New Insights Into the Noise Reduction Wiener Filter

Jingdong Chen, *Member, IEEE*, Jacob Benesty, *Senior Member, IEEE*, Yiteng (Arden) Huang, *Member, IEEE*, and Simon Doclo, *Member, IEEE*

***Abstract*—The problem of noise reduction has attracted a considerable amount of research attention over the past several decades. Among the numerous techniques that were developed, the optimal Wiener filter can be considered as one of the most fundamental noise reduction approaches, which has been delineated in different forms and adopted in various applications. Although it is not a secret that the Wiener filter may cause some detrimental effects to the speech signal (appreciable or even significant degradation in quality or intelligibility), few efforts have been reported to show the inherent relationship between noise reduction and speech distortion. By defining a speech-distortion index to measure the degree to which the speech signal is deformed and two noise-reduction factors to quantify the amount of noise being attenuated, this paper studies the quantitative performance behavior of the Wiener filter in the context of noise reduction. We show that in the single-channel case the *a posteriori* signal-to-noise ratio (SNR) (defined after the Wiener filter) is greater than or equal to the *a priori* SNR (defined before the Wiener filter), indicating that the Wiener filter is always able to achieve noise reduction. However, the amount of noise reduction is in general proportional to the amount of speech degradation. This may seem discouraging as we always expect an algorithm to have maximal noise reduction without much speech distortion. Fortunately, we show that speech distortion can be better managed in three different ways. If we have some *a priori* knowledge (such as the linear prediction coefficients) of the clean speech signal, this *a priori* knowledge can be exploited to achieve noise reduction while maintaining a low level of speech distortion. When no *a priori* knowledge is available, we can still achieve a better control of noise reduction and speech distortion by properly manipulating the Wiener filter, resulting in a suboptimal Wiener filter. In case that we have multiple microphone sensors, the multiple observations of the speech signal can be used to reduce noise with less or even no speech distortion.**

***Index Terms*—Microphone arrays, noise reduction, speech distortion, Wiener filter.**

I. Introduction

SINCE we are living in a natural environment where noise is inevitable and ubiquitous, speech signals are generally immersed in acoustic ambient noise and can seldom be recorded in pure form. Therefore, it is essential for speech processing and communication systems to apply effective noise reduction/speech enhancement techniques in order to extract the desired speech signal from its corrupted observations.

Noise reduction techniques have a broad range of applications, from hearing aids to cellular phones, voice-controlled systems, multiparty teleconferencing, and automatic speech recognition (ASR) systems. The choice between using and not using a noise reduction technique may have a significant impact on the functioning of these systems. In multiparty conferencing, for example, the background noise picked up by the microphone at each point of the conference combines additively at the network bridge with the noise signals from all other points. The loudspeaker at each location of the conference therefore reproduces the combined sum of the noise processes from all other locations. Clearly, this problem can be extremely serious if the number of conferees is large, and without noise reduction, communication is almost impossible in this context.

Noise reduction is a very challenging and complex problem due to several reasons. First of all, the nature and the characteristics of the noise signal change significantly from application to application, and moreover vary in time. It is therefore very difficult—if not impossible—to develop a versatile algorithm that works in diversified environments. Secondly, the objective of a noise reduction system is heavily dependent on the specific context and application. In some scenarios, for example, we want to increase the intelligibility or improve the overall speech perception quality, while in other scenarios, we expect to ameliorate the accuracy of an ASR system, or simply reduce the listeners' fatigue. It is very hard to satisfy all objectives at the same time. In addition, the complex characteristics of speech and the broad spectrum of constraints make the problem even more complicated.

Research on noise reduction/speech enhancement can be traced back to 40 years ago with 2 patents by Schroeder [1], [2] where an analog implementation of the spectral magnitude subtraction method was described. Since then it has become an area of active research. Over the past several decades, researchers and engineers have approached this challenging problem by exploiting different facets of the properties of the speech and noise signals. Some good reviews of such efforts can be found in [3]–[7]. Principally, the solutions to the problem can be classified from the following points of view.

- The number of channels available for enhancement; i.e., single-channel and multichannel techniques.
- How the noise is mixed to the speech; i.e., additive noise, multiplicative noise, and convolutional noise.
- Statistical relationship between the noise and speech; i.e., uncorrelated or even independent noise, and correlated noise (such as echo and reverberation).
- How the processing is carried out; i.e., in the time domain or in the frequency domain.

Manuscript received December 20, 2004; revised September 2, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Li Deng.

J. Chen and Y. Huang are with the Bell Labs, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: jingdong@research.bell-labs.com; arden@research.bell-labs.com).

J. Benesty is with the Université du Québec, INRS-EMT, Montréal, QC, H5A 1K6, Canada (e-mail: benesty@emt.inrs.ca).

S. Doclo is with the Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Leuven 3001, Belgium (e-mail: simon.doclo@esat.kuleuven.be).

Digital Object Identifier 10.1109/TSA.2005.860851

In general, the more microphones are available, the easier the task of noise reduction. For example, when multiple realizations of the signal can be accessed, beamforming, source separation, or spatio-temporal filtering techniques can be applied to extract the desired speech signal or to attenuate the unwanted noise [8]–[13].

If we have two microphones, where the first microphone picks up the noisy signal, and the second microphone is able to measure the noise field, we can use the second microphone signal as a noise reference and eliminate the noise in the first microphone by means of adaptive noise cancellation. However, in most situations, such as mobile communications, only one microphone is available. In this case, noise reduction techniques need to rely on assumptions about the speech and noise signals, or need to exploit aspects of speech perception, speech production, or a speech model. A common assumption is that the noise is additive and slowly varying, so that the noise characteristics estimated in the absence of speech can be used subsequently in the presence of speech. If in reality this premise does not hold, or only partially holds, the system will either have less noise reduction, or introduce more speech distortion.

Even with the limitations outlined above, single-channel noise reduction has attracted a tremendous amount of research attention because of its wide range of applications and relatively low cost. A variety of approaches have been developed, including *Wiener filter* [3], [14]–[19], *spectral* or *cepstral restoration* [17], [20]–[27], *signal subspace* [28]–[35], *parametric-model-based method* [36]–[38], and *statistical-model-based method* [5], [39]–[46].

Most of these algorithms were developed independently of each other and generally their noise reduction performance was evaluated by assessing the improvement of signal-to-noise ratio (SNR), subjective speech quality, or ASR performance (when the ASR system is trained in clean conditions and additive noise is the only distortion source). Almost with no exception, these algorithms achieve noise reduction by introducing some distortion to the speech signal. Some algorithms, such as the subspace method, are even explicitly formulated based on the tradeoff between noise reduction and speech distortion. However, so far, few efforts have been devoted to analyzing such a tradeoff behavior even though it is a very important issue. In this paper, we attempt to provide an analysis about the compromise between noise reduction and speech distortion. On one hand, such a study may offer us some insight into the range of existing algorithms that can be employed in practical noisy environments. On the other hand, a good understanding may help us to find new algorithms that can work more effectively than the existing ones.

Since there are so many algorithms in the literature, it is extremely difficult—if not impossible—to find a universal analytical tool that can be applied to any algorithm. In this paper, we choose the Wiener filter as the basis since it is one of the most fundamental approaches, and many algorithms are closely connected to this technique. For example, the minimum-mean-square-error (MMSE) estimator presented in [21], which belongs to the category of spectral restoration, converges to the Wiener filter at a high SNR. In addition, it is widely known that the Kalman filter is tightly related to the Wiener filter.

Starting from optimal Wiener filtering theory, we introduce a speech-distortion index to measure the degree to which the speech signal is deformed and two noise-reduction factors to quantify the amount of noise being attenuated. We then show that for the single-channel Wiener filter, the amount of noise reduction is in general proportional to the amount of speech degradation, implying that when the noise reduction is maximized, the speech distortion is maximized as well.

Depending on the nature of the application, some practical noise-reduction systems require very high-quality speech, but can tolerate a certain amount of residual noise, whereas other systems require the speech signal to be as clean as possible, but may allow some degree of speech distortion. Therefore, it is necessary that we have some management scheme to control the compromise between noise reduction and speech distortion in the context of Wiener filtering. To this end, we discuss three approaches. The first approach leads to a suboptimal filter where a parameter is introduced to control the tradeoff between speech distortion and noise reduction. The second approach leads to the well-known parametric-model-based noise reduction technique, where an AR model is exploited to achieve noise reduction, while maintaining a low level of speech distortion. The third approach pertains to a multichannel approach where spatio-temporal filtering techniques are employed to obtain noise reduction with less or even no speech distortion.

II. Estimation of the Clean Speech Samples

We consider a zero-mean clean speech signal $x(n)$ contaminated by a zero-mean noise process $v(n)$ [white or colored but uncorrelated with $x(n)$], so that the noisy speech signal at the discrete time sample n is

$$y(n) = x(n) + v(n). \tag{1}$$

Define the error signal between the clean speech sample at time n and its estimate

$$e_x(n) \triangleq x(n) - \hat{x}(n) = x(n) - \mathbf{h}^T\mathbf{y}(n) \tag{2}$$

where superscript ^{T} denotes transpose of a vector or a matrix,

$$\mathbf{h} = [\, h_0 \quad h_1 \quad \dots \quad h_{L-1} \,]^T$$

is an FIR filter of length L, and

$$\mathbf{y}(n) = [\, y(n) \quad y(n-1) \quad \dots \quad y(n-L+1) \,]^T$$

is a vector containing the L most recent samples of the observation signal $y(n)$.

We now can write the mean-square error (MSE) criterion

$$J_x(\mathbf{h}) \triangleq E\left\{e_x^2(n)\right\} \tag{3}$$

where $E\{\cdot\}$ denotes mathematical expectation. The optimal estimate $\hat{x}_{\mathrm{o}}(n)$ of the clean speech sample $x(n)$ tends to contain less noise than the observation sample $y(n)$, and the optimal filter that forms $\hat{x}_{\mathrm{o}}(n)$ is the Wiener filter which is obtained as follows:

$$\mathbf{h}_{\mathrm{o}} = \arg\min_{\mathbf{h}} J_x(\mathbf{h}). \tag{4}$$

Consider the particular filter

$$\mathbf{u}_1 = [\, 1 \quad 0 \quad \dots \quad 0 \,]^T.$$

This means that the observed signal $y(n)$ will pass this filter unaltered (no noise reduction), thus the corresponding MSE is

$$J_x(\mathbf{u}_1) = E\left\{\left[x(n) - \mathbf{u}_1^T\mathbf{y}(n)\right]^2\right\} = E\left\{[x(n) - y(n)]^2\right\} = E\left\{v^2(n)\right\} = \sigma_v^2. \quad (5)$$

In principle, for the optimal filter $\mathbf{h}_o$, we should have

$$J_x(\mathbf{h}_o) < J_x(\mathbf{u}_1) = \sigma_v^2. \quad (6)$$

In other words, the Wiener filter will be able to reduce the level of noise in the noisy speech signal $y(n)$.

From (4), we easily find the Wiener–Hopf equation

$$\mathbf{R}_y\mathbf{h}_o = \mathbf{p} \quad (7)$$

where

$$\mathbf{R}_y = E\left\{\mathbf{y}(n)\mathbf{y}^T(n)\right\} \quad (8)$$

is the correlation matrix of the observed signal $y(n)$ and

$$\mathbf{p} = E\left\{\mathbf{y}(n)x(n)\right\} \quad (9)$$

is the cross-correlation vector between the noisy and clean speech signals. However, $x(n)$ is unobservable; as a result, an estimation of $\mathbf{p}$ may seem difficult to obtain. But

$$\begin{aligned}\mathbf{p} &= E\left\{\mathbf{y}(n)x(n)\right\} = E\left\{\mathbf{y}(n)\left[y(n) - v(n)\right]\right\}\\ &= E\left\{\mathbf{y}(n)y(n)\right\} - E\left\{\left[\mathbf{x}(n) + \mathbf{v}(n)\right]v(n)\right\}\\ &= E\left\{\mathbf{y}(n)y(n)\right\} - E\{\mathbf{v}(n)v(n)\}\\ &= \mathbf{r}_y - \mathbf{r}_v. \end{aligned} \quad (10)$$

Now $\mathbf{p}$ depends on the correlation vectors $\mathbf{r}_y$ and $\mathbf{r}_v$. The vector $\mathbf{r}_y$ (which is also the first column of $\mathbf{R}_y$) can be easily estimated during speech and noise periods while $\mathbf{r}_v$ can be estimated during noise-only intervals assuming that the statistics of the noise do not change much with time.

Using (10) and the fact that $\mathbf{u}_1 = \mathbf{R}_y^{-1}\mathbf{r}_y$, we obtain the optimal filter

$$\begin{aligned}\mathbf{h}_o &= \mathbf{u}_1 - \mathbf{R}_y^{-1}\mathbf{r}_v = \left[\mathbf{I} - \mathbf{R}_y^{-1}\mathbf{R}_v\right]\mathbf{u}_1\\ &= \left[\frac{\mathbf{I}}{\mathrm{SNR}} + \tilde{\mathbf{R}}_v^{-1}\tilde{\mathbf{R}}_x\right]^{-1}\tilde{\mathbf{R}}_v^{-1}\tilde{\mathbf{R}}_x\mathbf{u}_1 \end{aligned} \quad (11)$$

where

$$\mathrm{SNR} \triangleq \frac{\sigma_x^2}{\sigma_v^2} \quad (12)$$

is the signal-to-noise ratio, $\mathbf{I}$ is the identity matrix, and

$$\tilde{\mathbf{R}}_x \triangleq \frac{\mathbf{R}_x}{\sigma_x^2},$$
$$\tilde{\mathbf{R}}_v \triangleq \frac{\mathbf{R}_v}{\sigma_v^2}.$$

We have

$$\lim_{\mathrm{SNR}\to\infty}\mathbf{h}_o = \mathbf{u}_1 \quad (13)$$
$$\lim_{\mathrm{SNR}\to 0}\mathbf{h}_o = \mathbf{0} \quad (14)$$

where $\mathbf{0}$ has the same size as $\mathbf{h}_o$ and consists of all zeros. The minimum MSE (MMSE) is

$$J_x(\mathbf{h}_o) = \sigma_x^2 - \mathbf{p}^T\mathbf{h}_o = \sigma_v^2 - \mathbf{r}_v^T\mathbf{R}_y^{-1}\mathbf{r}_v = \mathbf{r}_v^T\mathbf{h}_o. \quad (15)$$

We see clearly from the previous expression that $J_x(\mathbf{h}_o) < J_x(\mathbf{u}_1)$; therefore, noise reduction is possible.

