Digital Signal Processing 33 (2014) 169–179

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

A minimum variance distortionless response filter based on the bifrequency spectrum for single-channel noise reduction ☆

Hai Huang [a,*], Liheng Zhao [b], Jingdong Chen [a], Jacob Benesty [b]

[a] Center of Immersive and Intelligent Acoustics, Northwestern Polytechnical University, 127 Youyi West Road, Xi'an, Shaanxi 710072, China
[b] INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Suite 6900, Montreal, QC H5A 1K6, Canada

ARTICLE INFO

Article history:
Available online 1 July 2014

Keywords:
Single-channel noise reduction
Speech enhancement
Minimum variance distortionless response (MVDR) filter
Bifrequency spectrum
Interband correlation

ABSTRACT

This paper deals with the problem of single-channel noise reduction in the short-time Fourier transform (STFT) domain. Many algorithms have been developed to solve this important problem, most of which generally assume that the STFT coefficients in different frequency bands are uncorrelated, so the noise reduction is achieved by applying a gain function to the STFT of the noisy speech in each frequency band. However, this assumption is not accurate and the STFT coefficients of speech signals between neighboring frequency bands are correlated in practice due to the use of small lengths of the fast Fourier transform (FFT) and overlap add/save techniques in implementation. This paper formulates the noise reduction problem by taking into account the interband correlation using the so-called bifrequency spectrum. Based on this formulation, a single-channel minimum variance distortionless response (MVDR) filter is derived, which is shown to be able to significantly improve the signal-to-noise ratio (SNR) and meanwhile maintain the desired speech not much distorted. Simulations are presented to justify the claimed merits of the developed MVDR filter.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Noise reduction, a term also used interchangeably with speech enhancement, refers to the problem of recovering a speech signal of interest from the microphone observations contaminated by some additive noise. This problem has long been one of the major focuses in acoustic signal processing for voice communications and significant efforts have been devoted to solving it from different perspectives [1–10]. While some attempts have been made to tackle this problem with multiple microphones, leading to the so-called multichannel noise reduction techniques, most efforts in the literature focus on the single-sensor case as a large portion of current voice communication devices are equipped with only one microphone. Therefore, this paper is dedicated to the single-channel noise reduction problem with the objective of reducing the noise from a noisy microphone signal, thereby improving the perceptual speech quality and signal-to-noise ratio (SNR) [5–8].

With a single microphone, noise reduction is typically accomplished by linear filtering, i.e., passing the noisy signal through a filter. Since both the clean speech and noise are filtered at the same time, the most critical, yet most challenging, issue of noise reduction becomes one of designing a proper filter that can significantly mitigate the noise effect while maintaining the filtered speech signal close to its original form. While the filters can be designed in the time domain [3,5,11–14], the frequency-domain approaches are preferred. There are many practical reasons for this. First of all, most of our knowledge and understanding of speech production and perception are related to frequencies. In the frequency domain, it is not only easier to design noise reduction filters, but it is more straightforward to analyze and tune their performance as well. Secondly, thanks to the fast Fourier transform (FFT), the implementation of frequency-domain filters can be made, in general, computationally more efficient than filters in the time domain. Furthermore, the statistics of both the speech and noise signals can be better exploited in the frequency domain to optimize performance.

Since speech signals are nonstationary and noise can be nonstationary as well, frequency-domain approaches are implemented with the short-time Fourier transform (STFT). The fundamental paradigm of this structure consists of four basic steps. First, the noisy speech is divided into short-time frames. Then, each frame is transformed into the frequency domain via the STFT. This step is often called the analysis part of the process. Next, the STFT coefficients of speech and noise from different frequency bands (or STFT bins) are assumed to be uncorrelated and a gain is designed

☆ This work is partially supported by the Chinese Specialized Research Fund for the Doctoral Program of Higher Education (#: 20136102110010).

* Corresponding author.
E-mail addresses: huanghai@mail.nwpu.edu.cn (H. Huang), zhaoliheng@gmail.com (L. Zhao), jingdongchen@ieee.org (J. Chen), benesty@emt.inrs.ca (J. Benesty).

http://dx.doi.org/10.1016/j.dsp.2014.06.008
1051-2004/© 2014 Elsevier Inc. All rights reserved.

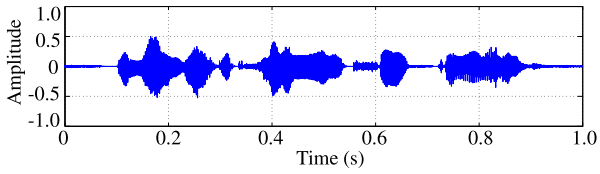

Fig. 1. A clean speech signal sampled at 8 kHz.

and applied in each subband to obtain an estimate of the clean speech STFT coefficients. Finally, the time-domain enhanced speech is synthesized from the estimated clean speech STFT coefficients using the inverse STFT. In this paradigm, the most critical step is the design of the noise reduction gain. This issue has been extensively studied over the last four decades and many algorithms have been developed, such as the spectral subtraction method [15,16], the Wiener gain [1,17], the maximum likelihood (ML) spectral amplitude estimator [17], the minimum-mean-square-error (MMSE) estimator [2], the maximum-a-posteriori (MAP) estimator [18], to name a few. A common assumption made by all these algorithms is that the STFT coefficients from different frequency bands are uncorrelated so that the noise reduction can be processed in every subband independently. This assumption may be true if the signals to be dealt with are stationary and the frame length in the short-time analysis is sufficiently large. However, it is well known that speech is nonstationary and the frame length cannot be too large. Moreover, overlap between neighboring frames is needed to avoid aliasing. As a result, the STFT coefficients from neighboring frequency bands generally exhibit strong correlation. To illustrate this, we recorded a 1-s 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 overlapping frames with a frame length of 16 ms and 75% overlap (a typical configuration example of noise reduction). Every frame is transformed into the STFT domain using a 128-point FFT. We then computed the normalized cross-correlation coefficients between different STFT frequency bands. [If we denote by A_i and A_j the FFT coefficients from the ith and jth frequency bands, the cross-correlation coefficient between the STFT coefficients from the two bands is defined as $\rho_{ij} = E(A_i A_j^*)/\sqrt{E(|A_i|^2)E(|A_j|^2)}$, where $E(\cdot)$ and * denote, respectively, mathematical expectation and complex conjugate.] Fig. 2 plots the results for the 4th, 8th, and 16th FFT bands [8,19]. It is clearly seen that there is a strong correlation between frequency bands that are next to each other. Then, one may ask some legitimate questions: is the interband correlation important for noise reduction? If so, how can we use such correlation information to improve the noise reduction performance?

Early attempts to answer the above questions can be found in [20] and [21], where an MMSE estimator and a Wiener filter were derived, respectively, based on the use of correlation among all the STFT frequency bands. While these algorithms are interesting from a theoretical viewpoint, they suffer from some practical drawbacks. First, to implement them, one would need to compute the inverse of a correlation matrix for each time frame, whose dimension depends on the FFT length. This makes the implementation of the algorithms computationally prohibitive. Second, a large number of signal frames is required to estimate the needed correlation matrices; otherwise, those matrices would be either rank deficient or ill conditioned. However, when a large number of frames are used, the estimate of these matrices would not follow the true statistics of the nonstationary speech signal, causing degradation in noise reduction performance. As shown in Fig. 2, correlation only exists between neighboring frequency bands and there is not much correlation between distant bands. Based on this observation, a framework based on the bifrequency spectrum

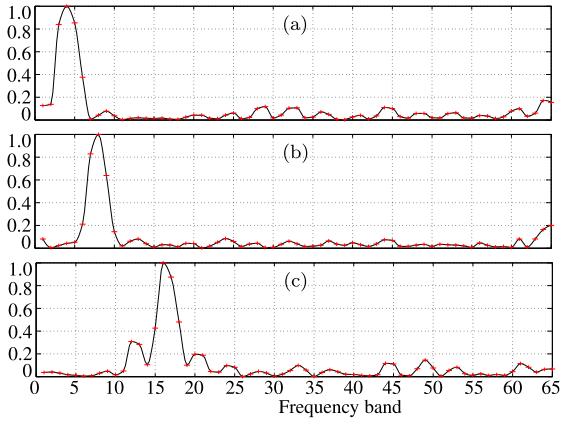

Fig. 2. The magnitude of the cross-correlation coefficients between: (a) the 4th and other frequency bins, (b) the 8th and other bins, and (c) the 16th and other bins. The sampling rate is 8 kHz, the frame length is 16 ms (128 points), the FFT length is 128, and the overlap is 75%.

was developed in [8] and [19]. Under this framework, Wiener and tradeoff filters were deduced, which are more practical than the algorithms in [20] and [21] in terms of noise reduction performance, implementation, and robustness.

As we know, the celebrated minimum variance distortionless response (MVDR) algorithm, originally proposed by Capon [22], has been found effective in dealing with noise reduction with microphone arrays. However, the design of an MVDR filter with conventional approaches using only a single microphone is not possible. Recently, a framework considering inter-frame correlation information was introduced in noise reduction and a multi-frame MVDR filter in the STFT domain was developed in [23] and [24], which can significantly improve noise reduction performance as compared to the traditional approaches. However, it also introduces much longer delay (depending on the number of frames used), which makes it difficult to deploy in many real-time applications such as telecommunications and hearing aids, where the delay introduced by a noise reduction processor is expected to be only a few milliseconds.

