ANALYSIS OF THE FREQUENCY-DOMAIN WIENER FILTER WITH THE PREDICTION GAIN

Jingdong Chen[1]*, Jacob Benesty*[2]*, and Yiteng (Arden) Huang*[1]

[1]: WeVoice, Inc.
9 Sylvan Dr.
Bridgewater, NJ 08807, USA

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

ABSTRACT

This paper presents a theoretical analysis on the performance of the optimal noise-reduction filter in the frequency domain. Using the autoregressive (AR) model to model both the clean speech and noise, we build the relationship between the Wiener filter and the AR parameters of the clean speech and noise signals. We show that if noise is not predictable, the Wiener filter is mostly related to the AR parameters of the desired speech signal. On the contrary, if the desired signal is not predictable, the Wiener filter is then mostly related to the AR parameters of the noise signal. More importantly, we provide the bounds for noise reduction, speech distortion, and SNR improvement, and show that the performance of the Wiener filter in terms of SNR improvement and degree of noise reduction and speech distortion is closely related to the prediction gain of the desired speech and noise signals.

Index Terms— Noise reduction, Wiener filter, autoregressive model, prediction gain.

1. PROBLEM FORMULATION, SIGNAL MODEL, AND PREDICTION GAIN

The noise reduction problem considered in this paper is one of recovering the desired signal (clean speech) $x(k)$, k being the discrete-time index, of zero mean from the noisy observation (microphone signal)

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

where $v(k)$ is the unwanted additive noise, which is assumed to be a zero-mean random process (white or colored) and uncorrelated with $x(k)$. Using the z-transform, (1) can be rewritten as

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

where $Y(z)$, $X(z)$, and $V(z)$ are the z-transforms of $y(k)$, $x(k)$, and $v(k)$, respectively. In the rest, we will always take $z = e^{j\omega}$, where j is the imaginary unit ($j^2 = -1$) and ω ($-\pi < \omega \leq \pi$) is the angular frequency.

Since $x(k)$ and $v(k)$ are assumed to be uncorrelated, we have

$$\phi_y(\omega) = \phi_x(\omega) + \phi_v(\omega), \tag{3}$$

where

$$\phi_a(\omega) = E\left[|A(e^{j\omega})|^2\right] \tag{4}$$

is the power spectral density (PSD) of the signal $a(k)$, $a \in \{x, v, y\}$, and $E[\cdot]$ denotes the mathematical expectation.

An estimate of $X(e^{j\omega})$ can be obtained by multiplying $Y(e^{j\omega})$ with a complex gain, $H(e^{j\omega})$, i.e.,

$$Z(e^{j\omega}) = H(e^{j\omega})Y(e^{j\omega}) = X_{\rm fd}(e^{j\omega}) + V_{\rm rn}(e^{j\omega}), \tag{5}$$

where $Z(e^{j\omega})$ is the frequency-domain representation of the signal $z(k)$, $X_{\rm fd}(e^{j\omega}) \triangleq H(e^{j\omega})X(e^{j\omega})$ is the filtered clean speech, and $V_{\rm rn}(e^{j\omega}) \triangleq H(e^{j\omega})V(e^{j\omega})$ is the residual noise. From (5), we deduce the PSD of $z(k)$:

$$\phi_z(\omega) = |H(e^{j\omega})|^2 \phi_y(\omega) = |H(e^{j\omega})|^2 \left[\phi_x(\omega) + \phi_v(\omega)\right]. \tag{6}$$

The objective of noise reduction in the frequency domain is then to find an optimal gain $H(e^{j\omega})$ at each frequency ω that would attenuate the noise as much as possible with as little distortion as possible to the desired signal (speech).

To better understand and analyze the optimal gains in the context of noise reduction, we propose to use the autoregressive (AR) model [5] for both the clean speech and noise signals. With this popular linear stochastic model, these two signals can be written as linear combination of their past values, i.e.,

$$x(k) = \sum_{l=1}^{L_x} a_{x,l}\, x(k-l) + u_x(k), \tag{7}$$

$$v(k) = \sum_{l=1}^{L_v} a_{v,l}\, v(k-l) + u_v(k), \tag{8}$$

where $a_{x,l}$ and $a_{v,l}$ are the AR parameters of $x(k)$ and $v(k)$, respectively, and $u_x(k)$ and $u_v(k)$ are two zero-mean random white noise signals. In the frequency domain, (7) and (8) are

$$X(e^{j\omega}) = \frac{U_x(e^{j\omega})}{A_x(e^{j\omega})}, \tag{9}$$

$$V(e^{j\omega}) = \frac{U_v(e^{j\omega})}{A_v(e^{j\omega})}, \tag{10}$$

where $U_x(e^{j\omega})$ and $U_v(e^{j\omega})$ are the frequency-domain representations of $u_x(k)$ and $u_v(k)$, respectively, and