The normalized MMSE is

$$\tilde{J}_x(\mathbf{h}_o) \triangleq \frac{J_x(\mathbf{h}_o)}{J_x(\mathbf{u}_1)} = \frac{J_x(\mathbf{h}_o)}{\sigma_v^2} \quad (16)$$

and $0 < \tilde{J}_x(\mathbf{h}_o) < 1$.

III. Estimation of the Noise Samples

In this section, we will estimate the noise samples from the observations $y(n)$. Define the error signal between the noise sample at time n and its estimate

$$e_v(n) \triangleq v(n) - \hat{v}(n) = v(n) - \mathbf{g}^T\mathbf{y}(n) \quad (17)$$

where

$$\mathbf{g} = \begin{bmatrix} g_0 & g_1 & \cdots & g_{L-1}\end{bmatrix}^T$$

is an FIR filter of length L. The MSE criterion associated with (17) is

$$J_v(\mathbf{g}) \triangleq E\left\{e_v^2(n)\right\}. \quad (18)$$

The estimation of $v(n)$ in the MMSE sense will tend to attenuate the clean speech.

The minimization of (18) leads to the Wiener–Hopf equation

$$\begin{aligned}\mathbf{g}_o &= \mathbf{R}_y^{-1}\mathbf{r}_v = \mathbf{R}_y^{-1}\mathbf{R}_v\mathbf{u}_1\\ &= \left[\mathrm{SNR}\cdot\mathbf{I} + \tilde{\mathbf{R}}_x^{-1}\tilde{\mathbf{R}}_v\right]^{-1}\tilde{\mathbf{R}}_x^{-1}\tilde{\mathbf{R}}_v\mathbf{u}_1. \end{aligned} \quad (19)$$

We have

$$\lim_{\mathrm{SNR}\to\infty}\mathbf{g}_o = \mathbf{0} \quad (20)$$
$$\lim_{\mathrm{SNR}\to 0}\mathbf{g}_o = \mathbf{u}_1. \quad (21)$$

The MSE for the particular filter $\mathbf{u}_1$ (no clean speech reduction) is

$$J_v(\mathbf{u}_1) = E\left\{x^2(n)\right\} = \sigma_x^2. \quad (22)$$

Therefore, the MMSE and the normalized MMSE are, respectively,

$$J_v(\mathbf{g}_o) = \sigma_v^2 - \mathbf{r}_v^T\mathbf{R}_y^{-1}\mathbf{r}_v = \sigma_v^2 - \mathbf{r}_v^T\mathbf{g}_o, \quad (23)$$
$$\tilde{J}_v(\mathbf{g}_o) \triangleq \frac{J_v(\mathbf{g}_o)}{J_v(\mathbf{u}_1)} = \frac{J_v(\mathbf{g}_o)}{\sigma_x^2}. \quad (24)$$

Since $J_v(\mathbf{g}_o) < J_v(\mathbf{u}_1)$, the Wiener filter will be able to reduce the level of the clean speech in the signal $y(n)$. As a result, $0 < \tilde{J}_v(\mathbf{g}_o) < 1$.

In Section IV, we will see that while the normalized MMSE, $\tilde{J}_x(\mathbf{h}_o)$, of the clean speech estimation plays a key role in noise reduction, the normalized MMSE, $\tilde{J}_v(\mathbf{g}_o)$, of the noise process estimation plays a key role in speech distortion.

IV. Important Relationships Between Noise Reduction and Speech Distortion

Obviously, there are some important relationships between the estimation of the clean speech and noise samples. From (11) and (19), we get a relation between the two optimal filters

$$\mathbf{h}_{\mathrm{o}} = \mathbf{u}_1 - \mathbf{g}_{\mathrm{o}}. \tag{25}$$

In fact, minimizing $J_x(\mathbf{h})$ or $J_v(\mathbf{u}_1 - \mathbf{h})$ with respect to $\mathbf{h}$ is equivalent. In the same manner, minimizing $J_v(\mathbf{g})$ or $J_x(\mathbf{u}_1 - \mathbf{g})$ with respect to $\mathbf{g}$ is the same thing. At the optimum, we have

$$\begin{aligned} e_{x,o}(n) &= x(n) - \mathbf{h}_{\mathrm{o}}^T \mathbf{y}(n) \\ &= x(n) - [\mathbf{u}_1 - \mathbf{g}_{\mathrm{o}}]^T [\mathbf{x}(n) + \mathbf{v}(n)] \\ &= -v(n) + \mathbf{g}_{\mathrm{o}}^T \mathbf{y}(n) = -e_{v,o}(n). \end{aligned} \tag{26}$$

From (15) and (23), we see that the two MMSEs are equal

$$J_x(\mathbf{h}_{\mathrm{o}}) = J_v(\mathbf{g}_{\mathrm{o}}). \tag{27}$$

However, the normalized MMSE's are not, in general. Indeed, we have a relation between the two

$$\begin{aligned} \tilde{J}_v(\mathbf{g}_{\mathrm{o}}) &= \frac{J_v(\mathbf{g}_{\mathrm{o}})}{\sigma_x^2} = \frac{J_x(\mathbf{h}_{\mathrm{o}})}{\sigma_x^2} \\ &= \frac{\sigma_v^2}{\sigma_x^2} \frac{J_x(\mathbf{h}_{\mathrm{o}})}{\sigma_v^2} = \frac{\tilde{J}_x(\mathbf{h}_{\mathrm{o}})}{\mathrm{SNR}}. \end{aligned} \tag{28}$$

So the only situation where the two normalized MMSE's are equal is when the SNR is equal to 1. For $\mathrm{SNR} < 1$, $\tilde{J}_x(\mathbf{h}_{\mathrm{o}}) < \tilde{J}_v(\mathbf{g}_{\mathrm{o}})$ and for $\mathrm{SNR} > 1$, $\tilde{J}_v(\mathbf{g}_{\mathrm{o}}) < \tilde{J}_x(\mathbf{h}_{\mathrm{o}})$. Also, $\tilde{J}_x(\mathbf{h}_{\mathrm{o}}) < \mathrm{SNR}$ and $\tilde{J}_v(\mathbf{g}_{\mathrm{o}}) < 1/\mathrm{SNR}$.

It can easily be verified that

$$J_v(\mathbf{h}_{\mathrm{o}}) = J_x(\mathbf{g}_{\mathrm{o}}) = \sigma_y^2 - 3J_x(\mathbf{h}_{\mathrm{o}}) \tag{29}$$

which implies that $J_x(\mathbf{h}_{\mathrm{o}}) < \sigma_y^2/3$. We already know that $J_x(\mathbf{h}_{\mathrm{o}}) < \sigma_v^2$ and $J_x(\mathbf{h}_{\mathrm{o}}) < \sigma_x^2$.

The optimal estimation of the clean speech, in the Wiener sense, is in fact what we call noise reduction

$$\hat{x}_{\mathrm{o}}(n) = \mathbf{h}_{\mathrm{o}}^T \mathbf{y}(n) \tag{30}$$

or equivalently, if the noise is estimated first

$$\hat{v}_{\mathrm{o}}(n) = \mathbf{g}_{\mathrm{o}}^T \mathbf{y}(n) \tag{31}$$

we can use this estimate to reduce the noise from the observed signal

$$\hat{x}_{\mathrm{o}}(n) = y(n) - \hat{v}_{\mathrm{o}}(n). \tag{32}$$

The power of the estimated clean speech signal with the optimal Wiener filter is

$$\begin{aligned} E\left\{\hat{x}_{\mathrm{o}}^2(n)\right\} &= \mathbf{h}_{\mathrm{o}}^T \mathbf{R}_y \mathbf{h}_{\mathrm{o}} = \sigma_x^2 - J_x(\mathbf{h}_{\mathrm{o}}) \\ &= \mathbf{h}_{\mathrm{o}}^T \mathbf{R}_x \mathbf{h}_{\mathrm{o}} + \mathbf{h}_{\mathrm{o}}^T \mathbf{R}_v \mathbf{h}_{\mathrm{o}} \end{aligned} \tag{33}$$

which is the sum of two terms. The first one is the power of the attenuated clean speech and the second one is the power of the residual noise (always greater than zero). While noise reduction is feasible with the Wiener filter, expression (33) shows that the price to pay for this is also a reduction of the clean speech [by a quantity equal to $J_x(\mathbf{h}_{\mathrm{o}}) + \mathbf{h}_{\mathrm{o}}^T \mathbf{R}_v \mathbf{h}_{\mathrm{o}}$ and this implies distortion], since $\mathbf{h}_{\mathrm{o}}^T \mathbf{R}_x \mathbf{h}_{\mathrm{o}} < \sigma_x^2$. In other words, the power of the attenuated clean speech signal is, obviously, always smaller than the power of the clean speech itself; this means that parts of the clean speech are attenuated in the process and as a result, distortion is unavoidable with this approach.

We now define the speech-distortion index due to the optimal filtering operation as

$$\begin{aligned} \upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}}) &\triangleq \frac{E\left\{\left[x(n) - \mathbf{h}_{\mathrm{o}}^T \mathbf{x}(n)\right]^2\right\}}{\sigma_x^2} \\ &= \frac{\mathbf{g}_{\mathrm{o}}^T \mathbf{R}_x \mathbf{g}_{\mathrm{o}}}{\sigma_x^2} = \frac{1}{\mathrm{SNR}} \left[\tilde{J}_x(\mathbf{h}_{\mathrm{o}}) - \mathbf{h}_{\mathrm{o}}^T \tilde{\mathbf{R}}_v \mathbf{h}_{\mathrm{o}}\right] \\ &< \tilde{J}_v(\mathbf{g}_{\mathrm{o}}). \end{aligned} \tag{34}$$

Clearly, this index is always between 0 and 1 for the optimal filter. Also

$$\lim_{\mathrm{SNR} \to 0} \upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}}) = 1 \tag{35}$$

$$\lim_{\mathrm{SNR} \to \infty} \upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}}) = 0. \tag{36}$$

So when $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$ is close to 1, the speech signal is highly distorted and when $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$ is near 0, the speech signal is lowly distorted. We deduce that for low SNRs, the Wiener filter can have a disastrous effect on the speech signal.

Similarly, we define the noise-reduction factor due to the Wiener filter as

$$\begin{aligned} \xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}}) &\triangleq \frac{\sigma_v^2}{E\left\{\left[\mathbf{h}_{\mathrm{o}}^T \mathbf{v}(n)\right]^2\right\}} \\ &= \frac{\sigma_v^2}{\mathbf{h}_{\mathrm{o}}^T \mathbf{R}_v \mathbf{h}_{\mathrm{o}}} = \frac{1}{\mathrm{SNR}\left[\tilde{J}_v(\mathbf{g}_{\mathrm{o}}) - \mathbf{g}_{\mathrm{o}}^T \tilde{\mathbf{R}}_x \mathbf{g}_{\mathrm{o}}\right]} \\ &> \frac{1}{\tilde{J}_x(\mathbf{h}_{\mathrm{o}})} \end{aligned} \tag{37}$$

and $\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}}) > 1$. The greater is $\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})$, the more noise reduction we have. Also

$$\lim_{\mathrm{SNR} \to 0} \xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}}) = \infty, \tag{38}$$

$$\lim_{\mathrm{SNR} \to \infty} \xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}}) = 1. \tag{39}$$

Using (34) and (37), we obtain important relations between the speech-distortion index and the noise-reduction factor

$$\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}}) = \frac{1}{\mathrm{SNR}} \left[\tilde{J}_x(\mathbf{h}_{\mathrm{o}}) - \frac{1}{\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})}\right] \tag{40}$$

$$\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}}) = \frac{1}{\mathrm{SNR}\left[\tilde{J}_v(\mathbf{g}_{\mathrm{o}}) - \upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})\right]}. \tag{41}$$

Therefore, for the optimum filter, when the SNR is very large, there is little speech distortion and little noise reduction (which is not really needed in this situation). On the other hand, when the SNR is very small, speech distortion is large as well as noise reduction.

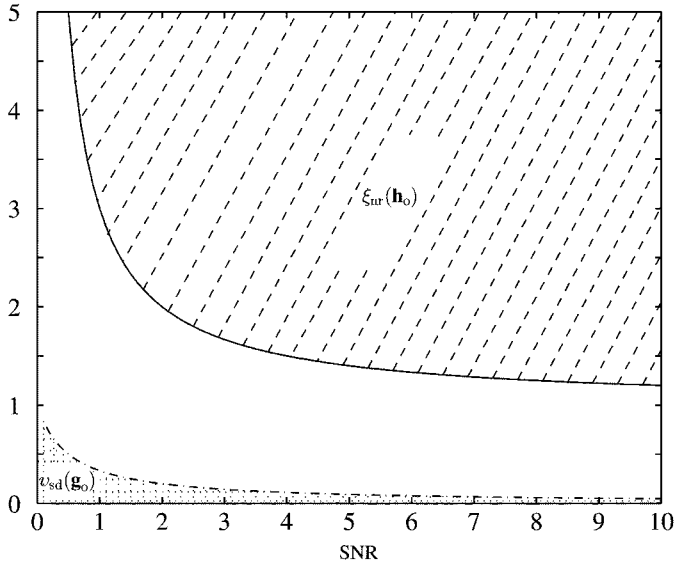

Fig. 1. Illustration of the areas where $\xi_{\rm nr}(\mathbf{h}_{\rm o})$ and $\upsilon_{\rm sd}(\mathbf{g}_{\rm o})$ take their values as a function of the SNR. $\xi_{\rm nr}(\mathbf{h}_{\rm o})$ can take any value above the solid line while $\upsilon_{\rm sd}(\mathbf{g}_{\rm o})$ can take any value under the dotted line.