To take into consideration of both the noise reduction performance and the delay issue, we attempt here to extend the basic idea in [8] and [19] to develop an MVDR filter that can exploit the interband information. In comparison with the interband Wiener and tradeoff filters developed in [8] and [19] that may introduce much speech distortion, the MVDR filter developed in this paper can improve the output SNR and meanwhile maintain the speech distortion at a very low level. Compared with the multi-frame MVDR filter given in [23], the proposed MVDR filter achieves noise reduction through only filtering one frame (the current processing frame) at a time. It, therefore, introduces much less delay.

The rest of this paper is organized as follows. Section 2 describes the signal model and the traditional formulation of the noise reduction problem in the STFT domain. Section 3 introduces the concept and definition of the bifrequency spectrum that can measure the degree of correlation between the STFT coefficients from different frequency bands. Then, in Section 4, an MVDR filter based on the correlation among all the frequency bands is developed. A more practical version of the MVDR filter is deduced in Section 5. Section 6 is concerned with the Wiener filter. Section 7 presents some useful performance measures for the evaluation of the developed MVDR filters. Simulation results are presented in Section 8 and finally conclusions are drawn in Section 9.

2. Signal model and traditional formulation of the noise reduction problem

2.1. Signal model

The noise reduction problem considered in this paper is one of recovering the desired signal (clean speech) $x(t)$, t being the time index, of zero mean from the noisy observation (microphone signal) [5,7,11]:

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

where $v(t)$ is the unwanted additive noise, which is assumed to be a zero-mean random process, white or colored, but uncorrelated with $x(t)$. The signals $x(t)$ and $v(t)$ are considered to be real and broadband. Using the STFT, (1) can be rewritten in the time-frequency domain as

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

where $Y(k,m)$, $X(k,m)$, and $V(k,m)$ are the STFTs of $y(t)$, $x(t)$, and $v(t)$, respectively, at the frequency band (or bin) $k \in \{0,1,\ldots,K-1\}$ and the time frame m. Since $x(t)$ and $v(t)$ are uncorrelated by assumption, the variance of $Y(k,m)$ is

$$\begin{aligned}\phi_Y(k,m) &\triangleq E\big[|Y(k,m)|^2\big] \\ &= \phi_X(k,m) + \phi_V(k,m),\end{aligned} \tag{3}$$

where $\phi_X(k,m) \triangleq E[|X(k,m)|^2]$ and $\phi_V(k,m) \triangleq E[|V(k,m)|^2]$ are the variances of $X(k,m)$ and $V(k,m)$, respectively.

2.2. Traditional formulation

With the signal model given in (2), the traditional approach assumes that the STFT coefficients from different subbands are mutually uncorrelated. In this case, the noise reduction problem consists of estimating $X(k,m)$ from $Y(k,m)$. This estimation is accomplished by applying a complex gain to the observation signal, $Y(k,m)$, [7], i.e.,

$$\begin{aligned}\widehat{X}(k,m) &= H(k,m)Y(k,m) \\ &= X_{\mathrm{f}}(k,m) + V_{\mathrm{rn}}(k,m),\end{aligned} \tag{4}$$

where $X_{\mathrm{f}}(k,m) \triangleq H(k,m)X(k,m)$ is the filtered version of the desired signal and $V_{\mathrm{rn}}(k,m) \triangleq H(k,m)V(k,m)$ is the residual noise. With this formulation, the aim of the traditional noise reduction approach is to find an optimal gain at every time frame m and frequency band k, i.e., $H_{\mathrm{o}}(k,m)$, so that the level of the residual noise after filtering [synthesized from $V_{\mathrm{rn}}(k,m)$] is significantly smaller than that of the original noise, $v(t)$, and meanwhile the filtered version of the desired signal [synthesized from $X_{\mathrm{f}}(k,m)$] is (perceptually) as close as possible to the original signal, $x(t)$. See [5–8,11] and many references therein on how different optimal noise reduction filters are obtained. The issue with the traditional formulation is that the interband correlation is neglected. In the next section, we show how this interband correlation can be measured with the so-called bifrequency spectrum.

3. Bifrequency spectrum

Before we discuss how the interband correlation can be used, 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 [25,26]

$$\phi_A\big(k,k',m\big) \triangleq E\big[A(k,m)A^*\big(k',m\big)\big], \tag{5}$$

where k and k' are possibly two different frequency bands. Basically, the bifrequency spectrum is a measure of the correlation between two different frequency bands of the same signal. If $a(t)$ is a wide-sense stationary signal and a long FFT length is used to represent $A(k,m)$, the bifrequency spectrum reduces to

$$\phi_A\big(k,k',m\big) = \begin{cases}\phi_A(k,m), & k = k' \\ 0, & k \neq k'\end{cases} \tag{6}$$

where $\phi_A(k,m) = \phi_A(k,k,m)$ is the variance of $A(k,m)$. Thus, for a stationary random process, the Fourier coefficients from two different bands are uncorrelated. However, for a nonstationary random process such as speech, the bifrequency spectrum will exhibit nonzero correlations along the so-called support curves other than the main diagonal $k = k'$ as it was shown in Section 1. It seems then appropriate, when deriving noise reduction algorithms in the STFT domain, to take into account the spectral correlation that are not negligible in this context.

4. MVDR filter with correlation among all frequency bands

Let us first concatenate all the K frequency bands of the observation signal in a vector:

$$\begin{aligned}\mathbf{y}(m) &\triangleq \big[\,Y(0,m) \quad Y(1,m) \quad \cdots \quad Y(K-1,m)\,\big]^T \\ &= \mathbf{x}(m) + \mathbf{v}(m),\end{aligned} \tag{7}$$

where the superscript ^{T} is the transpose operator, and $\mathbf{x}(m)$ and $\mathbf{v}(m)$ are also vectors of length K, which concatenate all frequency bins of the desired and noise signals, respectively.

To estimate the desired signal, $X(k,m)$, from $\mathbf{y}(m)$, we first decompose $\mathbf{x}(m)$ into two orthogonal components:

$$\mathbf{x}(m) = X(k,m)\boldsymbol{\gamma}_{X,k}(m) + \mathbf{x}_{\mathrm{i},k}(m), \tag{8}$$

where

$$\begin{aligned}\boldsymbol{\gamma}_{X,k}(m) &= \big[\,\gamma_X(0,k,m) \quad \cdots \quad \gamma_X(K-1,k,m)\,\big]^T \\ &= \frac{E[\mathbf{x}(m)X^*(k,m)]}{\phi_X(k,m)}\end{aligned} \tag{9}$$

is the (normalized) interband correlation vector,

$$\mathbf{x}_{\mathrm{i},k}(m) = \mathbf{x}(m) - X(k,m)\boldsymbol{\gamma}_{X,k}(m) \tag{10}$$

is the interference signal vector with respect to the desired signal, $X(k,m)$, and

$$\gamma_X\big(k,k',m\big) \triangleq \frac{\phi_X(k,k',m)}{\phi_X(k,m)} \tag{11}$$

is a function of the bifrequency spectrum of $X(k,m)$. By taking the definition of $\mathbf{x}_{\mathrm{i},k}(m)$ in (10) and using (9), we can easily obtain

$$E\big[X^*(k,m)\mathbf{x}_{\mathrm{i},k}(m)\big] = \mathbf{0}. \tag{12}$$

Now, we propose to estimate the desired signal as follows:

$$\widehat{X}(k,m) = \mathbf{h}_k^H(m)\mathbf{y}(m), \tag{13}$$

where $\mathbf{h}_k(m)$ is a complex-valued filter of length K and the superscript ^{H} is the conjugate-transpose operator. Using (8), we can rewrite (13) as

$$\begin{aligned}\widehat{X}(k,m) &= \mathbf{h}_k^H(m)\big[\mathbf{x}(m) + \mathbf{v}(m)\big] \\ &= X(k,m)\mathbf{h}_k^H(m)\boldsymbol{\gamma}_{X,k}(m) \\ &\quad + \mathbf{h}_k^H(m)\mathbf{x}_{\mathrm{i},k}(m) + \mathbf{h}_k^H(m)\mathbf{v}(m) \\ &= X_{\mathrm{fd}}(k,m) + X_{\mathrm{ri}}(k,m) + V_{\mathrm{rn}}(k,m),\end{aligned} \tag{14}$$

where

$$X_{\mathrm{fd}}(k,m) \triangleq X(k,m)\mathbf{h}_k^H(m)\boldsymbol{\gamma}_{X,k}(m) \tag{15}$$

is the filtered desired signal,

$$X_{\mathrm{ri}}(k,m) \triangleq \mathbf{h}_k^H(m)\mathbf{x}_{\mathrm{i},k}(m) \tag{16}$$

is the residual interference, and

$$V_{\mathrm{rn}}(k,m) \triangleq \mathbf{h}_k^H(m)\mathbf{v}(m) \tag{17}$$

is the residual noise. It can be checked that the estimate of the desired signal is the sum of three terms that are mutually uncorrelated. Therefore, the variance of $\widehat{X}(k,m)$ is

$$\begin{aligned}\phi_{\widehat{X}}(k,m) &= \mathbf{h}_k^H(m)\boldsymbol{\Phi}_{\mathbf{y}}(m)\mathbf{h}_k(m)\\ &= \phi_{X_{\mathrm{fd}}}(k,m) + \phi_{X_{\mathrm{ri}}}(k,m) + \phi_{V_{\mathrm{rn}}}(k,m),\end{aligned} \tag{18}$$

