,m) \quad \cdots \quad X(K-1,m)]^T, \tag{5}$$

where the superscript ^{T} denotes transpose. We can then estimate $\mathbf{x}(m)$ with

$$\mathbf{z}(m) = \mathbf{H}(m)\mathbf{y}(m), \tag{6}$$

where

$$\mathbf{H}(m) = \begin{bmatrix} \mathbf{h}_0^H(m) \\ \mathbf{h}_1^H(m) \\ \vdots \\ \mathbf{h}_{K-1}^H(m) \end{bmatrix} \tag{7}$$

is a square filtering matrix of size $K \times K$, the superscript ^{H} denotes transpose-conjugate, $\mathbf{h}_k(m)$, $k = 0, 1, \ldots, K-1$ are FIR filters of length K, and $\mathbf{y}(m)$ is defined in a similar way to $\mathbf{x}(m)$. We can rewrite (6) as

$$\mathbf{z}(m) = \mathbf{H}(m)\left[\mathbf{x}(m) + \mathbf{v}(m)\right] = \mathbf{x}_{\mathrm{fd}}(m) + \mathbf{v}_{\mathrm{rn}}(m), \tag{8}$$

where $\mathbf{v}(m)$ is defined similarly to $\mathbf{x}(m)$,

$$\mathbf{x}_{\mathrm{fd}}(m) = \mathbf{H}(m)\mathbf{x}(m) \tag{9}$$

is the filtered desired signal vector, and

$$\mathbf{v}_{\mathrm{rn}}(m) = \mathbf{H}(m)\mathbf{v}(m) \tag{10}$$

is the residual noise signal vector.

The correlation matrix of $\mathbf{z}(m)$ is then

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{z}}(m) &= E\left[\mathbf{z}(m)\mathbf{z}^H(m)\right] = \mathbf{H}(m)\mathbf{\Phi}_{\mathbf{y}}(m)\mathbf{H}^H(m) \\ &= \mathbf{H}(m)\mathbf{\Phi}_{\mathbf{x}}(m)\mathbf{H}^H(m) + \mathbf{H}(m)\mathbf{\Phi}_{\mathbf{v}}(m)\mathbf{H}^H(m),\end{aligned} \tag{11}$$

where

$$\mathbf{\Phi}_{\mathbf{y}}(m) = \begin{bmatrix} \phi_Y(0,m) & \phi_Y(0,1,m) & \cdots & \phi_Y(0,K-1,m) \\ \phi_Y(1,0,m) & \phi_Y(1,m) & \cdots & \phi_Y(1,K-1,m) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_Y(K-1,0,m) & \phi_Y(K-1,1,m) & \cdots & \phi_Y(K-1,m) \end{bmatrix},$$

and $\mathbf{\Phi}_{\mathbf{x}}(m)$ and $\mathbf{\Phi}_{\mathbf{v}}(m)$ are the correlation matrices of $\mathbf{x}(m)$ and $\mathbf{v}(m)$, respectively. We see now that the spectral correlation is taken into account in the estimator $\mathbf{z}(m)$. If the spectral correlation is negligible for both the speech and noise, then all three correlation matrices $\mathbf{\Phi}_{\mathbf{y}}(m)$, $\mathbf{\Phi}_{\mathbf{x}}(m)$, and $\mathbf{\Phi}_{\mathbf{v}}(m)$ are diagonal and this approach is identical to the noise reduction with a gain as explained in [1].

The error signal vector between the estimated and desired signals is

$$\mathbf{e}(m) = \mathbf{z}(m) - \mathbf{x}(m) = \mathbf{H}(m)\mathbf{y}(m) - \mathbf{x}(m), \tag{12}$$

which can also be written as the sum of two orthogonal error signal vectors:

$$\mathbf{e}(m) = \mathbf{e}_{\mathrm{d}}(m) + \mathbf{e}_{\mathrm{r}}(m), \tag{13}$$

where

$$\mathbf{e}_{\mathrm{d}}(m) = \left[\mathbf{H}(m) - \mathbf{I}_K\right]\mathbf{x}(m) \tag{14}$$

is the speech distortion due to the filtering matrix, $\mathbf{I}_K$ is the identity matrix of size $K \times K$, and

$$\mathbf{e}_{\mathrm{r}}(m) = \mathbf{H}(m)\mathbf{v}(m) \tag{15}$$

represents the residual noise.