Another way to examine the noise-reduction performance is to inspect the SNR improvement. Let us define the *a posteriori* SNR, after noise reduction with the Wiener filter as

$$\begin{aligned}\mathrm{SNR}_{\rm o} &\triangleq \frac{\mathbf{h}_{\rm o}^T\mathbf{R}_x\mathbf{h}_{\rm o}}{\mathbf{h}_{\rm o}^T\mathbf{R}_v\mathbf{h}_{\rm o}}\\ &= \mathrm{SNR}\,\frac{\mathbf{h}_{\rm o}^T\tilde{\mathbf{R}}_x\mathbf{h}_{\rm o}}{\mathbf{h}_{\rm o}^T\tilde{\mathbf{R}}_v\mathbf{h}_{\rm o}}\\ &= -1+\mathrm{SNR}\,\xi_{\rm nr}(\mathbf{h}_{\rm o})\left[1-\tilde{J}_v(\mathbf{g}_{\rm o})\right]\\ &= -1+\frac{1-\tilde{J}_v(\mathbf{g}_{\rm o})}{\tilde{J}_v(\mathbf{g}_{\rm o})-\upsilon_{\rm sd}(\mathbf{g}_{\rm o})}. \end{aligned} \tag{42}$$

It can be shown that the *a posteriori* SNR and the *a priori* SNR satisfy $\mathrm{SNR}_{\rm o}\geq\mathrm{SNR}$ (see Appendix), indicating that the Wiener filter is always able to improve the SNR of the noisy speech signal.

Knowing that $\mathrm{SNR}_{\rm o}\geq\mathrm{SNR}$, we can now give the lower bound for $\xi_{\rm nr}(\mathbf{h}_{\rm o})$. As a matter of fact, it follows from (42) that

$$\mathrm{SNR}_{\rm o}=-1+\frac{1-\tilde{J}_v(\mathbf{g}_{\rm o})}{\tilde{J}_v(\mathbf{g}_{\rm o})-\upsilon_{\rm sd}(\mathbf{g}_{\rm o})}\geq\mathrm{SNR}. \tag{43}$$

Since $\upsilon_{\rm sd}(\mathbf{g}_{\rm o})<\tilde{J}_v(\mathbf{g}_{\rm o})$, and $0\leq\upsilon_{\rm sd}(\mathbf{g}_{\rm o})\leq 1$, it can be easily shown that

$$\xi_{\rm nr}(\mathbf{h}_{\rm o})\geq\frac{\mathrm{SNR}+2}{\mathrm{SNR}}. \tag{44}$$

Similarly, we can derive the upper bound for $\upsilon_{\rm sd}(\mathbf{g}_{\rm o})$, i.e.,

$$\upsilon_{\rm sd}(\mathbf{g}_{\rm o})\leq\frac{1}{2\mathrm{SNR}+1}. \tag{45}$$

Fig. 1 illustrates expressions (44) and (45).

We now introduce another index for noise reduction

$$\zeta_{\rm nr}(\mathbf{h}_{\rm o})\triangleq 1-\tilde{J}_x(\mathbf{h}_{\rm o})<1. \tag{46}$$

The closer is $\zeta_{\rm nr}(\mathbf{h}_{\rm o})$ to 1, the more noise reduction we get. This index will be helpful to use in Sections V–VII.

V. Particular Case: White Gaussian Noise

In this section, we assume that the additive noise is white, so that,

$$\mathbf{r}_v=\sigma_v^2\mathbf{u}_1. \tag{47}$$

From (16) and (24), we observe that the two normalized MMSEs are

$$\tilde{J}_x(\mathbf{h}_{\rm o})=h_{\rm o,0} \tag{48}$$

$$\tilde{J}_v(\mathbf{g}_{\rm o})=\frac{1-g_{\rm o,0}}{\mathrm{SNR}}=\frac{h_{\rm o,0}}{\mathrm{SNR}} \tag{49}$$

where $h_{\rm o,0}$ and $g_{\rm o,0}$ are the first components of the vectors $\mathbf{h}_{\rm o}$ and $\mathbf{g}_{\rm o}$, respectively. Clearly, $0<h_{\rm o,0}<1$ and $0<g_{\rm o,0}<1$. Hence, the normalized MMSE $\tilde{J}_x(\mathbf{h}_{\rm o})$ is completely governed by the first element of the Wiener filter $\mathbf{h}_{\rm o}$.

Now, the speech-distortion index and the noise-reduction factor for the optimal filter can be simplified

$$\begin{aligned}\upsilon_{\rm sd}(\mathbf{g}_{\rm o})&=\frac{1}{\mathrm{SNR}}\left[h_{\rm o,0}-\mathbf{h}_{\rm o}^T\mathbf{h}_{\rm o}\right]\\ &=\frac{\mathbf{g}_{\rm o}^T\mathbf{h}_{\rm o}}{\mathrm{SNR}}=\frac{1}{\mathrm{SNR}}\left[g_{\rm o,0}-\mathbf{g}_{\rm o}^T\mathbf{g}_{\rm o}\right], \end{aligned} \tag{50}$$

$$\xi_{\rm nr}(\mathbf{h}_{\rm o})=\frac{1}{\mathbf{h}_{\rm o}^T\mathbf{h}_{\rm o}}. \tag{51}$$

We also deduce from (50) that $h_{\rm o,0}>\mathbf{h}_{\rm o}^T\mathbf{h}_{\rm o}$ and $g_{\rm o,0}>\mathbf{g}_{\rm o}^T\mathbf{g}_{\rm o}$.

We know from linear prediction theory that [47]

$$\mathbf{R}_y\begin{bmatrix}1\\ -\mathbf{a}_y\end{bmatrix}=\begin{bmatrix}E_y\\ \mathbf{0}_{(L-1)\times 1}\end{bmatrix} \tag{52}$$

where $\mathbf{a}_y$ is the forward linear predictor and E_y is the corresponding error energy. Replacing the previous equation in (11), we obtain

$$\mathbf{h}_{\rm o}=\mathbf{u}_1-\sigma_v^2\mathbf{R}_y^{-1}\mathbf{u}_1=\begin{bmatrix}h_{\rm o,0}\\ \frac{\sigma_v^2}{E_y}\mathbf{a}_y\end{bmatrix} \tag{53}$$

where

$$h_{\rm o,0}=\tilde{J}_x(\mathbf{h}_{\rm o})=1-\frac{\sigma_v^2}{E_y}. \tag{54}$$

Equation (53) shows how the Wiener filter is related to the forward predictor of the observed signal $y(n)$. This expression also gives a hint on how to choose the length of the optimal filter $\mathbf{h}_{\rm o}$: it should be equal to the length of the predictor $\mathbf{a}_y$ required to have a good prediction of the observed signal $y(n)$. Equation (54) contains some very interesting information. Indeed, if the clean speech signal is completely predictable, this means that $E_y\approx\sigma_v^2$ and $\tilde{J}_x(\mathbf{h}_{\rm o})\approx 0$. On the other hand, if $x(n)$ is not predictable, we have $E_y\approx\sigma_y^2$ and $\tilde{J}_x(\mathbf{h}_{\rm o})\approx 1-\sigma_v^2/\sigma_y^2$. This implies that the Wiener filter is more efficient to reduce the level of noise for predictable signals than for unpredictable ones.

VI. Better Ways to Manage Noise Reduction and Speech Distortion

For a noise-reduction/speech-enhancement system, we always expect that it can achieve maximal noise reduction without much speech distortion. From the previous section, however, it follows that while noise reduction is maximized with the

optimal Wiener filter, speech distortion is also maximized. One may ask the legitimate question: are there better ways to control the tradeoff between the conflicting requirements of noise reduction and speech distortion? Examining (34), one can see that to control the speech distortion, we need to minimize $E\left\{\left[x(n)-\mathbf{h}_{\mathrm{o}}^{T}\mathbf{x}(n)\right]^{2}\right\}$. This can be achieved in different ways. For example, a speech signal can be modeled as an AR process. If the AR coefficients are known *a priori* or can be estimated from the noisy speech, these coefficients can be exploited to minimize $E\left\{\left[x(n)-\mathbf{h}_{\mathrm{o}}^{T}\mathbf{x}(n)\right]^{2}\right\}$, while simultaneously achieving a reasonable level of noise attenuation. This is often referred to as the parametric-model-based technique [36], [37]. We will not discuss the details of this technique here. Instead, in what follows we will discuss two other approaches to manage noise reduction and speech distortion in a better way.

A. A Suboptimal Filter

Consider the suboptimal filter

$$\mathbf{h}_{\mathrm{s}}=\mathbf{u}_{1}-\mathbf{g}_{\mathrm{s}}=\mathbf{u}_{1}-\alpha\mathbf{g}_{\mathrm{o}} \tag{55}$$

where α is a real number. The MSE of the clean speech estimation corresponding to $\mathbf{h}_{\mathrm{s}}$ is

$$\begin{aligned}J_{x}(\mathbf{h}_{\mathrm{s}})&=E\left\{\left[x(n)-\mathbf{h}_{\mathrm{s}}^{T}y(n)\right]^{2}\right\}\\&=\sigma_{v}^{2}-\alpha(2-\alpha)\mathbf{r}_{v}^{T}\mathbf{R}_{y}^{-1}\mathbf{r}_{v}\end{aligned} \tag{56}$$

and, obviously, $J_{x}(\mathbf{h}_{\mathrm{s}})\geq J_{x}(\mathbf{h}_{\mathrm{o}})$, $\forall\ \alpha$; we have equality for $\alpha=1$. In order to have noise reduction, α must be chosen in such a way that $J_{x}(\mathbf{h}_{\mathrm{s}})<J_{x}(\mathbf{u}_{1})$, therefore

$$0<\alpha<2. \tag{57}$$

We can check that

$$J_{v}(\mathbf{g}_{\mathrm{s}})=E\left\{\left[v(n)-\alpha\mathbf{g}_{\mathrm{o}}^{T}\mathbf{y}(n)\right]^{2}\right\}=J_{x}(\mathbf{h}_{\mathrm{s}}). \tag{58}$$

Let

$$\hat{x}_{\mathrm{s}}(n)=\mathbf{h}_{\mathrm{s}}^{T}\mathbf{y}(n) \tag{59}$$

denote the estimation of the clean speech at time n with respect to $\mathbf{h}_{\mathrm{s}}$. The power of $\hat{x}_{\mathrm{s}}(n)$ is

$$\begin{aligned}E\left\{\hat{x}_{\mathrm{s}}^{2}(n)\right\}&=\mathbf{h}_{\mathrm{s}}^{T}\mathbf{R}_{y}\mathbf{h}_{\mathrm{s}}=\left[\mathbf{u}_{1}-\alpha\mathbf{R}_{y}^{-1}\mathbf{r}_{v}\right]^{T}\left[\mathbf{r}_{y}-\alpha\mathbf{r}_{v}\right]\\&=\sigma_{x}^{2}+(1-2\alpha)\sigma_{v}^{2}+\alpha^{2}\mathbf{r}_{v}^{T}\mathbf{R}_{y}^{-1}\mathbf{r}_{v}\\&=\mathbf{h}_{\mathrm{s}}^{T}\mathbf{R}_{x}\mathbf{h}_{\mathrm{s}}+\mathbf{h}_{\mathrm{s}}^{T}\mathbf{R}_{v}\mathbf{h}_{\mathrm{s}}.\end{aligned} \tag{60}$$

The speech-distortion index corresponding to the filter $\mathbf{h}_{\mathrm{s}}$ is

$$\begin{aligned}\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})&=\frac{E\left\{\left[x(n)-\mathbf{h}_{\mathrm{s}}^{T}\mathbf{x}(n)\right]^{2}\right\}}{\sigma_{x}^{2}}\\&=\alpha^{2}\mathbf{g}_{\mathrm{o}}^{T}\tilde{\mathbf{R}}_{x}\mathbf{g}_{\mathrm{o}}=\alpha^{2}\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}}).\end{aligned} \tag{61}$$

The previous expression shows that the ratio of the speech-distortion indices corresponding to the two filters $\mathbf{g}_{\mathrm{s}}$ and $\mathbf{g}_{\mathrm{o}}$ depends on α only.

In order to have less distortion with the suboptimal filter $\mathbf{h}_{\mathrm{s}}$ than with the Wiener filter $\mathbf{h}_{\mathrm{o}}$, we must find α in such a way that

$$\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})<\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}}) \tag{62}$$

hence, the condition on α should be

$$-1<\alpha<1. \tag{63}$$

Finally, the suboptimal filter $\mathbf{h}_{\mathrm{s}}$ can reduce the level of noise of the observed signal $y(n)$ but with less distortion than the Wiener filter $\mathbf{h}_{\mathrm{o}}$ if α is taken such as

$$0<\alpha<1. \tag{64}$$

For the extreme cases $\alpha=0$ and $\alpha=1$ we obtain respectively $\mathbf{h}_{\mathrm{s}}=\mathbf{u}_{1}$, no noise reduction at all but no additional distortion added, and $\mathbf{h}_{\mathrm{s}}=\mathbf{h}_{\mathrm{o}}$, maximum noise reduction with maximum speech distortion.

Since

$$\begin{aligned}J_{v}(\mathbf{g}_{\mathrm{s}})&=\mathbf{g}_{\mathrm{s}}^{T}\mathbf{R}_{x}\mathbf{g}_{\mathrm{s}}+\mathbf{h}_{\mathrm{s}}^{T}\mathbf{R}_{v}\mathbf{h}_{\mathrm{s}}\\&=\sigma_{x}^{2}\mathbf{g}_{\mathrm{s}}^{T}\tilde{\mathbf{R}}_{x}\mathbf{g}_{\mathrm{s}}+\sigma_{v}^{2}\mathbf{h}_{\mathrm{s}}^{T}\tilde{\mathbf{R}}_{v}\mathbf{h}_{\mathrm{s}}\\&=J_{x}(\mathbf{h}_{\mathrm{s}})\end{aligned} \tag{65}$$

it follows immediately that the speech-distortion index and the noise-reduction factor due to $\mathbf{h}_{\mathrm{s}}$ are

$$\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})=\frac{1}{\mathrm{SNR}}\left[\tilde{J}_{x}(\mathbf{h}_{\mathrm{s}})-\frac{1}{\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})}\right] \tag{66}$$

$$\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})=\frac{\sigma_{v}^{2}}{\mathbf{h}_{\mathrm{s}}^{T}\mathbf{R}_{v}\mathbf{h}_{\mathrm{s}}}=\frac{1}{\mathrm{SNR}\left[\tilde{J}_{v}(\mathbf{g}_{\mathrm{s}})-\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})\right]}. \tag{67}$$

From (61), one can see that $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})/\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})=\alpha^{2}$, which is a function of α only. Unlike $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})/\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$, $\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})/\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})$ does not only depend on α, but on the characteristics of both the speech and noise signal as well.