where

$$\begin{aligned}\boldsymbol{\Phi}_{\mathbf{y}}(m) &\triangleq E\left[\mathbf{y}(m)\mathbf{y}^H(m)\right]\\ &= \begin{bmatrix} \phi_Y(0,m) & \phi_Y(0,1,m) & \cdots & \phi_Y(0,K-1,m)\\ \phi_Y(1,0,m) & \phi_Y(1,m) & \cdots & \phi_Y(1,K-1,m)\\ \vdots & \vdots & \ddots & \vdots\\ \phi_Y(K-1,0,m) & \phi_Y(K-1,1,m) & \cdots & \phi_Y(K-1,m)\end{bmatrix}\end{aligned} \tag{19}$$

is the correlation matrix of $\mathbf{y}(m)$,

$$\begin{aligned}\phi_{X_{\mathrm{fd}}}(k,m) &\triangleq E\left[\left|X_{\mathrm{fd}}(k,m)\right|^2\right]\\ &= \phi_X(k,m)\left|\mathbf{h}_k^H(m)\boldsymbol{\gamma}_{X,k}(m)\right|^2,\end{aligned} \tag{20}$$

$$\begin{aligned}\phi_{X_{\mathrm{ri}}}(k,m) &\triangleq E\left[\left|X_{\mathrm{ri}}(k,m)\right|^2\right]\\ &= \mathbf{h}_k^H(m)\boldsymbol{\Phi}_{\mathbf{x}_{\mathrm{i},k}}(m)\mathbf{h}_k(m)\\ &= \mathbf{h}_k^H(m)\boldsymbol{\Phi}_{\mathbf{x}}(m)\mathbf{h}_k(m)\\ &\quad - \phi_X(k,m)\left|\mathbf{h}_k^H(m)\boldsymbol{\gamma}_{X,k}(m)\right|^2,\end{aligned} \tag{21}$$

$$\begin{aligned}\phi_{V_{\mathrm{rn}}}(k,m) &\triangleq E\left[\left|V_{\mathrm{rn}}(k,m)\right|^2\right]\\ &= \mathbf{h}_k^H(m)\boldsymbol{\Phi}_{\mathbf{v}}(m)\mathbf{h}_k(m),\end{aligned} \tag{22}$$

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

Now, we can derive an MVDR filter by minimizing $\phi_{\widehat{X}}(k,m)$ with the constraint that the desired signal is not distorted. Mathematically, this is equivalent to

$$\min_{\mathbf{h}_k(m)} \mathbf{h}_k^H(m)\boldsymbol{\Phi}_{\mathbf{y}}(m)\mathbf{h}_k(m) \quad \text{subject to } \mathbf{h}_k^H(m)\boldsymbol{\gamma}_{X,k}(m) = 1, \tag{23}$$

for which the solution is

$$\mathbf{h}_{\mathrm{MVDR},k}(m) = \frac{\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(m)\boldsymbol{\gamma}_{X,k}(m)}{\boldsymbol{\gamma}_{X,k}^H(m)\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(m)\boldsymbol{\gamma}_{X,k}(m)}. \tag{24}$$

Using (9), the interband correlation vector $\boldsymbol{\gamma}_{X,k}(m)$ can be rewritten as

$$\begin{aligned}\boldsymbol{\gamma}_{X,k}(m) &= \frac{E[\mathbf{x}(m)X^*(k,m)]}{\phi_X(k,m)}\\ &= \frac{\boldsymbol{\Phi}_{\mathbf{x}}(m)\mathbf{i}_{k+1}}{\phi_X(k,m)} = \frac{[\boldsymbol{\Phi}_{\mathbf{y}}(m) - \boldsymbol{\Phi}_{\mathbf{v}}(m)]\mathbf{i}_{k+1}}{\phi_Y(k,m) - \phi_V(k,m)},\end{aligned} \tag{25}$$

where $\mathbf{i}_{k+1}$ is the $(k+1)$th column of the $K \times K$ identity matrix, $\mathbf{I}_K$. Now, $\boldsymbol{\gamma}_{X,k}(m)$ depends on the second-order statistics of the observation and noise signals. The statistics of the noise signal can be estimated in practice, as in the traditional approach, with the help of a voice activity detector (VAD).

In the particular scenario where the interband correlations of both speech and noise are negligible, the correlation matrix $\boldsymbol{\Phi}_{\mathbf{y}}(m)$ degenerates to a diagonal one and (8) simplifies to

$$\mathbf{x}(m) = X(k,m)\mathbf{i}_{k+1} + \mathbf{x}_{\mathrm{i},k}(m), \tag{26}$$

so $\mathbf{x}_{\mathrm{i},k}(m)$ resembles $\mathbf{x}(m)$ except for its $(k+1)$th component which is 0. In this case, the MVDR filter becomes

$$\mathbf{h}_{\mathrm{MVDR},k}(m) = \mathbf{i}_{k+1}, \tag{27}$$

which does not modify the observation signal. As a result, there is neither noise reduction nor speech distortion.

5. Suboptimal MVDR filter using correlation among neighboring bands

In the previous section, we derived an optimal MVDR filter that exploits the correlation among all the frequency bands. However, this filter may have some practical drawbacks. First, to implement it, we need to compute the inverse of a $K \times K$ matrix at 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 matrix $\boldsymbol{\Phi}_{\mathbf{y}}(m)$; otherwise, it would be ill conditioned or even rank deficient. Furthermore, when a large number of frames is used, the estimate of this matrix (as well as other quantities) would not follow the true statistics of the nonstationary speech signal, causing degradation in noise reduction performance. As shown in Section 1, correlation is strong only between neighboring frequency bands, while it is negligible between distant bands. Given this, we introduce a suboptimal, yet more practical, approach in this section that considers correlation between only neighboring frequency bands. So, instead of estimating $X(k,m)$ from the noisy vector, $\mathbf{y}(m)$, of length K, we now estimate it from a lower dimensional vector:

$$\begin{aligned}\mathbf{y}_k(m) \triangleq \big[&Y\left(k-K_k^-,m\right) \quad \cdots \quad Y(k-1,m)\\ &Y(k,m) \quad Y(k+1,m) \quad \cdots \quad Y\left(k+K_k^+,m\right)\big]^T\end{aligned} \tag{28}$$

of length $L = K_k^- + K_k^+ + 1 \ll K$, where K_k^- and K_k^+ are, respectively, the numbers of bins before and after the frequency bin k. We can also write (28) as

$$\mathbf{y}_k(m) = \mathbf{x}_k(m) + \mathbf{v}_k(m), \tag{29}$$

where $\mathbf{x}_k(m)$ and $\mathbf{v}_k(m)$ are defined in a similar way to $\mathbf{y}_k(m)$. Again, we decompose $\mathbf{x}_k(m)$ as follows:

$$\mathbf{x}_k(m) = X(k,m)\boldsymbol{\rho}_{X,k}(m) + \mathbf{x}'_{\mathrm{i},k}(m), \tag{30}$$

where

$$\boldsymbol{\rho}_{X,k}(m) \triangleq \frac{E[\mathbf{x}_k(m)X^*(k,m)]}{\phi_X(k,m)} \tag{31}$$

is the (normalized) interband correlation vector, $\mathbf{x}'_{\mathrm{i},k}(m) = \mathbf{x}_k(m) - X(k,m)\boldsymbol{\rho}_{X,k}(m)$ is the interference signal vector, and

$$E\left[X^*(k,m)\mathbf{x}'_{\mathrm{i},k}(m)\right] = \mathbf{0}. \tag{32}$$

We can estimate the desired signal as

$$\begin{aligned}\widehat{X}(k,m) &= \mathbf{h}_k'^H(m)\mathbf{y}_k(m)\\ &= X(k,m)\mathbf{h}_k'^H(m)\boldsymbol{\rho}_{X,k}(m)\\ &\quad + \mathbf{h}_k'^H(m)\mathbf{x}'_{\mathrm{i},k}(m) + \mathbf{h}_k'^H(m)\mathbf{v}_k(m),\end{aligned} \tag{33}$$

where $\mathbf{h}'_k(m)$ is a filter of length $L = K_k^- + K_k^+ + 1$. Following the same steps as in Section 4, we easily derive the suboptimal MVDR filter:

$$\mathbf{h}'_{\mathrm{MVDR},k}(m) = \frac{\boldsymbol{\Phi}_{\mathbf{y}_k}^{-1}(m)\boldsymbol{\rho}_{X,k}(m)}{\boldsymbol{\rho}_{X,k}^H(m)\boldsymbol{\Phi}_{\mathbf{y}_k}^{-1}(m)\boldsymbol{\rho}_{X,k}(m)}, \tag{34}$$

where $\boldsymbol{\Phi}_{\mathbf{y}_k}(m) \triangleq E[\mathbf{y}_k(m)\mathbf{y}_k^H(m)]$ is the correlation matrix of $\mathbf{y}_k(m)$.

6. Wiener filter using correlation among neighboring bands

The optimal Wiener filter can be considered as one of the most fundamental and widely used noise reduction approaches in the literature, which has been delineated in different forms and adopted in various applications. It has been shown that many methods such as the spectral subtraction [15,16] and the MMSE estimator [2] are closely related to this filter. With the bifrequency spectrum, the Wiener filter can be derived as in [8,19], i.e.,

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

where $\boldsymbol{\Phi}_{\mathbf{y}_k}(m)$ and $\boldsymbol{\Phi}_{\mathbf{x}_k}(m)$ are the correlation matrix of $\mathbf{y}_k(m)$ and $\mathbf{x}_k(m)$ respectively, as defined in Section 5, $\mathbf{I}$ 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 $\mathbf{I}$.