Having defined the error signal, we can now write the fullband MSE criterion:

$$\begin{aligned}J\left[\mathbf{H}(m)\right] &= \mathrm{tr}\left\{E\left[\mathbf{e}(m)\mathbf{e}^H(m)\right]\right\} \\ &= \mathrm{tr}\left[\mathbf{\Phi}_{\mathbf{x}}(m)\right] - \mathrm{tr}\left[\mathbf{H}(m)\mathbf{\Phi}_{\mathbf{y}}(m)\mathbf{H}^H(m)\right] \\ &\quad -\mathrm{tr}\left[\mathbf{H}(m)\mathbf{\Phi}_{\mathbf{x}}(m)\right] - \mathrm{tr}\left[\mathbf{\Phi}_{\mathbf{x}}(m)\mathbf{H}^H(m)\right],\end{aligned} \tag{16}$$

where tr[·] denotes the trace of a square matrix. Using the fact that $E\left[\mathbf{e}_{\mathrm{d}}(m)\mathbf{e}_{\mathrm{r}}^H(m)\right] = \mathbf{0}_{K\times K}$, $J\left[\mathbf{H}(m)\right]$ can be expressed as the sum of two other fullband MSEs, i.e.,

$$\begin{aligned}J\left[\mathbf{H}(m)\right] &= \mathrm{tr}\left\{E\left[\mathbf{e}_{\mathrm{d}}(m)\mathbf{e}_{\mathrm{d}}^H(m)\right]\right\} + \mathrm{tr}\left\{E\left[\mathbf{e}_{\mathrm{r}}(m)\mathbf{e}_{\mathrm{r}}^H(m)\right]\right\} \\ &= J_{\mathrm{d}}\left[\mathbf{H}(m)\right] + J_{\mathrm{r}}\left[\mathbf{H}(m)\right].\end{aligned} \tag{17}$$

Given the above MSEs, we can now derive important optimal filtering matrices.

3.1. Wiener

If we differentiate the fullband MSE criterion, $J\left[\mathbf{H}(m)\right]$, with respect to $\mathbf{H}(m)$ and equate the result to zero, we find the Wiener filtering matrix

$$\mathbf{H}_{\mathrm{W}}(m) = \mathbf{\Phi}_{\mathbf{x}}(m)\mathbf{\Phi}_{\mathbf{y}}^{-1}(m) = \mathbf{I}_K - \mathbf{\Phi}_{\mathbf{v}}(m)\mathbf{\Phi}_{\mathbf{y}}^{-1}(m), \tag{18}$$

which is identical to the one derived in [4], [5]. This matrix depends only on the second-order statistics of the noisy and noise signals.

If the spectral correlation of the signals can be neglected, then $\mathbf{H}_{\mathrm{W}}(m)$ is a diagonal matrix whose components are the Wiener gains [1], [6].

3.2. Tradeoff

In the tradeoff approach, we minimize the MSE of speech distortion with the constraint that the residual noise level is smaller than that of the noise in the original noisy signal. Mathematically, this is equivalent to [6]

$$\min_{\mathbf{H}(m)} J_{\mathrm{d}}\left[\mathbf{H}(m)\right] \quad \text{subject to} \quad J_{\mathrm{r}}\left[\mathbf{H}(m)\right] = \beta \mathrm{tr}\left[\mathbf{\Phi}_{\mathbf{v}}(m)\right], \quad (19)$$

where $0 < \beta < 1$ to insure that we get some noise reduction. By using a Lagrange multiplier, $\mu \geq 0$, to adjoin the constraint to the cost function and assuming that the matrix $\mathbf{\Phi}_{\mathbf{x}}(m) + \mu\mathbf{\Phi}_{\mathbf{v}}(m)$ is invertible, we easily deduce the tradeoff filtering matrix

$$\mathbf{H}_{\mathrm{T},\mu}(m) = \mathbf{\Phi}_{\mathbf{x}}(m)\left[\mathbf{\Phi}_{\mathbf{x}}(m) + \mu\mathbf{\Phi}_{\mathbf{v}}(m)\right]^{-1}. \quad (20)$$

For $\mu = 1$, we get the Wiener filtering matrix. For $\mu = 0$, we see that $\mathbf{H}_{\mathrm{T},0}(m) = \mathbf{I}_K$. For μ greater or smaller than 1, we obtain a filtering matrix that reduces more or less noise than the Wiener filtering matrix.