However, using (56) and (15), we find that

$$\frac{\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})}{\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})}=\frac{1-\tilde{J}_{x}(\mathbf{h}_{\mathrm{s}})}{1-\tilde{J}_{x}(\mathbf{h}_{\mathrm{o}})}=\alpha(2-\alpha). \tag{68}$$

Fig. 2 plots $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})/\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$ and $\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})/\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})$, both as a function of α. We can see that when $\alpha=0.7$, the suboptimal filter achieves 91% of the noise reduction with the Wiener filter, while the speech distortion is only 49% of that of the Wiener filter. In real applications, we may want the system to achieve maximal noise reduction, while keeping the speech distortion as low as possible. If we define a cost function to measure the compromise between the noise reduction and the speech distortion as

$$\begin{aligned}J_{\zeta\upsilon}(\alpha)&\triangleq\frac{\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})}{\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})}-\frac{\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})}{\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})}\\&=2\alpha-2\alpha^{2}.\end{aligned} \tag{69}$$

It is trivial to see that the α that maximizes $J_{\zeta\upsilon}(\alpha)$ is

$$\alpha_{\mathrm{o}}=\arg\max_{\alpha}J_{\zeta\upsilon}(\alpha)=\frac{1}{2}. \tag{70}$$

In this case, the suboptimal filter achieves 75% of the noise reduction with the Wiener filter, while the speech-distortion is

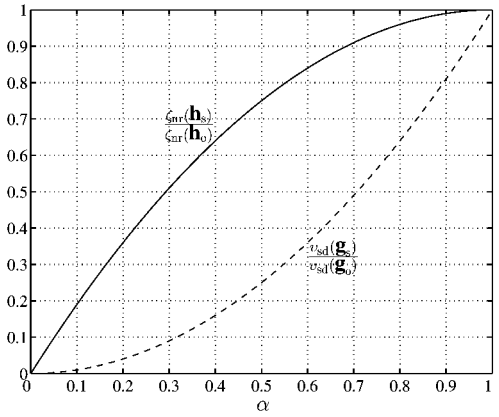

Fig. 2. $v_{\rm sd}(\mathbf{g}_{\rm s})/v_{\rm sd}(\mathbf{g}_{\rm o})$ (dashed line) and $\zeta_{\rm nr}(\mathbf{h}_{\rm s})/\zeta_{\rm nr}(\mathbf{h}_{\rm o})$ (solid line), both as a function of α.

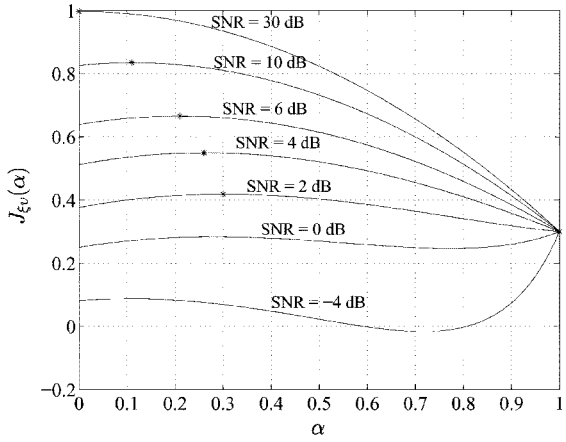

Fig. 3. Illustration of $J_{\xi v}(\alpha)$ in different SNR conditions, where both the signal and the noise are assumed to be Gaussian random processes, and $\beta = 0.7$. The "*" symbol in each curve represents the maximum of $J_{\xi v}(\alpha)$ in the corresponding condition.

only 25% of that of the Wiener filter. The parameter $\alpha_{\rm o}$, which is optimal in terms of the tradeoff between noise reduction and speech distortion, can be used as a guidance in designing a practical noise reduction system for applications like ASR.

Another way to obtain an optimal α is to define a discriminative cost function between $\xi_{\rm nr}(\mathbf{h}_{\rm s})/\xi_{\rm nr}(\mathbf{h}_{\rm o})$ and $v_{\rm sd}(\mathbf{g}_{\rm s})/v_{\rm sd}(\mathbf{g}_{\rm o})$, i.e.,

$$\begin{aligned} J_{\xi v}(\alpha) &\triangleq \frac{\xi_{\rm nr}(\mathbf{h}_{\rm s})}{\xi_{\rm nr}(\mathbf{h}_{\rm o})} - \beta\frac{v_{\rm sd}(\mathbf{g}_{\rm s})}{v_{\rm sd}(\mathbf{g}_{\rm o})} \\ &= \frac{(\mathbf{u}_1 - \mathbf{g}_{\rm o})^T\mathbf{R}_v(\mathbf{u}_1 - \mathbf{g}_{\rm o})}{(\mathbf{u}_1 - \alpha\mathbf{g}_{\rm o})^T\mathbf{R}_v(\mathbf{u}_1 - \alpha\mathbf{g}_{\rm o})} - \beta\alpha^2 \\ &= \frac{\sigma_v^2 + \mathbf{g}_{\rm o}^T\mathbf{R}_v\mathbf{g}_{\rm o} - 2\mathbf{r}_v^T\mathbf{g}_{\rm o}}{\sigma_v^2 + \alpha^2\mathbf{g}_{\rm o}^T\mathbf{R}_v\mathbf{g}_{\rm o} - 2\alpha\mathbf{r}_v^T\mathbf{g}_{\rm o}} - \beta\alpha^2 \end{aligned} \quad (71)$$

where β is an application-dependent constant and determines the relative importance between the improvement in speech distortion and degradation in noise reduction (e.g., in hearing aid applications we may tune this parameter using subjective intelligibility tests).

In contrast to $J_{\zeta v}(\alpha)$, which is a function of α only, the cost function $J_{\xi v}(\alpha)$ does not only depend on α, but on the characteristics of the speech and noise signal as well. Fig. 3 plots $J_{\xi v}(\alpha)$ as a function of α in different SNR conditions, where both the signal and the noise are assumed to be Gaussian random processes and $\beta = 0.7$. This figure shows that for the same α, $J_{\xi v}(\alpha)$ decreases with SNR, indicating that the higher the SNR, the better the suboptimal filter is able to control the compromise between noise reduction and speech distortion.

In order for the suboptimal filter to be able to control the tradeoff between noise reduction and speech distortion, α should be chosen in such a way that $\xi_{\rm nr}(\mathbf{h}_{\rm s})/\xi_{\rm nr}(\mathbf{h}_{\rm o}) > v_{\rm sd}(\mathbf{g}_{\rm s})/v_{\rm sd}(\mathbf{g}_{\rm o})$. Therefore, $J_{\xi v}(\alpha)$ should satisfy $J_{\xi v}(\alpha) > 0$. From Fig. 3, we notice that $J_{\xi v}(\alpha)$ is always positive if the SNR is above 1 (0 dB). When the SNR drops below 1 (0 dB), however, $J_{\xi v}(\alpha)$ may become negative, indicating that the suboptimal filter cannot work reliably in very noisy conditions [when $\rm SNR < 1\ (0\ dB)$].

Fig. 3 also shows the $\alpha_{\rm o}$ that maximizes $J_{\xi v}(\alpha)$ in different SNR situations. It is interesting to see that the $\alpha_{\rm o}$ approaches to 1 when $\rm SNR < 1\ (0\ dB)$, which means that the suboptimal filter converges to the Wiener filter in very low SNR conditions. As we increase the SNR, the $\alpha_{\rm o}$ begins to decrease. It goes to 0 when SNR is increased to 1000 (30 dB). This is understandable. When the SNR is very high, the speech signal is already very clean, so filtering is not really needed. By searching the $\alpha_{\rm o}$ that maximizes (71), the system can adaptively achieve the best tradeoff between noise reduction and speech distortion according to the characteristics of both the speech and noise signals.

B. Noise Reduction With Multiple Microphones

In more and more applications, multiple microphone signals are available. Therefore, it is interesting to investigate deeply the multichannel case, where various techniques such beamforming (nonadaptive and adaptive) and spatial-temporal filtering can be used to achieve noise reduction [13], [50]–[52]. One of the first papers to do so is a paper written by Doclo and Moonen [13], where the optimal filter is derived as well as a general class of estimators. The authors also show how the generalized singular value decomposition can be used in this spatio-temporal technique. In this section, we take a slightly different approach. We will see, in particular, that we can reduce the level of noise without distorting the speech signal.

We suppose that we have a linear array consisting of M microphones whose outputs are denoted as $y_m(n)$, $m = 0, 1, \ldots, M-1$. Without loss of generality, we select microphone 0 as the reference point and to simplify the analysis, we consider the following propagation model:

$$\begin{aligned} y_m(n) &= \beta_m s(n - t - \tau_m) + v_m(n), \\ m &= 0, 1, \ldots, M-1 \end{aligned} \quad (72)$$

where β_m is the attenuation factor (with $\beta_0 = 1$), t is the propagation time from the unknown speech source $s(n)$ to microphone 0, $v_m(n)$ is an additive noise signal at the mth microphone, and τ_m is the relative delay between microphones 0 and m, with $\tau_0 = 0$.

In the following, we assume that the relative delays τ_m, $m = 1, \ldots, M-1$, are known or can easily be estimated. So our first step is the design of a simple delay-and-sum beamformer, which spatially aligns the microphone signals to the direction of the

speech source. From now on, we will work on the time-aligned signals

$$\begin{aligned} z_m(n) &= y_m(n+\tau_m) \\ &= \beta_m s(n-t) + v_m(n+\tau_m), \\ &= x_m(n) + v_m(n+\tau_m), \\ m &= 0, 1, \ldots, M-1. \end{aligned} \tag{73}$$

A straightforward approach for noise reduction is to average the M signals $z_m(n)$

$$\begin{aligned} z_{\mathrm{a}}(n) &= \frac{1}{M}\sum_{m=0}^{M-1} z_m(n) \\ &= \frac{\beta_{\mathrm{a}}}{M} s(n-t) + \frac{1}{M}\sum_{m=0}^{M-1} v_m(n+\tau_m) \end{aligned} \tag{74}$$

where $\beta_{\mathrm{a}} = \sum_{m=0}^{M-1}\beta_m$. If the noises are added incoherently, the output SNR will, in principle, increase [48]. We can further reduce the noise by passing the signal $z_{\mathrm{a}}(n)$ through a Wiener filter as was shown in the previous sections. This approach has, however, two drawbacks. The first one is that, since for $m \neq i$, $E\{v_m(n+\tau_m)v_i(n+\tau_i)\} \neq 0$ in general, the output SNR will not improve that much; and the second one, as we know already, is speech distortion introduced by the optimal filter.

Let us now define the error signal, for the mth microphone, between the clean speech sample $x_m(n)$ and its estimate as

$$\begin{aligned} e_{x_m}(n) &\triangleq x_m(n) - \mathbf{h}_{:m}^T\mathbf{z}(n) \\ &= x_m(n) - \sum_{i=0}^{M-1}\mathbf{h}_{i:m}^T\mathbf{z}_i(n) \end{aligned} \tag{75}$$

where $\mathbf{h}_{i:m}$ are filters of length L and

$$\begin{aligned} \mathbf{h}_{:m} &\triangleq [\,\mathbf{h}_{0:m}^T \quad \mathbf{h}_{1:m}^T \quad \ldots \quad \mathbf{h}_{M-1:m}^T\,]^T, \\ \mathbf{z}(n) &\triangleq [\,\mathbf{z}_0^T(n) \quad \mathbf{z}_1^T(n) \quad \ldots \quad \mathbf{z}_{M-1}^T(n)\,]^T. \end{aligned}$$

Since $\mathbf{z}_i(n) = \beta_i\mathbf{s}(n-t) + \mathbf{v}_i(n+\tau_i)$, (75) becomes

$$\begin{aligned} e_{x_m}(n) &= \mathbf{s}^T(n-t)\left[\beta_m\mathbf{u}_1 - \sum_{i=0}^{M-1}\beta_i\mathbf{h}_{i:m}\right] \\ &\quad - \sum_{i=0}^{M-1}\mathbf{v}_i^T(n+\tau_i)\mathbf{h}_{i:m} \\ &= \mathbf{s}^T(n-t)\left[\beta_m\mathbf{u}_1 - \mathbf{D}\mathbf{h}_{:m}\right] - \mathbf{v}^T(n)\mathbf{h}_{:m} \\ &= e_{s,m}(n) - e_{v,m}(n) \end{aligned} \tag{76}$$

where

$$\begin{aligned} \mathbf{D} &\triangleq [\,\beta_0\mathbf{I} \quad \beta_1\mathbf{I} \quad \ldots \quad \beta_{M-1}\mathbf{I}\,], \\ \mathbf{v}(n) &\triangleq [\,\mathbf{v}_0^T(n+\tau_0) \quad \mathbf{v}_1^T(n+\tau_1) \quad \ldots \quad \mathbf{v}_{M-1}^T(n+\tau_{M-1})\,]^T \end{aligned}$$

Expression (76) is the difference between two error signals; $e_{s,m}(n)$ represents signal distortion and $e_{v,m}(n)$ represents the residual noise. The MSE corresponding to the residual noise with the mth microphone as the reference signal is

$$\begin{aligned} J_{v,m}(\mathbf{h}_{:m}) &= E\left\{e_{v,m}^2(n)\right\} \\ &= \mathbf{h}_{:m}^T E\left\{\mathbf{v}(n)\mathbf{v}^T(n)\right\}\mathbf{h}_{:m} \\ &= \mathbf{h}_{:m}^T\mathbf{R}_v\mathbf{h}_{:m}. \end{aligned} \tag{77}$$

Usually, in the single-channel case, the minimization of the MSE corresponding to the residual noise is done while keeping the signal distortion below a threshold [28]. With no distortion, the optimal filter obtained from this optimization is $\mathbf{u}_1$, hence there is not any noise reduction either. The advantage of multiple microphones is that, actually, we can minimize $J_{v,m}(\mathbf{h}_{:m})$ with the constraint that $\beta_m\mathbf{u}_1 = \mathbf{D}\mathbf{h}_{:m}$ (no speech distortion at all). Therefore, our optimization problem is

$$\min_{\mathbf{h}_{:m}} J_{v,m}(\mathbf{h}_{:m}) \quad \text{subject to } \beta_m\mathbf{u}_1 = \mathbf{D}\mathbf{h}_{:m}. \tag{78}$$

By using a Lagrange multiplier, we easily find the optimal solution

$$\mathbf{h}_{\mathrm{o},:m} = \beta_m\mathbf{R}_v^{-1}\mathbf{D}^T\left[\mathbf{D}\mathbf{R}_v^{-1}\mathbf{D}^T\right]^{-1}\mathbf{u}_1 \tag{79}$$

where we assumed that the noise signals $v_i(n)$ are not perfectly coherent so that $\mathbf{R}_v$ is not singular. This result is very similar to the linearly constrained minimum variance (LCMV) beamformer [51], [52]; but in (79) additional attenuation factors β_m have been included. Note also that this formula has been derived from a different point of view as a multichannel extension of a single-channel MMSE noise-reduction algorithm.