The focus of this paper is on the MVDR filter. However, the Wiener filter will also be included in Section 8 for the purpose of comparison.

7. Performance measures

To facilitate the evaluation of the developed single-channel noise reduction MVDR filters based on the bifrequency spectrum, some performance measures are presented in this section.

Given the signal model in (1), we define the input SNR as the ratio of the variance of the desired signal over the variance of the background noise, i.e.,

$$\mathrm{iSNR} = \frac{E[x^2(t)]}{E[v^2(t)]}. \tag{36}$$

To quantify the level of noise remaining at the output of the noise reduction filter, we define the output SNR as the ratio of the variance of the filtered desired signal over the variance of the residual interference-plus-noise[1]:

$$\mathrm{oSNR} = \frac{E[x_{\mathrm{fd}}^2(t)]}{E\{[x_{\mathrm{ri}}(t) + v_{\mathrm{rn}}(t)]^2\}}, \tag{37}$$

where $x_{\mathrm{fd}}(t)$, $x_{\mathrm{ri}}(t)$, and $v_{\mathrm{rn}}(t)$ are the time-domain signals reconstructed from $X_{\mathrm{fd}}(k,m)$, $X_{\mathrm{ri}}(k,m)$, and $V_{\mathrm{rn}}(k,m)$, respectively. One of the most important goals of noise reduction is to improve the SNR.

To quantify the distortion level of the filtered desired signal, we borrow the concept of the speech distortion index, which is defined as [7]

$$\upsilon_{\mathrm{sd}} \triangleq \frac{E\{[x_{\mathrm{fd}}(t) - x(t)]^2\}}{E[x^2(t)]}. \tag{38}$$

The speech distortion index is always greater than or equal to 0. The higher the value of this index, the more the desired signal is distorted.

[1] In this paper, we consider the uncorrelated interference as part of the noise in the definition of the performance measures.

Besides the above two performance metrics, we also use the perceptual-evaluation-of-speech-quality (PESQ) measure [27], which has been found to have higher correlations than other widely known objective measures, with the subjective ratings of overall quality of enhanced speech signal.

8. Simulations

In this section, we evaluate the MVDR filters developed in Sections 4 and 5 through simulations.

8.1. Signals and noise

In our simulations, the desired clean speech signal consists of two parts, one from a male talker and the other from a female talker, both are approximately 30-s long and sampled at 8 kHz. The noisy signal is generated by adding noise to the desired signal. The noise signal is then properly scaled to control the input SNR. We study four different types of noise, i.e., Gaussian, car, F-16 cockpit, and NYSE babble noise, which are representative noise samples from white and stationary to highly nonstationary. The computer-generated Gaussian noise is white and stationary. The car noise is recorded from a sedan car running at 50 miles/hour on a highway (all the windows are closed); this noise is still close to stationary, but it is colored. The F-16 cockpit noise, taken from the Noisex92 database [28], is recorded at the co-pilot's seat in a two-seat F-16, traveling at a speed of 500 knots, and an altitude of 300–600 feet, which is colored and nonstationary. The babble noise is recorded in a New York Stock Exchange (NYSE) room; it consists of sounds from various sources such as speakers, telephone rings, electric fans, etc and is highly nonstationary.

8.2. Implementation of the MVDR and Wiener filters

The noise reduction filters are implemented using the overlap add technique. The overlap factor is always 75% between the neighboring frames. To overcome the aliasing problem, a Kaiser window is applied both before the analysis and after the resynthesis steps.

8.3. Estimation of the statistics

To implement the MVDR filter in (34) [(24) is a particular case of (34)] and the Wiener filter in (35), we need to know the parameters $\boldsymbol{\Phi}_{\mathbf{y}_k}(m)$, $\boldsymbol{\Phi}_{\mathbf{v}_k}(m)$, and $\boldsymbol{\rho}_{X,k}(m)$. In our simulations, we first assume that the noise signal is accessible and estimate $\boldsymbol{\Phi}_{\mathbf{y}_k}(m)$ and $\boldsymbol{\Phi}_{\mathbf{v}_k}(m)$ from the corresponding signals with a short-time average using the most recent N frames, i.e.,

$$\boldsymbol{\Phi}_{\mathbf{y}_k}(m) = \frac{1}{N}\sum_{i=0}^{N-1}\mathbf{y}_k(m-i)\mathbf{y}_k^H(m-i). \tag{39}$$

The matrix $\boldsymbol{\Phi}_{\mathbf{v}_k}(m)$ is estimated in a similar way to $\boldsymbol{\Phi}_{\mathbf{y}_k}(m)$. An estimate of $\boldsymbol{\Phi}_{\mathbf{x}_k}(m)$ is computed by subtracting $\boldsymbol{\Phi}_{\mathbf{v}_k}(m)$ from $\boldsymbol{\Phi}_{\mathbf{y}_k}(m)$, i.e., $\boldsymbol{\Phi}_{\mathbf{x}_k}(m) = \boldsymbol{\Phi}_{\mathbf{y}_k}(m) - \boldsymbol{\Phi}_{\mathbf{v}_k}(m)$. Then, $\boldsymbol{\rho}_{X,k}(m)$ is taken as the $(K_k^- + 1)$th column of $\boldsymbol{\Phi}_{\mathbf{x}_k}(m)$ normalized by the $(K_k^- + 1)$th element of this vector. To avoid numerical issues, the inverse of $\boldsymbol{\Phi}_{\mathbf{y}_k}(m)$ is computed as follows. We first compute the eigenvalue decomposition of $\boldsymbol{\Phi}_{\mathbf{y}_k}(m)$:

$$\boldsymbol{\Phi}_{\mathbf{y}_k}(m) = \mathbf{Q}(m)\boldsymbol{\Lambda}_{\mathbf{y}_k}(m)\mathbf{Q}^{-1}(m), \tag{40}$$

where $\boldsymbol{\Lambda}_{\mathbf{y}_k}(m)$ is a diagonal matrix whose diagonal elements correspond to the eigenvalues of $\boldsymbol{\Phi}_{\mathbf{y}_k}(m)$, i.e., $\lambda_i(m)$, $i = 1, 2, \ldots, K_k^- + K_k^+ + 1$ and $\mathbf{Q}(m)$ is the eigenvector matrix. Then, the inverse of $\boldsymbol{\Phi}_{\mathbf{y}_k}(m)$ is computed as

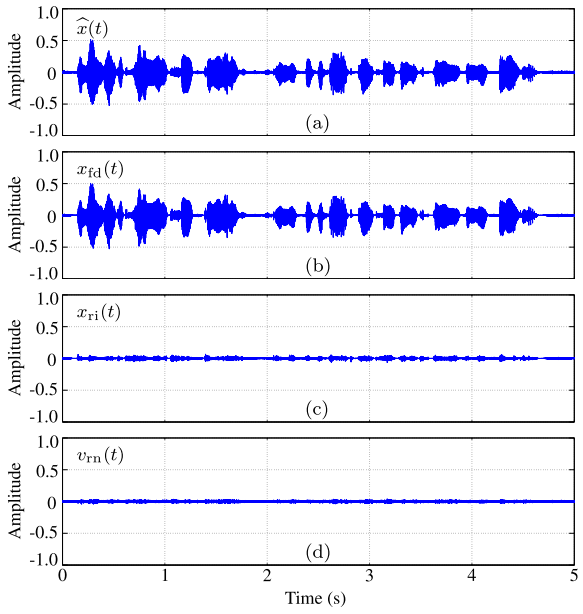

Fig. 3. Signals (first 5 s) after noise reduction by the MVDR filter (in white Gaussian noise): (a) the enhanced signal, $\widehat{x}(t)$, (b) the filtered desired signal, $x_{\mathrm{fd}}(t)$, (c) the residual interference, $x_{\mathrm{ri}}(t)$, and (d) the residual noise, $v_{\mathrm{rn}}(t)$. iSNR = 10 dB, $M = 128$, $K_k = 8$, and $N = 10$.

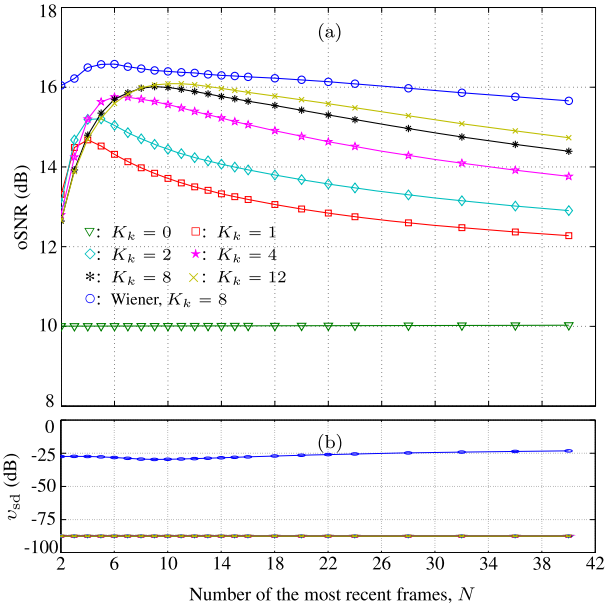

Fig. 4. Performance of the MVDR filter as a function of N for different values of K_k in white Gaussian noise. $M = 128$ and iSNR = 10 dB. For comparison, the performance of the Wiener filter with $K_k = 8$ is also plotted.

$$\mathbf{\Phi}_{\mathbf{y}_k}^{-1}(m) = \mathbf{Q}(m)\mathbf{\Lambda}_{\mathbf{y}_k}^{-1}(m)\mathbf{Q}^{-1}(m), \tag{41}$$

where $\mathbf{\Lambda}_{\mathbf{y}_k}^{-1}(m)$ is also a diagonal matrix whose ith diagonal element, denoted as $[\mathbf{\Lambda}_{\mathbf{y}_k}^{-1}(m)]_{ii}$, is computed as

$$\left[\mathbf{\Lambda}_{\mathbf{y}_k}^{-1}(m)\right]_{ii} = \begin{cases} \lambda_i^{-1}(m), & \text{if } \lambda_i(m) > \min\{\alpha\lambda_1(m), \delta\} \\ 0, & \text{else} \end{cases}, \tag{42}$$

where $\lambda_1(m)$ is the largest eigenvalue of $\mathbf{\Phi}_{\mathbf{y}_k}(m)$, $\alpha \in (0, 1)$ is a constant and $\delta > 0$ is a regularization parameter. In our simulations, these two parameters are empirically set to $\alpha = \delta = 0.01$.