4. A SUBOPTIMAL APPROACH

We discussed in the previous section two optimal filtering matrices that exploit the correlation among all the frequency bins. While they are optimal from a theoretical perspective, these filters have some practical drawbacks. First, to implement these filters, we need to compute the inverse of a $K \times K$ matrix for each time frame, which is computationally very expensive as K is usually large. Second, we need a large number of signal frames ($> K$) to estimate the correlation matrices $\mathbf{\Phi}_{\mathbf{y}}(m)$, $\mathbf{\Phi}_{\mathbf{v}}(m)$, and $\mathbf{\Phi}_{\mathbf{x}}(m)$; otherwise, they would be either rank deficient or ill conditioned. When a large number of frames are used, the estimates of these matrices would not follow the true statistics of the nonstationary speech signal, causing degradation in noise-reduction performance. As shown in Section 1, correlation only exists between neighboring frequency bins and there is not much correlation between distant bins. Given this, we discuss a suboptimal yet more practical approach in this section that considers correlation between only neighboring frequency bins. So, instead of estimating $\mathbf{x}(m)$ from the noisy vector $\mathbf{y}(m)$, we now estimate $X(k,m)$ on a sample basis, i.e.,

$$Z(k,m) = \mathbf{h}_k'^H(m)\mathbf{y}_k(m), \quad (21)$$

where

$$\mathbf{y}_k(m) = \big[Y(k-K_k^-,m) \quad \cdots \quad Y(k-1,m) \quad Y(k,m) \quad Y(k+1,m) \quad \cdots \quad Y(k+K_k^+,m)\big]^T \quad (22)$$

is a vector of length $K_k^- + K_k^+ + 1 \ll K$, K_k^- and K_k^+ are, respectively, the numbers of samples before and after the kth bin that are used to estimate $X(k,m)$, and $\mathbf{h}_k'(m)$ is an FIR filter of length $K_k^- + K_k^+ + 1$.

Apparently, we can rewrite (21) as

$$\begin{aligned} Z(k,m) &= \mathbf{h}_k'^H(m)\left[\mathbf{x}_k(m) + \mathbf{v}_k(m)\right] \\ &= X_{\mathrm{fd}}(k,m) + V_{\mathrm{rn}}(k,m), \end{aligned} \quad (23)$$

where $\mathbf{x}_k(m)$ and $\mathbf{v}_k(m)$ are defined similarly to $\mathbf{y}_k(m)$,

$$X_{\mathrm{fd}}(k,m) = \mathbf{h}_k'^H(m)\mathbf{x}_k(m) \quad (24)$$

is the filtered desired signal at the frequency-bin k and time-frame m, and

$$V_{\mathrm{rn}}(k,m) = \mathbf{h}_k'^H(m)\mathbf{v}_k(m) \quad (25)$$

is the residual noise.

It is easy to check that the two terms $X_{\mathrm{fd}}(k,m)$ and $V_{\mathrm{rn}}(k,m)$ are uncorrelated. So, the variance of $Z(k,m)$ is

$$\phi_Z(k,m) = \phi_{X_{\mathrm{fd}}}(k,m) + \phi_{V_{\mathrm{rn}}}(k,m), \quad (26)$$

where

$$\phi_{X_{\mathrm{fd}}}(k,m) = E\left[|X_{\mathrm{fd}}(k,m)|^2\right] = \mathbf{h}_k'^H(m)\mathbf{\Phi}_{\mathbf{x}_k}(m)\mathbf{h}_k'(m), \quad (27)$$

$$\phi_{V_{\mathrm{rn}}}(k,m) = \mathbf{h}_k'^H(m)\mathbf{\Phi}_{\mathbf{v}_k}(m)\mathbf{h}_k'(m), \quad (28)$$

and $\mathbf{\Phi}_{\mathbf{x}_k}(m)$ and $\mathbf{\Phi}_{\mathbf{v}_k}(m)$ are the correlation matrices of $\mathbf{x}_k(m)$ and $\mathbf{v}_k(m)$, respectively.