Given the optimal filter $\mathbf{h}_{\mathrm{o},:m}$, we can write the MMSE for the mth microphone as

$$J_{v,m}(\mathbf{h}_{\mathrm{o},:m}) = \beta_m^2\mathbf{u}_1^T\left[\mathbf{D}\mathbf{R}_v^{-1}\mathbf{D}^T\right]^{-1}\mathbf{u}_1. \tag{80}$$

Since we have M microphones, we have M MMSEs as well. The best MMSE from a noise reduction point of view is the smallest one, which is, according to (80), the microphone signal with the smallest attenuation factor.

The attenuation factors β_m can be easily determined, if the power of the noise signals is known, by using the formula

$$\begin{aligned} \beta_m^2 &= \frac{E\left\{z_m^2(n)\right\} - E\left\{v_m^2(n+\tau_m)\right\}}{E\left\{z_0^2(n)\right\} - E\left\{v_0^2(n)\right\}}, \\ m &= 1, 2, \ldots, M-1. \end{aligned} \tag{81}$$

For the particular case where the noise is spatio-temporally white with a power equal to σ_v^2, the MMSE and the normalized MMSE for the mth microphone are, respectively,

$$J_{v,m}(\mathbf{h}_{\mathrm{o},:m}) = \sigma_v^2\frac{\beta_m^2}{\sum_{i=0}^{M-1}\beta_i^2} \tag{82}$$

$$\tilde{J}_{v,m}(\mathbf{h}_{\mathrm{o},:m}) = \frac{\beta_m^2}{\sum_{i=0}^{M-1}\beta_i^2}. \tag{83}$$

As in the single-channel case, we can define for the mth microphone the speech-distortion index as

$$v_{\mathrm{sd}}(\mathbf{h}_{\mathrm{o},:m}) = \frac{E\left\{\left[x_m(n) - \sum_{i=0}^{M-1}\mathbf{h}_{i:m}^T\mathbf{x}_i(n)\right]^2\right\}}{\sigma_{x_m}^2} \tag{84}$$

and the noise-reduction factors as

$$\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o},:m}) = \frac{\sigma_{v_m}^2}{E\left\{\left[\sum_{i=0}^{M-1}\mathbf{h}_{i:m}^T\mathbf{v}_i(n+\tau_m)\right]^2\right\}} \tag{85}$$

$$\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o},:m}) = 1 - \tilde{J}_{v,m}(\mathbf{h}_{\mathrm{o},:m}). \tag{86}$$

With the optimal filter given in (79), for the particular case where the noise is spatio-temporally white with a power equal to σ_v^2, it can be easily shown that

$$\upsilon_{\rm sd}(\mathbf{h}_{{\rm o},:m}) = 0,$$
$$\xi_{\rm nr}(\mathbf{h}_{{\rm o},:m}) = \sum_{i=0}^{M-1} \beta_i^2,$$

and

$$\zeta_{\rm nr}(\mathbf{h}_{{\rm o},:m}) = 1 - \frac{\beta_m^2}{\sum_{i=0}^{M-1} \beta_i^2}.$$

It can be seen that when the number of microphones goes to infinity, $\xi_{\rm nr}(\mathbf{h}_{{\rm o},:m})$ and $\zeta_{\rm nr}(\mathbf{h}_{{\rm o},:m})$ approach, respectively, to infinity and 1, and meanwhile $\upsilon_{\rm sd}(\mathbf{h}_{{\rm o},:m}) = 0$, which indicates that the noise can be completely removed with no signal distortion at all.

VII. Simulation Experiments

By defining a speech-distortion index to measure the degree to which the speech signal is deformed and two noise-reduction factors to quantify the amount of noise being attenuated, we have analytically examined the performance behavior of the Wiener-filter-based noise reduction technique. It is shown that the Wiener filter achieves noise reduction by distorting the speech signal. The more the noise is reduced, the more the speech is distorted. We also proposed several approaches to better manage the tradeoff between noise reduction and speech distortion. To further verify the analysis, and to assess the noise-reduction-and-speech-distortion management schemes, we implemented a time-domain Wiener-filter system. The sampling rate is 8 kHz. The noise signal is estimated in the time-frequency domain using a sequential algorithm presented in [6], [7]. Briefly, this algorithm obtains an estimate of noise using the overlap-add technique on a frame-by-frame basis. The noisy speech signal $y(n)$ is segmented into frames with a frame width of 8 ms and an overlapping factor of 75%. Each frame is then transformed via a DFT into a block of spectral samples. Successive blocks of spectral samples form a two-dimensional time-frequency matrix denoted by $Y_t(j\omega)$, where subscript t is the frame index, denoting the time dimension, and ω is the angular frequency. Then an estimate of the magnitude of the noise spectrum is formulated as shown in (87) at the bottom of the page, where $\alpha_{\rm a}$ and $\alpha_{\rm d}$ are the "attack" and "decay" coefficients respectively. Meanwhile, to reduce its temporal fluctuation, the magnitude of the noisy speech spectrum is smoothed according to the following recursion (see (88), shown at the bottom of the page), where again $\beta_{\rm a}$ is the "attack" coefficient and $\beta_{\rm d}$ the "decay" coefficient. To further reduce the spectral fluctuation, both $\hat{V}_t(\omega)$ and $\bar{Y}_t(\omega)$ are averaged across the neighboring frequency bins around ω. Finally, an estimate of the noise spectrum is obtained by multiplying $\hat{V}_t(\omega)/\bar{Y}_t(\omega)$ with $Y_t(j\omega)$, and the time-domain noise signal is obtained through IDFT and the overlap-add technique. See [6], [7] for a more detailed description of this noise-estimation scheme.

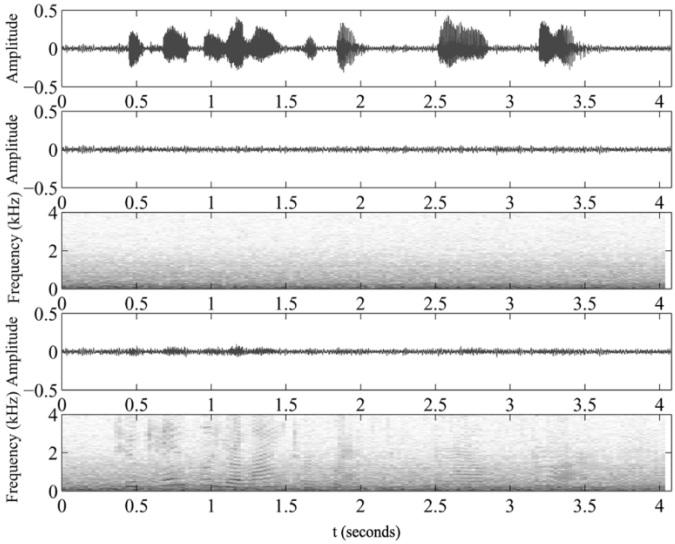

Fig. 4. Noise and its estimate. The first trace (from the top) shows the waveform of a speech signal corrupted by a car noise where SNR = 10 (10 dB). The second and third traces plot the waveform and spectrogram of the noise signal. The fourth and fifth traces display the waveform and spectrogram of the noise estimate.

Fig. 4 shows a speech signal corrupted by a car noise [SNR = 10 (10 dB)], the waveform and the spectrogram of the car noise that is added to the speech, and the waveform and spectrogram of the noise estimate. It can be seen that during the absence of speech, the estimate is a good approximation of the noise signal. It is also noticed from its spectrogram that the noise estimate consists of some minor speech components during the presence of speech. Our listening test, however, shows that the residual speech in the noise estimate is almost inaudible. An apparent advantage of this noise-estimation technique is that it does not require an explicit voice activity detector. In addition, our experimental investigation reveals that such a scheme is able to capture the noise characteristics in both the presence and absence of speech, therefore it does not rely on the assumption that the noise characteristics in the presence of speech stay the same as in the absence of speech.

$$\hat{V}_t(\omega) = \begin{cases} \alpha_{\rm a}\hat{V}_{t-1}(\omega) + (1-\alpha_{\rm a})|Y_t(j\omega)|, & \text{if } |Y_t(j\omega)| \geq \hat{V}_{t-1}(\omega) \\ \alpha_{\rm d}\hat{V}_{t-1}(\omega) + (1-\alpha_{\rm d})|Y_t(j\omega)|, & \text{if } |Y_t(j\omega)| < \hat{V}_{t-1}(\omega) \end{cases} \tag{87}$$

$$\bar{Y}_t(\omega) = \begin{cases} \beta_{\rm a}\bar{Y}_{t-1}(\omega) + (1-\beta_{\rm a})|Y_t(j\omega)|, & \text{if } |Y_t(j\omega)| \geq \bar{Y}_{t-1}(\omega) \\ \beta_{\rm d}\bar{Y}_{t-1}(\omega) + (1-\beta_{\rm d})|Y_t(j\omega)|, & \text{if } |Y_t(j\omega)| < \bar{Y}_{t-1}(\omega) \end{cases} \tag{88}$$

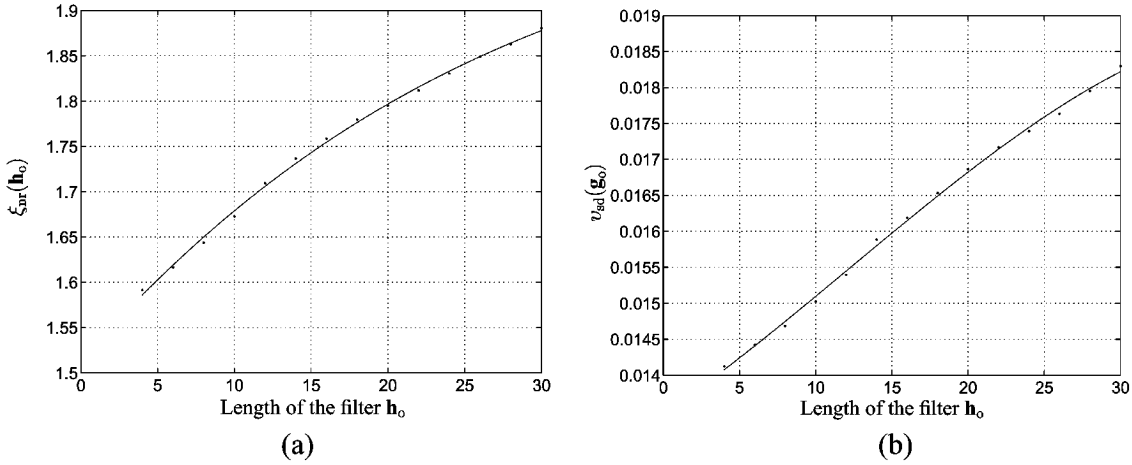

Fig. 5. Noise-reduction factor and signal-distortion index, both as a function of the filter length: (a) noise reduction and (b) signal distortion. The source is a signal recorded in a NYSE room; the background noise is a computer-generated white Gaussian random process; and $\mathrm{SNR} = 10$ (10 dB).

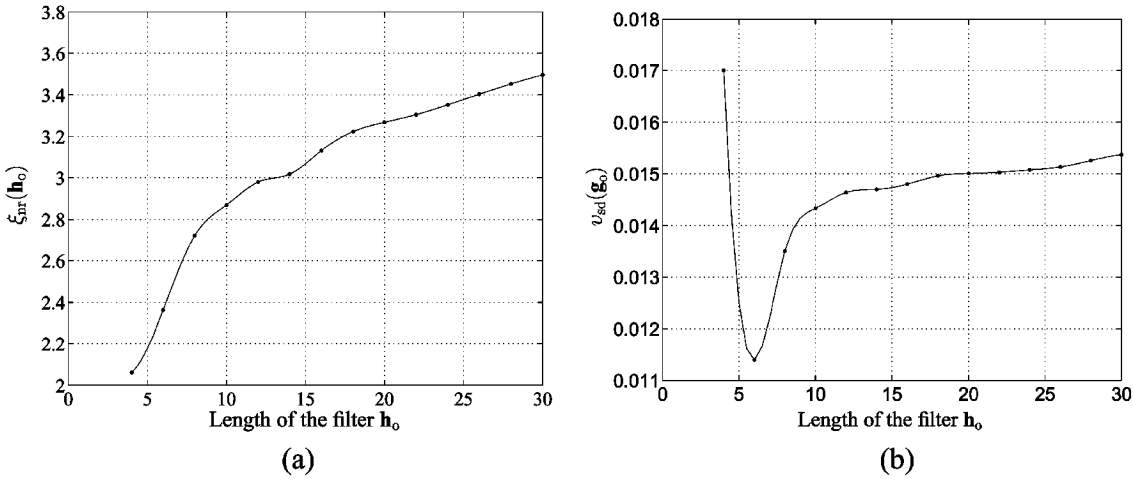

Fig. 6. Noise-reduction factor and signal-distortion index, both as a function of the filter length: (a) noise reduction and (b) speech distortion. The source signal is an /i:/ sound from a female speaker; the background noise is a computer-generated white Gaussian process; and $\mathrm{SNR} = 10$ (10 dB).

Based on the implemented system, we evaluate the Wiener filter for noise reduction. The first experiment investigates the influence of the filter length on the noise reduction performance. Instead of using the estimated noise, here we assume that the noise signal is known *a priori*. Therefore, this experiment demonstrates the upper limit of the performance of the Wiener filter. We consider two cases. In the first one, both the source signal and the background noise are random processes in which the current value of the signal cannot be predicted from its past samples. The source signal is a noise signal recorded from a New York Stock Exchange (NYSE) room. This signal consists of sound from various sources such as speakers, telephone rings, electric fans, etc. The background noise is a computer-generated Gaussian random process. The results for this case are graphically portrayed in Fig. 5. It can be seen that both the noise-reduction factor $[\xi_{\mathrm{nr}}(\mathbf{h}_o)]$ and the speech-distortion index increase linearly with the filter length. Therefore, a longer filter should be applied for more noise reduction. However, the more the noise is attenuated, the more the source signal is deformed, as shown in Fig. 5.