8.4. Noise reduction performance as a function of the number of neighboring bands

In the first simulation, we set $K_k^- = K_k^+ = K_k$ and investigate the performance of the MVDR filter given in (34) as a function of K_k. The input SNR is 10 dB, the frame length (the same as the FFT size) is $M = 128$, and the number of frames, N, used for computing the correlation matrices, varies between 2 and 40. We vary the value of K_k from 0 to 12. When $K_k = 0$, the MVDR filter becomes the identity filter, which does not change the noisy signal and, therefore, there is neither noise reduction nor speech distortion. When K_k is greater than 0, the MVDR filter takes into account the interband correlation.

To visualize the noise reduction performance and also the signal decomposition used in the derivation of the MVDR filter, we plot in Fig. 3, the different signals (only the first 5 s) processed by this filter. It is clearly seen from this figure that the MVDR filter mitigates the noise significantly while maintaining the filtered designed signal close to the original desired one. To examine the noise reduction performance, we computed the output SNR, i.e., oSNR, and the speech distortion index, i.e., υ_{sd}. The results, as a function of N and for different values of K_k, are presented in Fig. 4. The performance of the Wiener filter with $K_k = 8$ is also plotted for comparison. It is seen that the MVDR filter does not introduce much speech distortion since the value of the speech distortion index is very small. As for the output SNR, one can see that it improves when K_k increases, which proves the usefulness of the interband correlation in noise reduction. However, when K_k is greater than 8, the additional performance gain as compared to that for $K_k = 8$ is not much while the complexity increase can be significant. This result justifies the superiority of the suboptimal MVDR over the optimal one. For a fixed value of $K_k > 0$, one can see that the output SNR first increases with N, and then decreases. The underlying reason can be explained as follows. If the value of N is too small, we cannot get a reliable estimation of the signal statistics, which affects the noise reduction performance. But if the value of N is too large, the estimated statistics cannot follow the time-varying properties of the speech signal, leading to performance degradation. We see that the optimal value of N depends slightly on the value of K_k. But generally, good performance is achieved when N is around 8 in our simulation setup. The Wiener filter with $K_k = 8$ yields a higher output SNR than the MVDR filter; but the Wiener filter adds speech distortion as clearly seen in the figure.

8.5. Performance as a function of the frame length

In the second simulation, we study how the frame length, M (in our implementation, we always take $M = 2K - 2$), affects the performance of the MVDR filter. Following the previous simulation, we set iSNR = 10 dB, $K_k = 8$, and vary N from 2 to 40. Fig. 5 shows the noise reduction performance as a function of N for different values of M. As we know, the spectral resolution and interband correlation are generally affected by the frame length. But the noise reduction performance only changes slightly with the frame length, as seen in Fig. 5. Note that in practical applications, we generally prefer to use a small frame length as this would introduce less processing delay. Based on this result, we will set the frame length to 128 (i.e., 16 ms) in all the following simulations.

8.6. Performance in different types of noise and SNR conditions

In this simulation, we evaluate the MVDR filter in four different types of noise, i.e., Gaussian, car, F-16 cockpit, and NYSE babble. Following the previous simulations, we set K_k to 8, M to 128, and N to 10. Fig. 6 plots the output SNR and the speech distortion index of the MVDR filter, both as a function of the input SNR. It is seen that the performance of the MVDR filter in the different

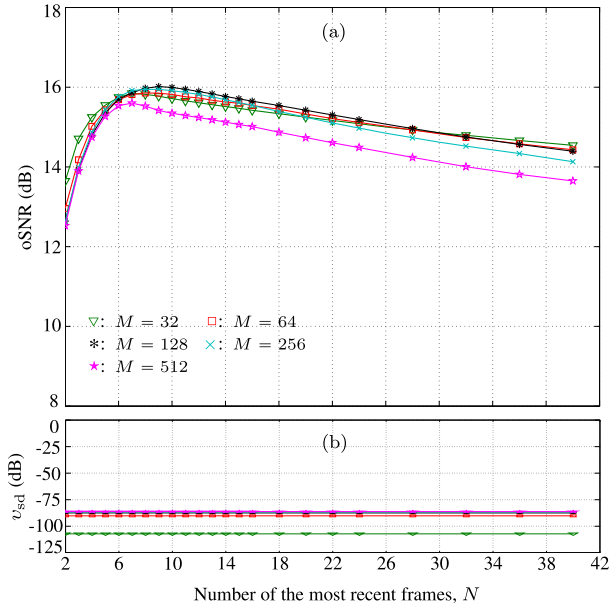

Fig. 5. Effect of the frame length, M, on the performance of the MVDR filter in white Gaussian noise. $K_k = 8$ and iSNR = 10 dB.

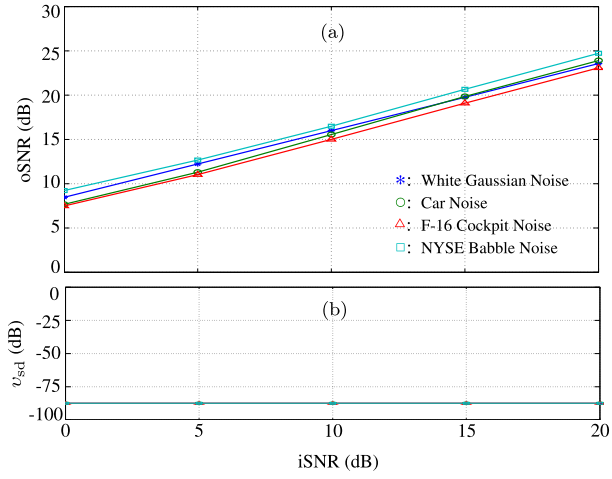

Fig. 6. Performance of the MVDR filter in different types of noise: white Gaussian, car, F-16 cockpit, and NYSE babble. $N = 10$, $K_k = 8$, and $M = 128$.

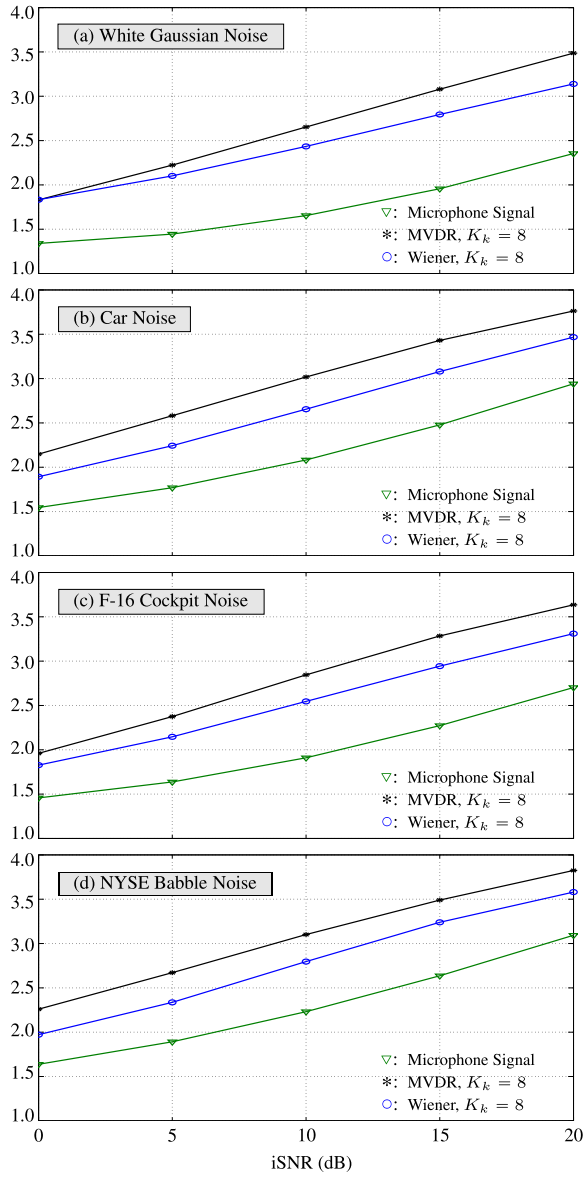

Fig. 7. The PESQ MOS-LQO scores of the noisy signal and the enhanced signal by the MVDR and Wiener filters in different types of noise: (a) white Gaussian, (b) car, (c) F-16 cockpit, and (d) NYSE babble. $M = 128$, $K_k = 8$, and $N = 10$.

types of noise is quite similar, which is somewhat unexpected since nonstationary noise is in general more difficult to deal with. The underlying reason may be due to the fact that the noise statistics are directly computed from the noise signal, thereby avoiding the statistics estimation error. In practical situations, however, the estimation error of noise statistics is unavoidable and would increase as the noise becomes more nonstationary. This estimation error will subsequently be translated into either more speech distortion or less noise reduction. We will come back to this point later.