Following the same line of ideas of the previous section, we can define the error signal between the estimated and desired signals as

$$\begin{aligned} e(k,m) &= Z(k,m) - X(k,m) \\ &= e_{\mathrm{d}}(k,m) + e_{\mathrm{r}}(k,m), \end{aligned} \quad (29)$$

where

$$e_{\mathrm{d}}(k,m) = X_{\mathrm{fd}}(k,m) - X(k,m) \quad (30)$$

is the speech distortion due to the filter and

$$e_{\mathrm{r}}(k,m) = V_{\mathrm{rn}}(k,m) \quad (31)$$

represents the residual noise.

Given the above error signal, we can now write the narrowband MSE criterion:

$$J\left[\mathbf{h}_k'(m)\right] = E\left[|e(k,m)|^2\right] = J_{\mathrm{d}}\left[\mathbf{h}_k'(m)\right] + J_{\mathrm{r}}\left[\mathbf{h}_k'(m)\right], \quad (32)$$

where

$$J_{\mathrm{d}}\left[\mathbf{h}_k'(m)\right] = E\left[|e_{\mathrm{d}}(k,m)|^2\right] \quad (33)$$

and

$$J_{\mathrm{r}}\left[\mathbf{h}_k'(m)\right] = E\left[|e_{\mathrm{r}}(k,m)|^2\right] = \phi_{V_{\mathrm{rn}}}(k,m). \quad (34)$$

4.1. Wiener

The Wiener filter can be derived by taking the gradient of the narrowband MSE, $J\left[\mathbf{h}_k'(m)\right]$, with respect to $\mathbf{h}_k'(m)$ and equating the result to zero:

$$\begin{aligned} \mathbf{h}_{k,\mathrm{W}}'(m) &= \mathbf{\Phi}_{\mathbf{y}_k}^{-1}(m)\mathbf{\Phi}_{\mathbf{x}_k}(m)\mathbf{i}_{K_k^-+1} \\ &= \left[\mathbf{I} - \mathbf{\Phi}_{\mathbf{y}_k}^{-1}(m)\mathbf{\Phi}_{\mathbf{v}_k}(m)\right]\mathbf{i}_{K_k^-+1}, \end{aligned} \quad (35)$$

where $\mathbf{\Phi}_{\mathbf{y}_k}(m)$ is the correlation matrix of $\mathbf{y}_k(m)$, $\mathbf{I}$ is the identity matrix of size $(K_k^- + K_k^+ + 1) \times (K_k^- + K_k^+ + 1)$, and $\mathbf{i}_{K_k^-+1}$ is the $(K_k^- + 1)$th column of $\mathbf{I}$.

4.2. Tradeoff

Similarly to the optimal approach in Section 3, we can obtain a tradeoff filter by minimizing the MSE of speech distortion with the constraint that the residual noise level is smaller than that of the noise in the original noisy signal. Mathematically, this can be written as

$$\min_{\mathbf{h}_k'(m)} J_{\mathrm{d}}\left[\mathbf{h}_k'(m)\right] \text{ subject to } J_{\mathrm{r}}\left[\mathbf{h}_k'(m)\right] = \beta_k\phi_V(k,m), \quad (36)$$

where $0 < \beta_k < 1$ is a frequency-dependent constant to insure that we get some noise reduction. By using a Lagrange multiplier, $\mu_k \geq 0$, to adjoin the constraint to the cost function and assuming that the matrix $\mathbf{\Phi}_{\mathbf{x}_k} + \mu_k\mathbf{\Phi}_{\mathbf{v}_k}(m)$ is invertible, we can deduce the tradeoff filter

$$\mathbf{h}_{k,\mathrm{T},\mu_k}'(m) = \left[\mathbf{\Phi}_{\mathbf{x}_k}(m) + \mu_k\mathbf{\Phi}_{\mathbf{v}_k}(m)\right]^{-1}\mathbf{\Phi}_{\mathbf{x}_k}(m)\mathbf{i}_{K_k^-+1}. \quad (37)$$