In the second case, we test the Wiener filter for noise reduction in the context of speech signals. It is known that a speech signal can be modeled as an AR process, where its current value can be predicted from its past samples. To simplify the situation for the ease of analysis, the source signal used here is an /i:/ sound recorded from a female speaker. Similarly as in the previous case, the background noise is a computer-generated white Gaussian random process. The results are plotted in Fig. 6. Again, the noise-reduction factor, which quantifies the amount of noise being attenuated, increases monotonically with the filter length; but unlike the previous case, the relationship between the noise reduction and the filter length is not linear. Instead, the curve at first grows quickly as the filter length is increased up to 10, and then continues to grow but with a slower rate. Unlike ξ_{nr}, the speech-distortion index, i.e., υ_{sd}, exhibits a nonmonotonic relationship with the filter length. It first decreases to its minimum, and then increases again as the filter length is increased. The reason, as we have explained in Section V, is that a speech signal can be modeled as an AR process. Particular to this experiment, the /i:/ sound used here can be well modeled with a sixth-order LPC (linear prediction coding) analysis. Therefore, when the filter length is increased to 6, the numerator of (34) is minimized, as a result, the speech-distortion index reaches its minimum. Continuing to increase the filter length leads to a

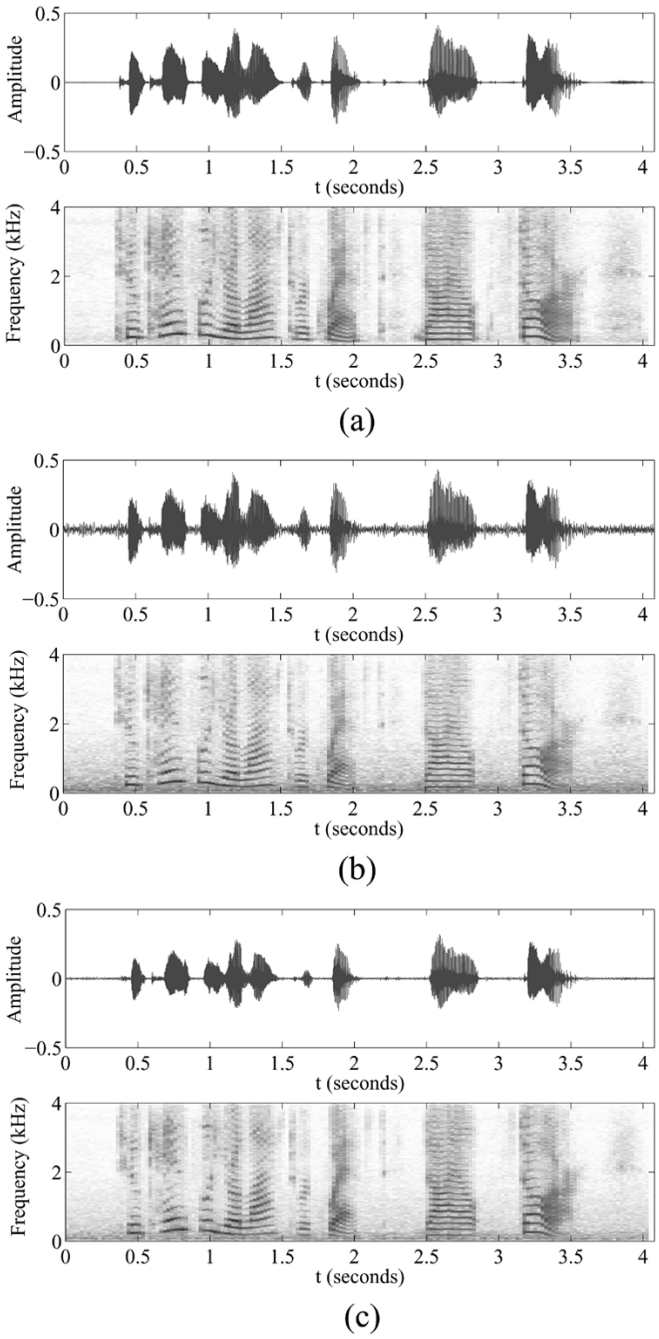

Fig. 7. Noise reduction in a car noise condition where $\mathrm{SNR} = 10\ (10\ \mathrm{dB})$: (a) clean speech and its spectrogram; (b) noisy speech and its spectrogram; and (c) noise reduced speech and its spectrogram.

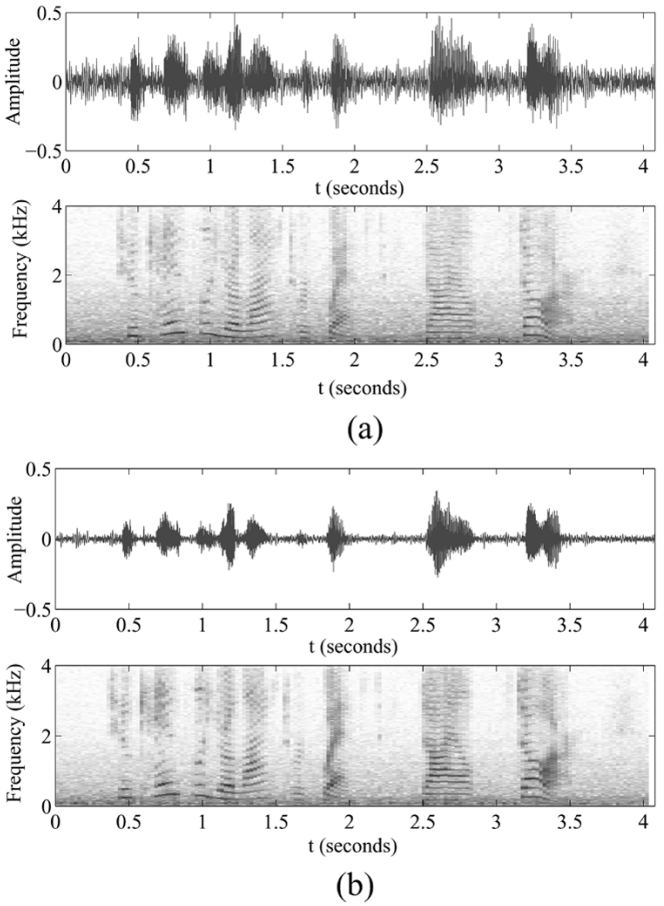

Fig. 8. Noise reduction in a car noise condition (same speech and noise signals as in Fig. 7) where $\mathrm{SNR} = 1\ (0\ \mathrm{dB})$: (a) noisy speech and its spectrogram and (b) noise reduced speech and its spectrogram.

higher distortion due to more noise reduction. To further verify this observation, we investigated several other vowels, and found that the curve of υ_{sd} versus filter length follows a similar shape, except that the minimum may appear in a slightly different location. Taking into account the sounds other than vowels in speech that may be less predicable, we find that good performance with the Wiener filter (in terms of the compromise between noise reduction and speech distortion) can be achieved when the filter length L is chosen around 20. Figs. 7 and 8 plot, respectively, the outputs of our Wiener filter system for $\mathrm{SNR} = 10\ (10\ \mathrm{dB})$ and $\mathrm{SNR} = 1\ (0\ \mathrm{dB})$, where the speech signal is from a female speaker, the background noise is a car noise signal, and $L = 20$.

The second experiment tests the noise reduction performance in different SNR conditions. Here the speech signal is recorded from a female speaker as shown in Fig. 7. The computer-generated random Gaussian noise is added to the speech signal to control the SNR. The length of the Wiener filter is set to $L = 20$. The results are presented in Fig. 9, where besides ξ_{nr} and υ_{sd}, we also plotted the Itakura–Saito (IS) distance, a widely used objective quality measure that performs a comparison of spectral envelopes (AR parameters) between the clean and the processed speech [53]. Studies have shown that the IS measure is highly correlated (0.59) with subjective quality judgements [54]. A recent report reveals that the difference in mean opinion score (MOS) between two processed speech signals would be less than 1.6 if their IS measure is less than 0.5 for various codecs [55]. Many other reported experiments confirmed that two spectra would be perceptually nearly identical if their IS distance is less than 0.1. All this evidence indicates that the IS distance is a reasonably good objective measure of speech quality.

As SNR decreases, the observation signal becomes more noisy. Therefore, the Wiener filter is expected to have more noise reduction for low SNRs. This is verified by Fig. 9(a), where significant noise reduction is obtained for low SNR conditions. However, more noise reduction would correspond to more speech distortion. This is confirmed by Fig. 9(b) and (d) where both the speech-distortion index and the IS distance increase as speech becomes more noisy. Comparing the IS

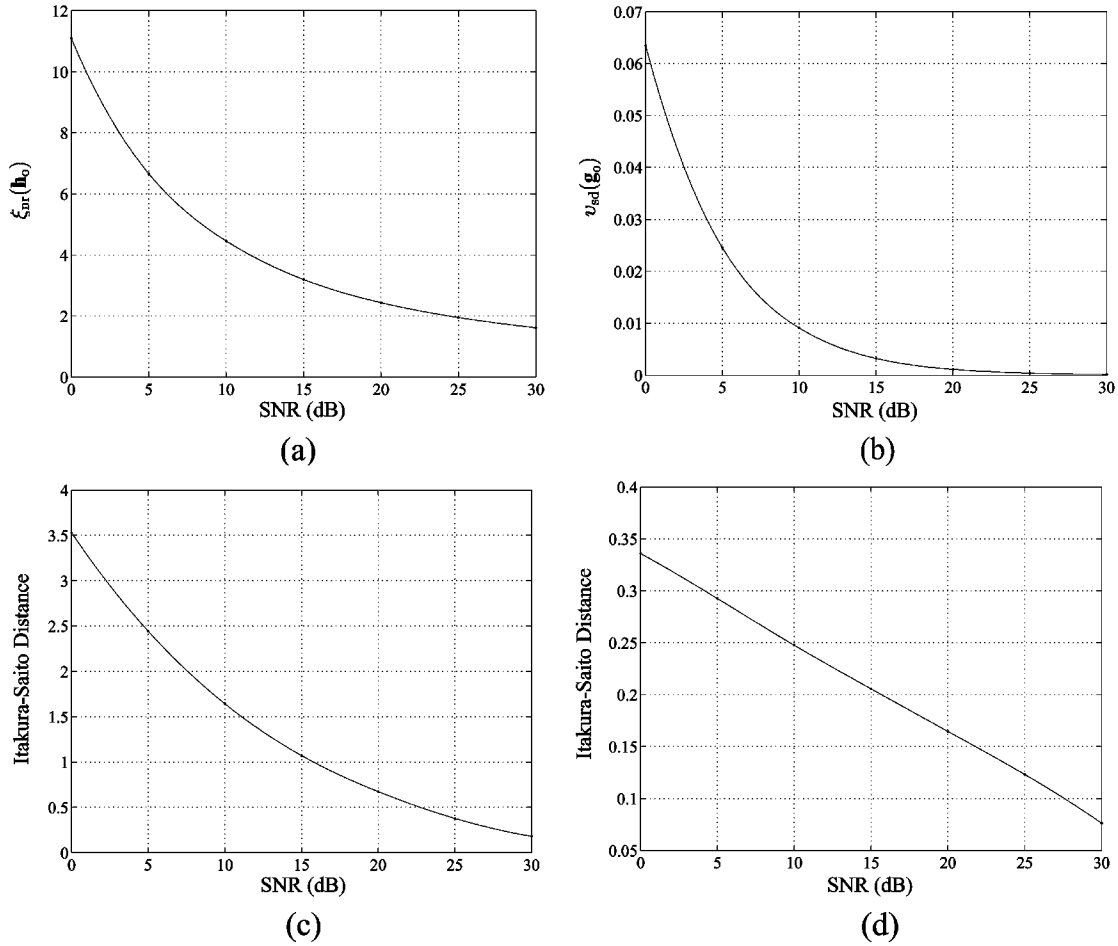

Fig. 9. Noise reduction performance as a function of SNR in white Gaussian noise: (a) noise-reduction factor; (b) speech-distortion index; (c) Itakura–Saito distance between the clean and noisy speeches; and (d) Itakura–Saito distance between the clean and noise-reduced speeches.

TABLE I
NOISE REDUCTION PERFORMANCE WITH THE SUBOPTIMAL FILTER, WHERE ISD^1 IS THE IS DISTANCE BETWEEN THE CLEAN SPEECH AND THE FILTERED VERSION OF THE CLEAN SPEECH, WHICH PURELY MEASURES THE SPEECH DISTORTION DUE TO THE FILTERING EFFECT; ISD^2 IS THE IS DISTANCE BETWEEN THE CLEAN AND NOISE-REDUCED SPEECHES; ISD^3 IS THE IS DISTANCE BETWEEN THE CLEAN AND NOISY SPEECH SIGNALS

SNR		v_{sd}	ξ_{nr}	ζ_{nr}	ISD^1	ISD^2	ISD^3
100 (20 dB)	Wiener filter	0.0011	2.4390	0.6152	0.1691	0.1471	0.6727
	Suboptimal filter ($\alpha = 0.8$)	0.0007	2.1753	0.5771	0.0423	0.2820	0.6727
	Suboptimal filter ($\alpha = 0.7$)	0.0006	2.0106	0.5422	0.0281	0.3476	0.6727
31.6 (15 dB)	Wiener filter	0.0033	3.1977	0.6903	0.2133	0.2032	1.0446
	Suboptimal filter ($\alpha = 0.8$)	0.0021	2.7379	0.6518	0.0488	0.5114	1.0446
	Suboptimal filter ($\alpha = 0.7$)	0.0016	2.4544	0.6139	0.0352	0.6034	1.0446
10 (10 dB)	Wiener filter	0.0092	4.4565	0.7610	0.2622	0.2652	1.5458
	Suboptimal filter ($\alpha = 0.8$)	0.0059	3.5896	0.7222	0.0582	0.7759	1.5458
	Suboptimal filter ($\alpha = 0.7$)	0.0045	3.0807	0.6816	0.0441	0.8917	1.5458

distance before [Fig. 9(c)] and after [Fig. 9(d)] noise reduction, one can see that significant gain in the IS distance has been achieved, indicating that the Wiener filter is able to reduce noise and improve speech quality (but not necessarily speech intelligibility).

The third experiment is to verify the performance behavior of the suboptimal filter derived in Section VI-A. The experimental conditions are the same as outlined in the previous experiment. The results are presented in Table I, where for the purpose of comparison, besides the speech-distortion index and the noise-reduction factor, we also show three IS distances (between the clean $[x(n)]$ and filtered speech $[\mathbf{h}^T\mathbf{x}(n)]$ signals denoted as ISD^1, between the clean and noise-reduced speech $[\mathbf{h}^T\mathbf{y}(n)]$ signals marked as ISD^2, and between the clean and noisy signals $[y(n)]$ denoted as ISD^3, respectively).