8.7. Speech quality evaluation with PESQ

In the previous simulations, we examined the output SNR and the speech distortion index, which provide some insight into the noise reduction performance of the filters. In this simulation, we evaluate the quality of the enhanced speech through the perceptual-evaluation-of-speech-quality (PESQ) measure, which has been found to have higher correlation with the subjective ratings of the overall quality of enhanced speech signal [27] than other widely known objective measures. In this simulation, we use the PESQ MOS-LQO (listening quality objective) score same as in [24]. This score is computed in three steps. First, in a given noise and input SNR condition, the PESQ score is computed for each talker. The average PESQ score is then calculated. This average PESQ score is finally mapped to the PESQ MOS-LQO as follows:

$$\text{PESQ}_{\text{MOS-LQO}} = 0.999 + \frac{4}{1 + e^{-1.4945 \times \text{PESQ}_{\text{MOS}} + 4.6607}}. \tag{43}$$

The results of this simulation are shown in Fig. 7. For comparison, we also plotted the performance of the Wiener filter with $K_k = 8$. It is clearly seen that both the MVDR and the Wiener filters improve the PESQ MOS-LQO in all the studied noise and input SNR conditions.

From Fig. 4, we observe that the output SNR of the Wiener filter is higher than that of the MVDR filter. However, it is seen that the MVDR filter produces a higher PESQ score. The underlying reason is that the PESQ score is affected by both noise reduction and speech distortion. While it produces a higher output SNR than the MVDR filter, the Wiener filter adds more speech distortion. As

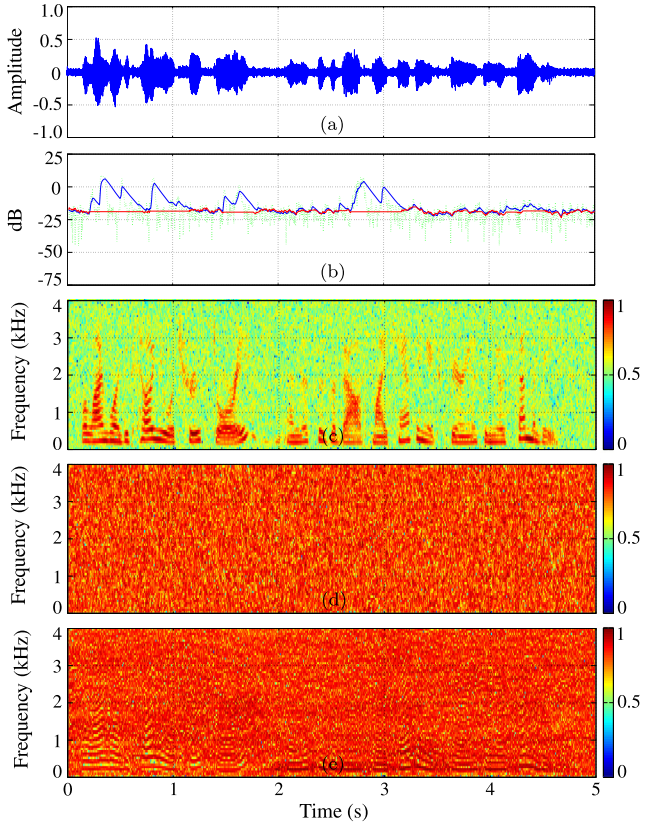

Fig. 8. Results (first 5 s) of noise estimation by IMCRA: (a) waveform of the noisy signal, $y(t)$; (b) magnitude square spectrum of the noisy signal (green dotted), smoothed magnitude square spectrum of the noisy signal (blue solid), and IMCRA noise estimate (red heavy solid) in the 20th frequency band; (c) spectrogram of the noisy signal; (d) spectrogram of the original noise signal; and (e) spectrogram of the estimated noise signal. The noisy signal is a clean speech from a female talker corrupted by a white Gaussian noise with iSNR = 10 dB. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a results, the overall speech quality of the Wiener filter is not as good as that of the MVDR filter, as indicated by the PESQ score.

8.8. Performance with noise estimation

In the previous simulations, we assumed that the noise signal is accessible and its correlation matrix in every frequency band is computed with a short-time average using the most recent time frames. While it provides a fair way to evaluate the derived noise reduction filter and compare its performance to other techniques, this way of computing the noise statistics makes the implementation impractical. In reality, the noise signal is generally not accessible and its spectrum has to be estimated from the noisy signal. Tremendous efforts have been devoted to this estimation problem and many useful methods have been developed. Representative ones include the recursive averaging algorithm [29,30], the minimum statistics tracking method [31,32], and the histogram-based approach [17,33], etc. In this paper, we adopt the improved minima controlled recursive averaging (IMCRA) approach in [30], which has been proved to be able to provide a robust noise power spectrum estimation in a broad range of noise environments.

Fig. 8 presents an example of the noise estimation using the IMCRA method. Fig. 8(a) plots the waveform of the noisy signal. The noisy signal is a female speech contaminated by a white Gaussian noise with an input SNR of 10 dB. Fig. 8(b) shows the magnitude square spectrum of the noisy signal, i.e., $|Y(k,m)|^2$, the smoothed magnitude square spectrum (smoothed using a single

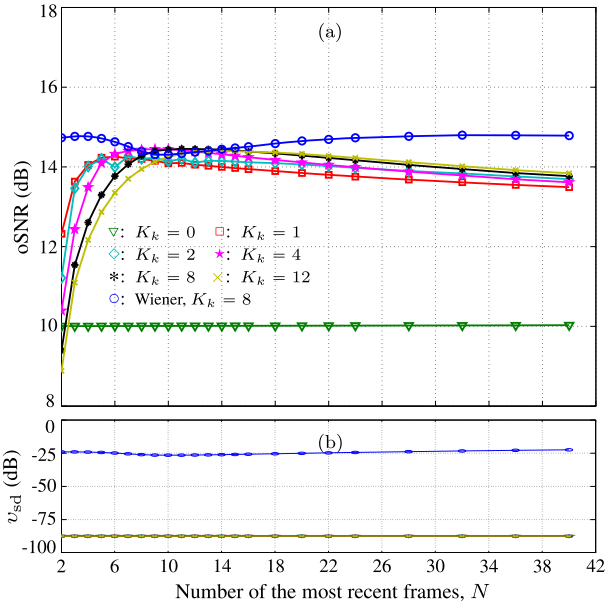

Fig. 9. Performance of the MVDR and Wiener filters (with IMCRA noise estimation) as a function of N for different values of K_k. iSNR = 10 dB and $M = 128$ in white Gaussian noise.

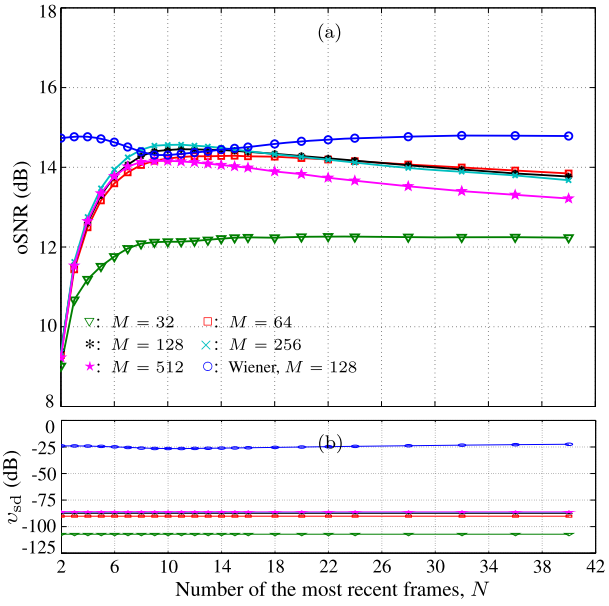

Fig. 10. Effect of the frame length, M, on the performance of the MVDR and Wiener filters (with IMCRA noise estimation) in white Gaussian noise. $K_k = 8$ and iSNR = 10 dB.

pole recursion with a smoothing factor of 0.9), and the noise spectrum estimated with the IMCRA approach in the 20th frequency band. Fig. 8(c), (d), and (e) plot the spectrograms of the noisy signal, the original noise signal, and the estimated noise signal with the IMCRA method. One can see that the IMCRA approach performs well in estimating the noise spectrum and signal.

With the IMCRA noise estimation, we can now assess both the MVDR and the Wiener filters for their practical performance. The evaluation results in terms of the output SNR and the speech distortion index are plotted in Figs. 9–11.