If $\mu_k = 1$, we get the Wiener filter. When $\mu_k = 0$, we see that $\mathbf{h}_{k,\mathrm{T},\mu_k}'(m) = \mathbf{i}_{K_k^-+1}$. For μ_k greater or smaller than 1, we obtain a filter that reduces more or less noise than the Wiener filter.

5. EXPERIMENTAL RESULTS

In this section, we study the effect of using interband correlation on noise reduction performance through experiments. Due to space limit, we only present the results of the Wiener filter given in (35). The clean speech signal used in the experiments was recorded from a female speaker in a quiet office room. It was sampled at 8 kHz. The overall length of the signal is 30 seconds. The noisy speech is obtained by adding a white Gaussian noise signal to the clean speech and the noise signal is properly scaled to control the input SNR to 10 dB. We divide the signals into overlap frames with a frame length of 16 ms and 75% overlap. Each frame is transformed into the STFT domain using a 128-point FFT.

Implementation of the Wiener filter given in (35) requires estimation of the correlation matrices $\mathbf{\Phi}_{\mathbf{y}_k}(m)$, $\mathbf{\Phi}_{\mathbf{v}_k}(m)$, and $\mathbf{\Phi}_{\mathbf{x}_k}(m)$. Computation of the matrix $\mathbf{\Phi}_{\mathbf{y}_k}(m)$ is relatively easy as the noisy signal spectrum $Y(k,m)$ is accessible. But we would need a noise estimator or voice activity detector (VAD) to compute the other two matrices in practice. However, in this paper, we will set aside the noise estimation issue and only focus on illustrating the effect of using interband correlation on noise reduction performance. So, we will not use any noise estimator in our experiments. Instead, we compute all the correlation matrices directly from the corresponding signal using a short time average.

In the first experiment, we set $K_k^- = K_k^+ = K_k$ and study the performance of the Wiener filter given in (35) as a function of K_k (note that the first and last a few bins are properly processed with the available bins before or after the kth bin). We compute the correlation matrices $\mathbf{\Phi}_{\mathbf{y}_k}(m)$, $\mathbf{\Phi}_{\mathbf{x}_k}(m)$, and $\mathbf{\Phi}_{\mathbf{v}_k}(m)$ from the corresponding signals using the most recent 100 spectrum samples (i.e., 100 frames). The inverse of the $\mathbf{\Phi}_{\mathbf{y}_k}(m)$ is computed using the eigenvalue decomposition based approach where all the non-positive eigenvalues (this can happen when K_k is large) of $\mathbf{\Phi}_{\mathbf{y}_k}(m)$ are set to zero during the inverse process. We use the output SNR to evaluate the performance of noise reduction and the Itakura-Saito distance (ISD) between the clean and filtered speech to evaluate the amount of speech distortion. These two measures are computed in a global manner, i.e., we use overlap-add technique to reconstruct the time-domain enhanced and filtered speech signals. The output SNR is computed as the ratio between the energy of the filtered speech and that of the residual noise. The ISD is computed between the clean and filtered speech. The result is plotted in Fig. 3. It is clearly seen that using the interband correlation can help improve noise reduction as the performance of the Wiener filter first increases (larger output SNR and smaller ISD value) with K_k. But when it is larger than roughly 10, further increasing K_k will lead to performance degradation. This degradation is mainly caused by the numerical issue as the estimation error of the noisy correlation matrix $\mathbf{\Phi}_{\mathbf{y}_k}(m)$ increases with K_k given the fixed number of samples (it is 100 in our experiment). Taking into account both the SNR improvement and speech distortion, the best performance is obtained with inclusion of only a few neighboring frequency bins. This result, on the one hand, validates the observation made in Section 1, and on the other hand, shows the advantage of the Wiener filter given in (35) over that given in (18) as (18) can be viewed as an extreme case of (35) when K_k^- and K_k^+ reach their maximum value to $K_k^- + K_k^+ + 1 = K$.