One can see that the IS distance between the clean and noisy speech signals increases as SNR drops. The reason for this is apparent. When SNR decreases, the speech signal becomes more

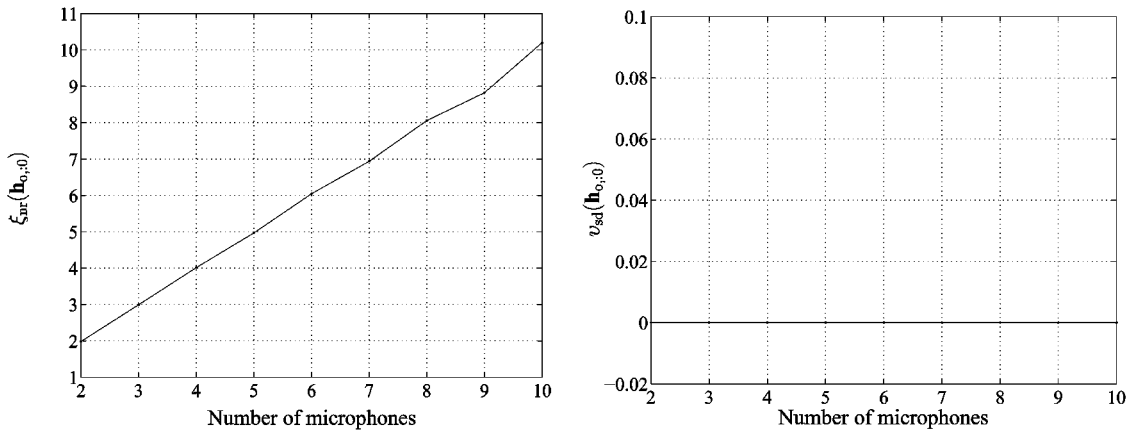

Fig. 10. Noise-reduction factor and signal-distortion index, both as a function of the number of microphone sensor: (a) noise reduction; (b) speech distortion. The source signal is a speech from a female speaker as shown in Fig. 7; the background noise is a computer-generated white Gaussian process; and $\mathrm{SNR} = 10\ (10\ \mathrm{dB})$.

noisy. As a result, the difference between the spectral envelope (or AR parameters) of the clean speech and that (or those) of the noisy speech tends to be more significant, which leads to a higher IS distance. It is noticed that ISD^2 is much smaller than ISD^3. This significant gain in IS distance indicates that the use of noise reduction technique is able to mitigate noise and improve speech quality. Comparing the results from both the Wiener and the suboptimal Wiener filters, we can see that a better compromise between noise reduction and speech distortion is accomplished by using the suboptimal filter. For example, when $\mathrm{SNR} = 100\ (20\ \mathrm{dB})$, the suboptimal filter with $\alpha = 0.7$ has achieved a noise reduction of 2.0106, which is 82% of that with the Wiener filter; but its speech-distortion index is 0.0006, which is only 54% of that of the Wiener filter; the corresponding IS distance between the clean and filtered speech is 0.0281, which is only 17% of that of the Wiener filter. From the analysis shown in Section VI-A, we know that both $\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{s}})/\upsilon_{\mathrm{sd}}(\mathbf{g}_{\mathrm{o}})$ and $\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})/\zeta_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})$ are independent of SNR. This can be easily verified from Table I. However, it is noted that $\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{s}})/\xi_{\mathrm{nr}}(\mathbf{h}_{\mathrm{o}})$ decreases with SNR, which may indicate that the suboptimal filter works more efficiently for higher SNR than for lower SNR conditions.

The last experiment is to investigate the performance of the multichannel optimal filter given in (79). Since the focus of this paper is on reduction of additive noise, the reverberation effect is not considered here. To simplify the analysis, we assume that we have an equispaced linear array, which consists of ten microphone sensors. The spacing between adjacent microphones is $d = 2.3$ cm. There is only a single speech source (a speech signal from a female speaker) propagating from the far field to the array with an incident angle (the angle between the wavefront and the line joining the sensors in the linear array) of $\theta = 30^{\circ}$. We further assume that all the microphone sensors have the same signal and noise power. The sampling rate is 16 kHz. For the experiment, we choose Microphone 0 as the reference sensor, and synchronize the observation signals according to the time-difference-of-arrival (TDOA) information estimated using the algorithm presented in [56]. We then pass the time-aligned observation signals through the optimal filter given in (79) to extract the desired speech signal. The results for this experiments are graphically portrayed in Fig. 10. It can be seen that the noise-reduction index increases linearly with the number of microphones, while the speech distortion is approximately 0. Comparing Fig. 10 with 9, one can see that in the condition where $\mathrm{SNR} = 10\ (10\ \mathrm{dB})$, the multichannel optimal filter with 4 sensors achieves a noise reduction similar to the optimal single-channel Wiener filter, but with no speech distortion, which shows the advantage of using multiple microphones.

VIII. Conclusion

The problem of speech enhancement has attracted a considerable amount of research attention over the past several decades. Among the numerous techniques that were developed, the optimal Wiener filter can be considered as one of the most fundamental noise-reduction approaches. It is widely known that the Wiener filter achieves noise reduction by deforming the speech signal. However, so far not much has been said on how the Wiener filter really works. In this paper we analyzed the inherent relationship between noise reduction and speech distortion with the Wiener filter. Starting from the speech and noise estimation using the Wiener theory, we introduced a speech-distortion index and two noise-reduction factors, and showed that for the single-channel Wiener filter, the amount of noise attenuation is in general proportional to the amount of speech degradation, i.e., more noise reduction incurs more speech distortion.

Depending on the nature of the application, some practical noise-reduction systems may require very high-quality speech, but can tolerate a certain amount of noise. While other systems may want speech as clean as possible even with some degree of speech distortion. Therefore, it is necessary to have some management schemes to control the contradicting requirements between noise reduction and speech distortion. To do so, we have discussed three approaches. If we know the linear prediction coefficients of the clean speech signal or they can be estimated from the noisy speech, these coefficients can be employed to achieve noise reduction while maintaining a low level of speech distortion. When no *a priori* knowledge is available, we can use a suboptimal filter in which a free parameter is introduced to control the compromise between noise reduction and speech

distortion. By setting the free parameter to 0.7, we showed that the suboptimal filter can achieve 90% of the noise reduction compared to the Wiener filter; but the resulting speech distortion is less than half compared to the Wiener filter. In case that we have multiple microphone sensors, the multiple observations of the speech signal can be used to reduce noise with less or even no speech distortion.

APPENDIX
RELATIONSHIP BETWEEN THE *A PRIORI* AND THE *A POSTERIORI* SNR

Theorem: With the Wiener filter in the context of noise reduction, the *a priori* SNR given in (12) and the *a posteriori* SNR defined in (42) satisfy

$$\mathrm{SNR}_{\mathrm{o}} \geq \mathrm{SNR}. \tag{89}$$

Proof: From their definitions, we know that all three matrices, $\mathbf{R}_x$, $\mathbf{R}_v$, and $\mathbf{R}_y$ are symmetric, and positive semi-definite. We further assume that $\mathbf{R}_v$ is positive definite so its inverse exists. In addition, based on the independence assumption between the speech signal and noise, we have $\mathbf{R}_y = \mathbf{R}_x + \mathbf{R}_v$. In case that both $\mathbf{R}_x$ and $\mathbf{R}_v$ are diagonal matrices, or $\mathbf{R}_v$ is a scaled version of $\mathbf{R}_x$ (i.e., $\mathbf{R}_x = \mathrm{SNR} \cdot \mathbf{R}_v$), it can be easily seen that $\mathrm{SNR}_{\mathrm{o}} = \mathrm{SNR}$. Here, we consider more complicated situations where at least one of the $\mathbf{R}_x$ and $\mathbf{R}_v$ matrices is not diagonal. In this case, according to [49], there exists a linear transformation that can simultaneously diagonalize $\mathbf{R}_x$, $\mathbf{R}_v$, and $\mathbf{R}_y$. The process is done as follows.

$$\begin{aligned}
\mathbf{R}_x &= (\mathbf{B}^T)^{-1}\mathbf{\Lambda}\mathbf{B}^{-1}, \\
\mathbf{R}_v &= (\mathbf{B}^T)^{-1}\mathbf{B}^{-1}, \\
\mathbf{R}_y &= (\mathbf{B}^T)^{-1}[\mathbf{I}+\mathbf{\Lambda}]\mathbf{B}^{-1}
\end{aligned} \tag{90}$$

where again $\mathbf{I}$ is the identity matrix

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & 0 & \lambda_L \end{bmatrix} \tag{91}$$

is the eigenvalue matrix of $\mathbf{R}_v^{-1}\mathbf{R}_x$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L \geq 0$, $\mathbf{B}$ is the eigenvector matrix of $\mathbf{R}_v^{-1}\mathbf{R}_x$, and

$$\mathbf{R}_v^{-1}\mathbf{R}_x\mathbf{B} = \mathbf{B}\mathbf{\Lambda}. \tag{92}$$

Note that $\mathbf{B}$ is not necessarily orthogonal since $\mathbf{R}_v^{-1}\mathbf{R}_x$ is not necessarily symmetric. Then from the definition of SNR and $\mathrm{SNR}_{\mathrm{o}}$, we immediately have

$$\mathrm{SNR} = \frac{\mathbf{u}_1^T\mathbf{R}_x\mathbf{u}_1}{\mathbf{u}_1^T\mathbf{R}_v\mathbf{u}_1} = \frac{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{\Lambda}\mathbf{B}^{-1}\mathbf{u}_1}{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{B}^{-1}\mathbf{u}_1} \tag{93}$$

and

$$\begin{aligned}
\mathrm{SNR}_{\mathrm{o}} &= \frac{\mathbf{h}_{\mathrm{o}}^T\mathbf{R}_x\mathbf{h}_{\mathrm{o}}}{\mathbf{h}_{\mathrm{o}}^T\mathbf{R}_v\mathbf{h}_{\mathrm{o}}} = \frac{\mathbf{u}_1^T\mathbf{R}_x^T\mathbf{R}_y^{-1}\mathbf{R}_x\mathbf{R}_y^{-1}\mathbf{R}_x\mathbf{u}_1}{\mathbf{u}_1^T\mathbf{R}_x^T\mathbf{R}_y^{-1}\mathbf{R}_v\mathbf{R}_y^{-1}\mathbf{R}_x\mathbf{u}_1} \\
&= \frac{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda}\mathbf{B}^{-1}\mathbf{u}_1}{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda}\mathbf{B}^{-1}\mathbf{u}_1} \\
&= \frac{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{\Sigma}_1\mathbf{B}^{-1}\mathbf{u}_1}{\mathbf{u}_1^T(\mathbf{B}^{-1})^T\mathbf{\Sigma}_2\mathbf{B}^{-1}\mathbf{u}_1}
\end{aligned} \tag{94}$$

where

$$\begin{aligned}
\mathbf{\Sigma}_1 &\triangleq \mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda} \\
&= \begin{bmatrix} \frac{\lambda_1^3}{(1+\lambda_1)^2} & 0 & \dots & 0 \\ 0 & \frac{\lambda_2^3}{(1+\lambda_2)^2} & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & 0 & \frac{\lambda_L^3}{(1+\lambda_L)^2} \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{\Sigma}_2 &\triangleq \mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{\Lambda} \\
&= \begin{bmatrix} \frac{\lambda_1^2}{(1+\lambda_1)^2} & 0 & \dots & 0 \\ 0 & \frac{\lambda_2^2}{(1+\lambda_2)^2} & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & 0 & \frac{\lambda_L^2}{(1+\lambda_L)^2} \end{bmatrix}
\end{aligned}$$

are two diagonal matrices. If for the ease of expression we denote $\mathbf{B}^{-1}$ as $\mathbf{A} = \mathbf{B}^{-1} = [a_{ij}]$, then both SNR and $\mathrm{SNR}_{\mathrm{o}}$ can be rewritten as

$$\begin{aligned}
\mathrm{SNR} &= \frac{\sum_{i=1}^{L}\lambda_i a_{i1}^2}{\sum_{i=1}^{L} a_{i1}^2}, \\
\mathrm{SNR}_{\mathrm{o}} &= \frac{\sum_{i=1}^{L}\frac{\lambda_i^3}{(1+\lambda_i)^2}a_{i1}^2}{\sum_{i=1}^{L}\frac{\lambda_i^2}{(1+\lambda_i)^2}a_{i1}^2}.
\end{aligned} \tag{95}$$

Since $\sum_{i=1}^{L}\left[\lambda_i^3/(1+\lambda_i)^2\right]a_{i1}^2$, $\sum_{i=1}^{L}\left[\lambda_i^2/(1+\lambda_i)^2\right]a_{i1}^2$, $\sum_{i=1}^{L}\lambda_i a_{i1}^2$, and $\sum_{i=1}^{L} a_{i1}^2$ all are nonnegative numbers, as long as we can show that the inequality

$$\sum_{i=1}^{L}\frac{\lambda_i^3}{(1+\lambda_i)^2}a_{i1}^2\sum_{i=1}^{L}a_{i1}^2 \geq \sum_{i=1}^{L}\frac{\lambda_i^2}{(1+\lambda_i)^2}a_{i1}^2\sum_{i=1}^{L}\lambda_i a_{i1}^2 \tag{96}$$

holds, then $\mathrm{SNR}_{\mathrm{o}} \geq \mathrm{SNR}$. Now we prove this inequality by way of induction.