Comparing Fig. 9 with Fig. 4, one can see that the MVDR and Wiener filters yield less SNR improvement after incorporating noise estimation. This is, of course, expected as any noise estimation inevitably introduces some estimation error in the noise

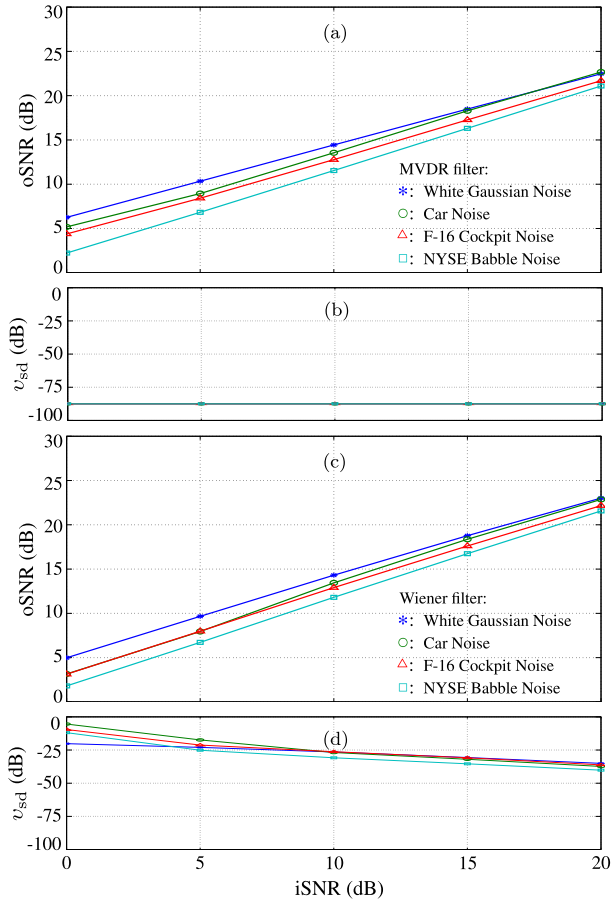

Fig. 11. Performance of the MVDR and Wiener filters (with IMCRA noise estimation) in different types of noise: white Gaussian, car, F-16 cockpit, and NYSE babble. $N = 10$, $K_k = 8$, and $M = 128$.

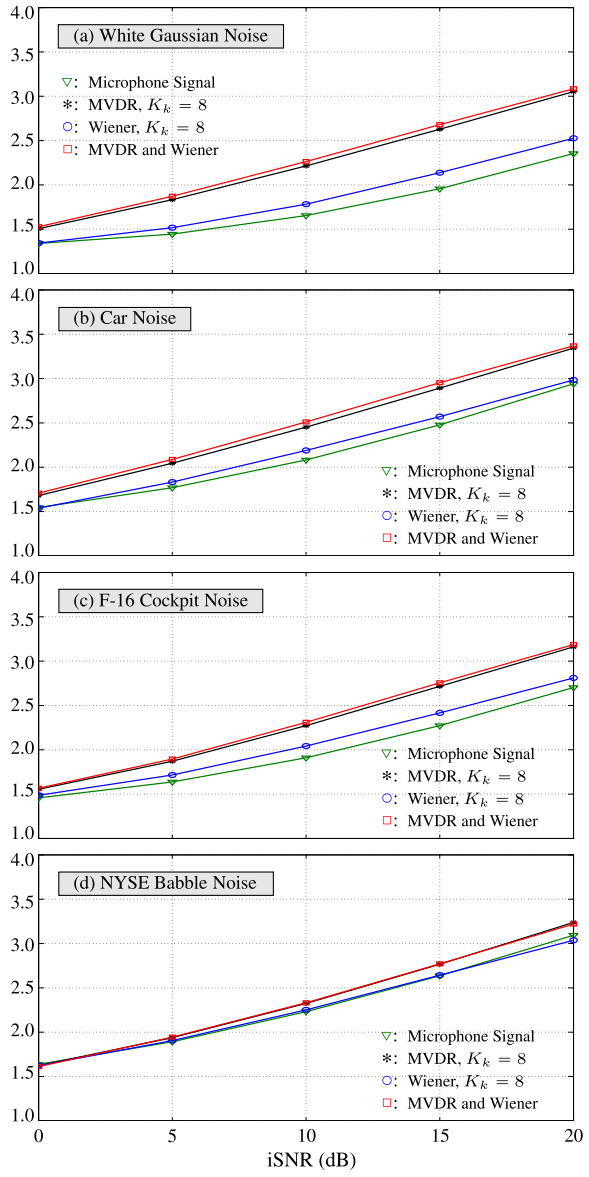

Fig. 12. PESQ MOS-LQO of the noisy signal and the enhanced signals by the MVDR and Wiener filters in four different types of noise: white Gaussian, car, F-16 cockpit, and NYSE babble. $M = 128$, $K_k = 8$, and $N = 10$.

statistics, which causes degradation in noise reduction performance. However, we can still see, from Fig. 9, that the MVDR filter improves the SNR significantly and meanwhile it preserves the desired speech with little distortion.

Fig. 11 plots the performance of the MVDR and Wiener filters in four different types of noise. In a given SNR condition, both filters yield the largest output SNR in white Gaussian noise and the smallest output SNR in the (highly nonstationary) NYSE babble noise. This is understandable as it is more difficult to estimate the statistics if the noise is nonstationary.

The PESQ evaluation results are plotted in Fig. 12. As seen, both the MVDR and Wiener filters improve the speech quality, as indicated by the PESQ MOS-LQO score. This, again, illustrates the importance of using the interband correlation in noise reduction. In comparison, the MVDR filter does a better job in improving speech quality than the Wiener filter since the former does not add speech distortion into the enhanced signal.

In some scenarios, if the application can tolerate some speech distortion but requires more noise reduction, one can achieve this by simply combining the MVDR and Wiener filters together, i.e., perform noise reduction using the MVDR filter first and then apply the Wiener to the output of the MVDR filter. This combination may provide at least two benefits. First, it should yield more noise reduction, thereby further improving the SNR. Second, it provides a better way for the Wiener filter to estimate the noise statistics since the MVDR filter would significantly improve the SNR as shown in the previous simulations. The results of this combination are also plotted in Fig. 12. One can see that combining the two filters can further improve speech quality as compared to the use of only MVDR filter, though the improvement is not significant. The reason that the quality improvement is not significant is due to the fact that the second-stage Wiener filter introduces speech distortion as it achieves noise reduction and the quality improvement is a compromise between the amount of noise reduction and the degree of speech distortion. Nevertheless, this combination gives another option for tradeoff between noise reduction and speech distortion. Many other filters can be used together with the MVDR filter.

8.8.1. Computational complexity

It can be checked that the computational complexity of the MVDR and Wiener filters is a function of the filter length L ($L = 2K_k + 1$, K_k is the number of the neighboring frequency bins). In this section, we analyze the computational complexity of these two filters in terms of the number of real-valued multiplications/divisions (the number of additions/subtractions are neglected because they are much quicker to compute in most generic hardware platforms). We assume that complex-valued multiplications are transformed into real-valued multiplications. The multiplica-

tion between a real number and complex number requires 2 real-valued multiplications. The multiplication between two complex numbers needs 4 real-valued multiplications. The division between a complex number and a real number requires 2 real-valued multiplications [34]. The results are summarized in Table 1. It is easy to see that the computational complexity of both filters increases L.

In comparison, the MVDR filter requires slightly more multiplications than the Wiener filter; but the difference is negligible with today's DSP processors.

Table 1
Computational complexity of the MVDR and Wiener filters in terms of the required number of multiplications.

Parameters:
$L = 2K_k + 1$: length of the MVDR and Wiener filters
K_k: number of the neighboring frequency bins for the MVDR and Wiener filters

Algorithm steps	Required (real-valued) multiplications
• Estimation of $\mathbf{\Phi}_{\mathbf{y}_k}(m)$ and $\mathbf{\Phi}_{\mathbf{v}_k}(m)$ (with short-time average)	$8L^2$
• Computing $\mathbf{\Phi}_{\mathbf{y}_k}^{-1}(m)$ (with Gauss–Jordan elimination method [35])	$4L^3$
• Computing $\mathbf{h}'_{\mathrm{MVDR},k}(m)$ in (34)	$4L^2 + 6L$
• Computing $\mathbf{h}'_{\mathrm{W},k}(m)$ in (35)	$4L^2$
• Computing $\widehat{X}(k,m)$	$4L$
• Total multiplications of the MVDR filter	$4L^3 + 12L^2 + 10L$
• Total multiplications of the Wiener filter	$4L^3 + 12L^2 + 4L$

9. Conclusions

This paper dealt with the problem of single-channel noise reduction in the STFT domain. Unlike the traditional methods that achieve noise reduction using only a gain and neglect the inter-band correlation information, the approach taken here exploits this information. The concept of bifrequency spectrum was introduced in this context. Noise reduction is then recast in the STFT domain as a filtering problem based on the bifrequency spectrum. Two versions of the MVDR filter were then deduced; one uses the interband correlation among all the STFT frequency bands while the other employs the correlation between only the neighboring bands. While the first version is optimal from a theoretical viewpoint, it is not robust. The second version is suboptimal, but is much more practical. A large number of simulations were carried out to evaluate the MVDR noise reduction filter with the noise spectrum being either directly computed from the noise signal or estimated from the noisy signal. The results showed that the developed MVDR filter can significantly improve the SNR while preserving the desired signal without much distortion. Moreover, it was also shown that the MVDR filter can dramatically improve the PESQ score in all the studied types of noise and SNR conditions.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.dsp.2014.06.008.