One can notice that the speech distortion in Fig. 3 is large (the ISD values are large). This is mainly due to the fact that a large number of frames (100) are used to compute the correlation matrices to ensure that the estimated $\mathbf{\Phi}_{\mathbf{y}_k}(m)$ matrix is full rank for different values of K_k. With the use of such a large number of frames, the estimated correlation matrices cannot follow the time-varying statistics of the speech signal. In the second experiment, we fix K_k to 2 and vary the number of frames used to compute the correlation matrices. The result is sketched in Fig. 4. Clearly, we see that if too many frames are used to compute the statistics, the output SNR decreases and the ISD value increases. This performance degradation again demonstrates the advantage of the optimal filters deduced in Section 4 over those derived in Section 3 for practical usage since it is easier to estimate the time-varying correlation matrices for the filters in Section 4 with a small K_k value.

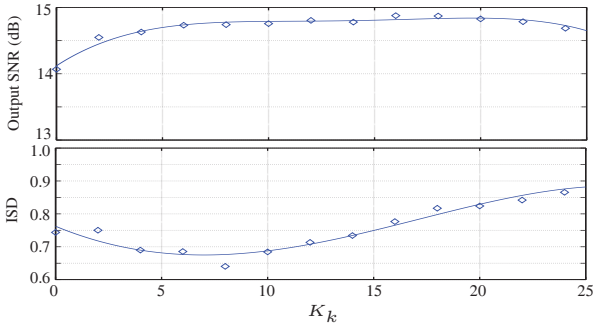

Fig. 3. The output SNR and ISD between the clean and filtered speech of the Wiener filter as a function of K_k ($K_k^- = K_k^+ = K_k$).

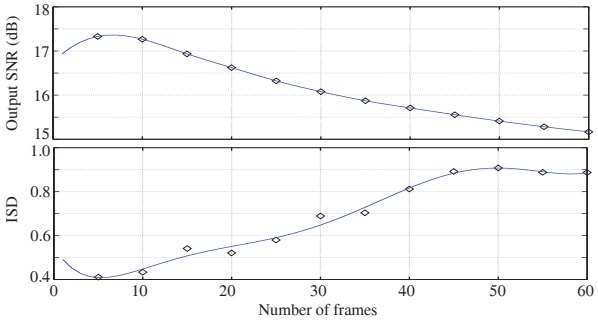

Fig. 4. The output SNR and ISD between the clean and filtered speech of the Wiener filter as a function of the number of frames used to compute the correlation matrices.

6. CONCLUSIONS

This paper studied the problem of noise reduction in the STFT domain. Unlike most traditional approaches that assume that STFT coefficients in different frequency bins are independent, we considered the interband correlation in deriving the noise reduction filters. Particularly, we discussed two approaches: one considers the cross-correlation between all the frequency bins and the other takes into account only the cross-correlation between neighboring bins. While the former is optimal from a theoretical viewpoint, the latter is more practical as we demonstrated that the time-varying correlation matrices involved in this category have a much lower dimensionality, and therefore, are easier to estimate.

7. REFERENCES

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

[2] N. L. Gerr and J. C. Allen, "The generalized spectrum and spectral coherence of a harmonizable time series," *Digit. Signal Process.*, vol. 4, pp. 222–238, 1994.

[3] A. Napolitano, "Uncertainty in measurements on spectrally correlated stochastic processes," *IEEE Trans. Signal Process.*, vol. 49, pp. 2172–2191, Sept. 2003.

[4] C. Li and S. V. Andersen, "A block-based linear MMSE noise reduction with a high temporal resolution modeling of the speech excitation," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 18, pp. 2965–2978, 2005.

[5] E. Plourde and B. Champagne, "Multidomentional STSA estimators for speech enhancement with correlated spectral components," *IEEE Trans. Signal Process.*, vol. 59, pp.3013–3024, July 2011.

[6] J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*. Berlin, Germany: Springer-Verlag, 2011.