- Basic Step: If $L = 2$,

$$\begin{aligned}
\sum_{i=1}^{2}\frac{\lambda_i^3}{(1+\lambda_i)^2}a_{i1}^2\sum_{i=1}^{2}a_{i1}^2 &= \frac{\lambda_1^3}{(1+\lambda_1)^2}a_{11}^4 \\
&\quad + \frac{\lambda_2^3}{(1+\lambda_2)^2}a_{21}^4 + \left[\frac{\lambda_1^3}{(1+\lambda_1)^2} + \frac{\lambda_2^3}{(1+\lambda_2)^2}\right]a_{11}^2a_{21}^2.
\end{aligned}$$

Since $\lambda_i \geq 0$, it is trivial to show that

$$\frac{\lambda_1^3}{(1+\lambda_1)^2} + \frac{\lambda_2^3}{(1+\lambda_2)^2} \geq \frac{\lambda_1^2\lambda_2}{(1+\lambda_1)^2} + \frac{\lambda_1\lambda_2^2}{(1+\lambda_2)^2}$$

where "=" holds when $\lambda_1 = \lambda_2$. Therefore

$$\begin{aligned}
&\sum_{i=1}^{2}\frac{\lambda_i^3}{(1+\lambda_i)^2}a_{i1}^2\sum_{i=1}^{2}a_{i1}^2 \\
&\quad \geq \frac{\lambda_1^3}{(1+\lambda_1)^2}a_{11}^4 + \frac{\lambda_2^3}{(1+\lambda_2)^2}a_{21}^4 \\
&\qquad + \left[\frac{\lambda_1^2\lambda_2}{(1+\lambda_1)^2} + \frac{\lambda_1\lambda_2^2}{(1+\lambda_2)^2}\right]a_{11}^2a_{21}^2 \\
&\quad = \sum_{i=1}^{2}\frac{\lambda_i^2}{(1+\lambda_i)^2}a_{i1}^2\sum_{i=1}^{2}\lambda_i a_{i1}^2
\end{aligned}$$

so the property is true for $L = 2$, where "=" holds when any one of a_{11} and a_{21} is equal to 0 (note that a_{11} and a_{21} cannot be zero at the same time since $\mathbf{A}$ is invertible) or when $\lambda_1 = \lambda_2$.

- Inductive Step: Assume that the property is true for $L = n$, i.e.,

$$\sum_{i=1}^{n} \frac{\lambda_i^3}{(1+\lambda_i)^2} a_{i1}^2 \sum_{i=1}^{n} a_{i1}^2 \geq \sum_{i=1}^{n} \frac{\lambda_i^2}{(1+\lambda_i)^2} a_{i1}^2 \sum_{i=1}^{n} \lambda_i a_{i1}^2.$$

We must prove that it is also true for $L = n+1$. As a matter of fact

$$\begin{aligned}
&\sum_{i=1}^{n+1} \frac{\lambda_i^3}{(1+\lambda_i)^2} a_{i1}^2 \sum_{i=1}^{n+1} a_{i1}^2 \\
&\quad = \left[\sum_{i=1}^{n} \frac{\lambda_i^3}{(1+\lambda_i)^2} a_{i1}^2 + \frac{\lambda_{n+1}^3}{(1+\lambda_{n+1})^2} a_{n+11}^2 \right] \\
&\qquad \times \left[\sum_{i=1}^{n} a_{i1}^2 + a_{n+11}^2 \right] \\
&\quad = \left[\sum_{i=1}^{n} \frac{\lambda_i^3}{(1+\lambda_i)^2} a_{i1}^2 \right] \left[\sum_{i=1}^{n} a_{i1}^2 \right] + \frac{\lambda_{n+1}^3}{(1+\lambda_{n+1})^2} a_{n+11}^4 \\
&\qquad + \sum_{i=1}^{n} \left[\frac{\lambda_i^3}{(1+\lambda_i)^2} + \frac{\lambda_{n+1}^3}{(1+\lambda_{n+1})^2} \right] a_{i1}^2 a_{n+11}^2. \qquad (97)
\end{aligned}$$

Using the induction hypothesis, and also the fact that

$$\frac{\lambda_i^3}{(1+\lambda_i)^2} + \frac{\lambda_{n+1}^3}{(1+\lambda_{n+1})^2} \geq \frac{\lambda_i^2 \lambda_{n+1}}{(1+\lambda_i)^2} + \frac{\lambda_i \lambda_{n+1}^2}{(1+\lambda_{n+1})^2}$$

hence

$$\begin{aligned}
&\sum_{i=1}^{n+1} \frac{\lambda_i^3}{(1+\lambda_i)^2} a_{i1}^2 \sum_{i=1}^{n+1} a_{i1}^2 \\
&\quad \geq \sum_{i=1}^{n} \frac{\lambda_i^2}{(1+\lambda_i)^2} a_{i1}^2 \sum_{i=1}^{n} \lambda_i a_{i1}^2 + \frac{\lambda_{n+1}^3}{(1+\lambda_{n+1})^2} a_{n+11}^4 \\
&\qquad + \sum_{i=1}^{n} \left[\frac{\lambda_i^2 \lambda_{n+1}}{(1+\lambda_i)^2} + \frac{\lambda_i \lambda_{n+1}^2}{(1+\lambda_{n+1})^2} \right] a_{i1}^2 a_{n+11}^2 \\
&\quad = \sum_{i=1}^{n+1} \frac{\lambda_i^2}{(1+\lambda_i)^2} a_{i1}^2 \sum_{i=1}^{n+1} \lambda_i a_{i1}^2 \qquad (98)
\end{aligned}$$

where "=" holds when all the λ_i's corresponding to nonzero a_{i1} are equal, where $i = 1, 2, \ldots, n+1$. That completes the proof. Even though it can improve the SNR, the Wiener filter does not maximize the *a posteriori* SNR. As a matter of fact, (42) is well known as the generalized Rayleigh quotient. So the filter that maximizes the *a posteriori* SNR is the eigenvector corresponding to the maximum eigenvalue of the matrix $\mathbf{R}_v^{-1}\mathbf{R}_x$. However, this filter typically gives rise to large speech distortion.

REFERENCES

[1] M. R. Schroeder, "Apparatus for supressing noise and distortion in communication signals," U.S. Patent 3 180 936, Apr., 27 1965.

[2] ——, "Processing of communication signals to reduce effects of noise," U.S. Patent 3 403 224, Sep., 24 1968.

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[4] J. S. Lim, *Speech Enhancement*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[5] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1554, Oct. 1992.

[6] E. J. Diethorn, "Subband noise reduction methods for speech enhancement," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Boston, MA: Kluwer, 2004, pp. 91–115.

[7] J. Chen, Y. Huang, and J. Benesty, "Filtering techniques for noise reduction and speech enhancement," in *Adaptive Signal Processing: Applications to Real-World Problems*, J. Benesty and Y. Huang, Eds. Berlin, Germany: Springer, 2003, pp. 129–154.

[8] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[9] S. E. Nordholm, I. Claesson, and N. Grbic, "Performance limits in subband beamforming," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 193–203, May 2003.

[10] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 497–507, Sep. 2000.

[11] F. Jabloun and B. Champagne, "A multi-microphone signal subspace approach for speech enhancement," in *Proc. IEEE ICASSP*, 2001, pp. 205–208.

[12] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2001.

[13] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[14] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985.

[15] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[16] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[17] P. Vary, "Noise suppression by spectral magnitude estimation-mechanism and theoretical limits," *Signal Process.*, vol. 8, pp. 387–400, Jul. 1985.

[18] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[19] W. Etter and G. S. Moschytz, "Noise reduction by noise-adaptive spectral magnitude expansion," *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.

[20] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 4, pp. 679–681, Aug. 1982.

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[22] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[23] N. Virag, "Single channel speech enhancement based on masking properties of human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[24] Y. M. Chang and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 1943–1954, Sep. 1991.

[25] T. F. Quatieri and R. B. Dunn, "Speech enhancement based on auditory spectral change," in *Proc. IEEE ICASSP*, vol. 1, May 2002, pp. 257–260.

[26] L. Deng, J. Droppo, and A. Acero, "Estimation cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 218–233, May 2004.
[27] ——, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 133–143, Mar. 2004.
[28] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.
[29] M. Dendrinos, S. Bakamidis, and G. Garayannis, "Speech enhancement from noise: A regenerative approach," *Speech Commun.*, vol. 10, pp. 45–57, Feb. 1991.
[30] P. S. K. Hansen, "Signal Subspace Methods for Speech Enhancement," Ph.D., Tech. Univ. Denmark, Lyngby, 1997.
[31] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, "Reduction of broad-band noise in speech by truncated qsvd," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 6, pp. 439–448, Nov. 1995.
[32] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Process. Lett.*, vol. 10, no. 4, pp. 104–106, Apr. 2003.
[33] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 2, pp. 87–95, Feb. 2001.
[34] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 159–167, Mar. 2000.
[35] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.
[36] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE ICASSP*, 1987, pp. 177–180.
[37] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.
[38] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 373–385, Jul. 1998.
[39] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 12, pp. 1846–1856, Dec. 1989.
[40] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, Apr. 1992.
[41] I. Cohen, "Modeling speech signals in the time-frequency domain using GARCH," *Signal Process.*, vol. 84, pp. 2453–2459, Dec. 2004.
[42] T. Lotter, "Single and Multichannel Speech Enhancement for Hearing Aids," Ph.D. dissertation, RWTH Aachen Univ., Aachen, Germany, 2004.
[43] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 173–185, Mar. 2002.
[44] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, Sep. 1998.
[45] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 341–351, Sep. 2002.
[46] J. Vermaak and M. Niranjan, "Markov Chain Monte Carlo methods for speech enhancement," in *Proc. IEEE ICASSP*, vol. 2, May 1998, pp. 1013–1016.
[47] S. Haykin, *Adaptive Filter Theory*, 4th Ed. ed. Upper Saddle River, NJ: Prentice-Hall, 2002.
[48] P. M. Clarkson, *Optimal and Adaptive Signal Processing*. Boca Raton, FL: CRC, 1993.
[49] K. Fukunaga, *Introduction to Statistial Pattern Recognition*. San Diego, CA: Academic, 1990.
[50] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[51] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.
[52] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 10, pp. 1365–1375, Oct. 1987.
[53] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[54] S. Quakenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[55] G. Chen, S. N. Koh, and I. Y. Soon, "Enhanced Itakura measure incorporating masking properties of human auditory system," *Signal Process.*, vol. 83, pp. 1445–1456, Jul. 2003.
[56] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, pp. 384–391, Jan. 2000.

Jingdong Chen (M'99) received the B.S. degree in electrical engineering and the M.S. degree in array signal processing from the Northwestern Polytechnic University in 1993 and 1995, respectively, and the Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences in 1998. His Ph.D. research focused on speech recognition in noisy environments. He studied and proposed several techniques covering speech enhancement and HMM adaptation by signal transformation.

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, as a Research Fellow, where he engaged in research in robust speech recognition, signal processing, and discriminative feature representation. From 2000 to 2001, he was with ATR Spoken Language Translation Research Laboratories, Kyoto, where he conducted research in robust speech recognition and speech enhancement. He joined Bell Laboratories as a Member of Technical Staff in July 2001. His current research interests include adaptive signal processing, speech enhancement, adaptive noise/echo cancellation, microphone array signal processing, signal separation, and source localization. He is a co-editor/co-author of the book *Speech Enhancement* (Berlin, Germany: Springer-Verlag, 2005).

Dr. Chen is the recipient of 1998–1999 research grant from the Japan Key Technology Center, and the 1996–1998 President's Award from the Chinese Academy of Sciences.

Jacob Benesty (SM'04) was born in Marrakech, Morocco, in 1963. He received the Masters degree in microwaves from Pierre & Marie Curie University, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, France, in April 1991.

During his Ph.D. program (from November 1989 to April 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecommunications (CNET), Paris, France. From January 1994 to July 1995, he worked at Telecom Paris 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. In May 2003, he joined the Université du Québec, INRS-EMT, in Montréal, QC, Canada, as an associate professor. His research interests are in acoustic signal processing and multimedia communications. He co-authored the book *Advances in Network and Acoustic Echo Cancellation* (Berlin, Germany: Springer-Verlag, 2001). He is also a co-editor/co-author of the books *Speech Enhancement* (Berlin, Germany: Springer-Verlag, 2005), *Audio Signal Processing for Next-Generation Multimedia Communication Systems* (Boston, MA: Kluwer, 2004), *Adaptive Signal Processing: Applications to Real-World Problems* (Berlin, Germany: Springer-Verlag, 2003), and *Acoustic Signal Processing for Telecommunication* (Boston, MA: Kluwer, 2000).

Dr. Benesty received the 2001 Best Paper Award from the IEEE Signal Processing Society. He is a member of the editorial board of the EURASIP Journal on Applied Signal Processing. He was the co-chair of the 1999 International Workshop on Acoustic Echo and Noise Control.

Yiteng (Arden) Huang (S'97–M'01) received the B.S. degree from the Tsinghua University in 1994, the M.S. and Ph.D. degrees from the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1998 and 2001, respectively, all in electrical and computer engineering.

During his doctoral studies from 1998 to 2001, he was a Research Assistant with the Center of Signal and Image Processing, Georgia Tech, and was a teaching assistant with the School of Electrical and Computer Engineering, Georgia Tech. In the summers from 1998 to 2000, he worked with Bell Laboratories, Murray Hill, NJ and engaged in research on passive acoustic source localization with microphone arrays. Upon graduation, he joined Bell Laboratories as a Member of Technical Staff in March 2001. His current research interests are in multichannel acoustic signal processing, multimedia and wireless communications. He is a co-editor/co-author of the books *Audio Signal Processing for Next-Generation Multimedia Communication Systems* (Boston, MA: Kluwer, 2004) and *Adaptive Signal Processing: Applications to Real-World Problems* (Berlin, Germany: Springer-Verlag, 2003).

Dr. Huang was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS. He received the 2002 Young Author Best Paper Award from the IEEE Signal Processing Society, the 2000–2001 Outstanding Graduate Teaching Assistant Award from the School Electrical and Computer Engineering, Georgia Tech, the 2000 Outstanding Research Award from the Center of Signal and Image Processing, Georgia Tech, and the 1997–1998 Colonel Oscar P. Cleaver Outstanding Graduate Student Award from the School of Electrical and Computer Engineering, Georgia Tech.

Simon Doclo (S'95–M'03) was born in Wilrijk, Belgium, in 1974. He received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in 1997 and 2003, respectively.

Currently, he is a Postdoctoral Fellow of the Fund for Scientific Research—Flanders, affiliated with the Electrical Engineering Department of the Katholieke Universiteit Leuven. In 2005, he was a Visiting Postdoctoral Fellow at the Adaptive Systems Laboratory, McMaster University, Hamilton, ON, Canada. His research interests are in microphone array processing for acoustic noise reduction, dereverberation and sound localization, adaptive filtering, speech enhancement, and hearing aid technology. He serves as Guest Editor for the *Journal on Applied Signal Processing*.

Dr. Doclo received the first prize "KVIV-Studentenprijzen" (with E. De Clippel) for the best M.Sc. engineering thesis in Flanders in 1997, a Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2001, and the EURASIP Signal Processing Best Paper Award 2003 (with M. Moonen). He was secretary of the IEEE Benelux Signal Processing Chapter (1998–2002).