References

[1] J.S. Lim, A.V. Oppenheim, Enhancement and bandwidth compression of noisy speech, Proc. IEEE 67 (12) (1979) 1586–1604.
[2] Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Trans. Acoust. Speech Signal Process. 32 (6) (1984) 1109–1121.
[3] S.H. Jensen, P.C. Hansen, S.D. Hansen, J.A. Sørensen, Reduction of broad-band noise in speech by truncated QSVD, IEEE Trans. Speech Audio Process. 3 (6) (1995) 439–448.
[4] V. Bray, M. Valente, Can omni-directional hearing aids improve speech understanding in noise? Audiology Online, available: http://www.audiologyonline.com/articles/can-omni-directional-hearing-aids-1214, 2001.
[5] J. Benesty, S. Makino, J. Chen, Speech Enhancement, Springer-Verlag, Berlin, Germany, 2005.
[6] P. Loizou, Speech Enhancement: Theory and Practice, CRC Press, Boca Raton, Florida, 2007.
[7] J. Benesty, J. Chen, Y. Huang, I. Cohen, Noise Reduction in Speech Processing, Springer-Verlag, Berlin, Germany, 2009.
[8] J. Benesty, J. Chen, E. Habets, Speech Enhancement in the STFT Domain, Springer-Verlag, Berlin, Germany, 2011.
[9] P. Kechichian, S. Srinivasan, Model-based speech enhancement using a bone-conducted signal, J. Acoust. Soc. Am. 131 (3) (2012) EL262–EL267.
[10] S. Srinivasan, D. Hanumantha, R. Naidu, Speech enhancement using a generic noise codebook, J. Acoust. Soc. Am. 132 (2) (2012) EL161–EL167.
[11] J. Chen, J. Benesty, Y. Huang, S. Doclo, New insights into the noise reduction Wiener filter, IEEE Trans. Audio Speech Lang. Process. 14 (4) (2006) 1218–1234.
[12] J. Chen, J. Benesty, Y. Huang, E.J. Diethorn, Fundamentals of noise reduction, in: J. Benesty, M.M. Sondhi, Y. Huang (Eds.), Springer Handbook of Speech Processing, Springer-Verlag, Berlin, Germany, 2007, pp. 843–871.
[13] J.R. Jensen, J. Benesty, M.G. Christensen, S.H. Jensen, Non-causal time-domain filters for single-channel noise reduction, IEEE Trans. Speech Audio Process. 20 (5) (2012) 1526–1541.
[14] J. Benesty, J. Chen, Y. Huang, T. Gaensler, Time-domain noise reduction based on an orthogonal decomposition for desired signal extraction, J. Acoust. Soc. Am. 132 (1) (2012) 452–464.
[15] M.R. Weiss, E. Aschkenasy, T.W. Parsons, Processing speech signals to attenuate interference, in: Proc. IEEE Symp. Speech Recognition, 1974, pp. 292–295.
[16] S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust. Speech Signal Process. 27 (2) (1979) 113–120.
[17] R.J. McAulay, M.L. Malpass, Speech enhancement using a soft-decision noise suppression filter, IEEE Trans. Acoust. Speech Signal Process. 28 (2) (1980) 137–145.
[18] P.J. Wolfe, S.J. Godsill, Efficient alternatives to the Ephraim–Malah suppression rule for audio signal enhancement, EURASIP J. Appl. Signal Process. 2003 (10) (2003) 1043–1051.
[19] J. Chen, J. Benesty, Single-channel noise reduction in the STFT domain based on the bifrequency spectrum, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP, 2012, pp. 97–100.
[20] C. Li, 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. 2005 (18) (2005) 2965–2978.
[21] E. Plourde, B. Champagne, Multidimentional STSA estimators for speech enhancement with correlated spectral components, IEEE Trans. Signal Process. 59 (7) (2011) 3013–3024.
[22] J. Capon, High-resolution frequency-wavenumber spectrum analysis, Proc. IEEE 57 (8) (1969) 1408–1418.
[23] J. Benesty, Y. Huang, A single-channel noise reduction MVDR filter, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP, 2011, pp. 273–276.
[24] Y. Huang, J. Benesty, A multi-frame approach to the frequency-domain single-channel noise reduction problem, IEEE Trans. Audio Speech Lang. Process. 20 (4) (2012) 1256–1269.
[25] N.L. Gerr, J.C. Allen, The generalised spectrum and spectral coherence of a harmonizable time series, Digit. Signal Process. 4 (4) (1994) 222–238.
[26] A. Napolitano, Uncertainty in measurements on spectrally correlated stochastic processes, IEEE Trans. Signal Process. 49 (9) (2003) 2172–2191.
[27] ITU-T P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T recommendation P.862.
[28] A. Varga, H.J.M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems, Speech Commun. 12 (3) (1993) 247–251.
[29] I. Cohen, B. Berdugo, Noise estimation by minima controlled recursive averaging for robust speech enhancement, IEEE Signal Process. Lett. 9 (1) (2002) 12–15.
[30] I. Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging, IEEE Trans. Speech Audio Process. 11 (5) (2003) 466–475.
[31] R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics, Speech Commun. 9 (5) (2001) 504–512.
[32] G. Doblinger, Computationally efficient speech enhancement by spectral minima tracking in subbands, in: Proc. Euro. Conf. Speech Commun. Tech., 1995, pp. 1513–1516.
[33] H.G. Hirsch, C. Ehrlicher, Noise estimation techniques for robust speech recognition, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., ICASSP, vol. 1, 1995, pp. 153–156.
[34] J. Chen, J. Benesty, Y. Huang, Time delay estimation in room acoustic environments: an overview, EURASIP J. Appl. Signal Process. 2006 (2006) 1–19.
[35] C. Meyer, Matrix Analysis and Applied Linear Algebra Book and Solutions Manual, Society for Industrial and Applied Mathematics, 2000.

Hai Huang received his B.S. and M.S. degrees from Department of Electronic and Communication Engineering at Northwestern Polytechnical University (NWPU), Xi'an, China, in 2009 and 2012 respectively. He is currently a Ph.D. student in Information and Communication Engineering at NWPU. His research interests are in speech enhancement, noise reduction, echo cancellation, and echo suppression for hands-free speech communications.

Liheng Zhao received the B.E. degree in automation, and the Ph.D. degree in pattern recognition and intelligent system from University of Science and Technology of China, Hefei, China, in 2007 and 2012, respectively. He is currently a Postdoctoral Fellow at INRS-EMT, University of Quebec, in Montreal, Canada. His research interests include microphone array signal processing, audio signal processing, and speech technologies.

Jingdong Chen received the Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences in 1998. From 1998 to 1999, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he conducted research on speech synthesis, speech analysis, as well as objective measurements for evaluating speech synthesis. He then joined the Griffith University, Brisbane, Australia, where he engaged in research on robust speech recognition and signal processing. From 2000 to 2001, he worked at ATR Spoken Language Translation Research Laboratories on robust speech recognition and speech enhancement. From 2001 to 2009, he was a Member of Technical Staff at Bell Laboratories, Murray Hill, New Jersey, working on acoustic signal processing for telecommunications. He subsequently joined WeVoice Inc. in New Jersey, serving as the Chief Scientist. He is currently a Professor at the Northwestern Polytechnical University in Xi'an, China. His research interests include acoustic signal processing, adaptive signal processing, speech enhancement, adaptive noise/echo control, microphone array signal processing, signal separation, and speech communication.

Dr. Chen received the 2008 Best Paper Award from the IEEE Signal Processing Society (with Benesty, Huang, and Doclo), the best paper award from the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2011 (with Benesty), the Bell Labs Role Model Teamwork Award twice, respectively, in 2009 and 2007, the NASA Tech Brief Award twice, respectively, in 2010 and 2009, the Japan Trust International Research Grant from the Japan Key Technology Center in 1998, and the Young Author Best Paper Award from the 5th National Conference on Man–Machine Speech Communications in 1998.

Jacob Benesty was born in 1963. He received a Master degree in microwaves from Pierre & Marie Curie University, France, in 1987, and a Ph.D. degree in control and signal processing from Orsay University, France, in April 1991. During his Ph.D. (from Nov. 1989 to Apr. 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecomunications (CNET), Paris, France. From January 1994 to July 1995, he worked at Telecom Paris University on multichannel adaptive filters and acoustic echo cancellation. From October 1995 to May 2003, he was first a Consultant and then a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ, USA. In May 2003, he joined the University of Quebec, INRS-EMT, in Montreal, Quebec, Canada, as a Professor. He is also an Adjunct Professor at Aalborg University, in Denmark and at Northwestern Polytechnical University, Xi'an, Shaanxi, in China.

His research interests are in signal processing, acoustic signal processing, and multimedia communications. He is the inventor of many important technologies. In particular, he was the lead researcher at Bell Labs who conceived and designed the world-first real-time hands-free full-duplex stereophonic teleconferencing system. Also, he conceived and designed the world-first PC-based multi-party hands-free full-duplex stereo conferencing system over IP networks.

He was the co-chair of the 1999 International Workshop on Acoustic Echo and Noise Control and the general co-chair of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. He is the recipient, with Morgan and Sondhi, of the IEEE Signal Processing Society 2001 Best Paper Award. He is the recipient, with Chen, Huang, and Doclo, of the IEEE Signal Processing Society 2008 Best Paper Award. He is also the co-author of a paper for which Huang received the IEEE Signal Processing Society 2002 Young Author Best Paper Award. In 2010, he received the "Gheorghe Cartianu Award" from the Romanian Academy. In 2011, he received the Best Paper Award from the IEEE WASPAA for a paper that he co-authored with Chen.