$$A_x(e^{j\omega}) = 1 - \sum_{l=1}^{L_x} a_{x,l} e^{-j\omega l}, \tag{11}$$

$$A_v(e^{j\omega}) = 1 - \sum_{l=1}^{L_v} a_{v,l} e^{-j\omega l}, \tag{12}$$

are two minimum-phase polynomials. With this AR modelling, the PSDs of the speech and noise signals become

$$\phi_x(\omega) = \frac{\sigma_{u_x}^2}{|A_x(e^{j\omega})|^2}, \tag{13}$$

$$\phi_v(\omega) = \frac{\sigma_{u_v}^2}{|A_v(e^{j\omega})|^2}, \tag{14}$$

where $\sigma^2_{u_x} = E\left[u^2_x(k)\right]$ and $\sigma^2_{u_v} = E\left[u^2_v(k)\right]$ are the variances of $u_x(k)$ and $u_v(k)$, respectively.

We are now ready to define a measure that naturally follows from the AR modelling. We define the subband prediction gain of the desired signal $x(k)$ as

$$\mathcal{P}_x(\omega) \triangleq \frac{\phi_x(\omega)}{\sigma^2_{u_x}} = \frac{1}{\left|A_x(e^{j\omega})\right|^2}. \quad (15)$$

If $\mathcal{P}_x(\omega) = 1$ then the signal $x(k)$ is completely unpredictable at frequency ω. A larger value of $\mathcal{P}_x(\omega)$ means that the signal $x(k)$ is more predictable at frequency ω.

In a similar manner, we define the fullband prediction gain [6] of the desired signal $x(k)$ as

$$\mathcal{P}_x \triangleq \frac{\sigma^2_x}{\sigma^2_{u_x}}, \quad (16)$$

where $\sigma^2_x = E\left[x^2(k)\right]$ is the variance of $x(k)$. Using the relation between the variance and the PSD of $x(k)$:

$$\sigma^2_x = \frac{1}{2\pi}\int_{-\pi}^{\pi}\phi_x(\omega)d\omega = \frac{\sigma^2_{u_x}}{2\pi}\int_{-\pi}^{\pi}\frac{d\omega}{\left|A_x(e^{j\omega})\right|^2}, \quad (17)$$

(16) can be rewritten as

$$\mathcal{P}_x = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{d\omega}{\left|A_x(e^{j\omega})\right|^2} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathcal{P}_x(\omega)d\omega. \quad (18)$$

It can be checked that $\mathcal{P}_x \geq 1$. As a matter of fact, using (7) we find that the variance of $x(k)$ is

$$\sigma^2_x = \mathbf{a}_x^T\mathbf{R}_x\mathbf{a}_x + \sigma^2_{u_x}, \quad (19)$$

where

$$\mathbf{a}_x = \left[\begin{array}{cccc} a_{x,1} & a_{x,2} & \cdots & a_{x,L_x}\end{array}\right]^T,$$

and $\mathbf{R}_x = E\left[\mathbf{x}(k)\mathbf{x}^T(k)\right]$ is the correlation matrix of

$$\mathbf{x}(k) = \left[\begin{array}{cccc} x(k) & x(k-1) & \cdots & x(k-L_x+1)\end{array}\right]^T.$$

From (19), we get another form of the fullband prediction gain:

$$\mathcal{P}_x = 1 + \frac{\mathbf{a}_x^T\mathbf{R}_x\mathbf{a}_x}{\sigma^2_{u_x}}. \quad (20)$$

Since $\mathbf{a}_x^T\mathbf{R}_x\mathbf{a}_x \geq 0$, we deduce from (20) that $\mathcal{P}_x \geq 1$.

If $\mathcal{P}_x = 1$, this means that the signal $x(k)$ is completely unpredictable (white noise). The larger is $\mathcal{P}_x$, the more predictable is the signal $x(k)$.

Obviously, the same definitions of subband and fullband prediction gains apply to the noise signal $v(k)$ by just replacing in the previous derivations $x(k)$ by $v(k)$.

By analogy to the fullband prediction gain defined in (18), we define the fullband prediction gains of the filtered desired and residual noise signals as, respectively,

$$\mathcal{P}_x(H) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H(e^{j\omega})\right|^2\mathcal{P}_x(\omega)d\omega \quad (21)$$

and

$$\mathcal{P}_v(H) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|H(e^{j\omega})\right|^2\mathcal{P}_v(\omega)d\omega. \quad (22)$$

We can verify that if $\left|H(e^{j\omega})\right|^2 \leq 1,\ \forall\omega$, then $\mathcal{P}_x(H) \leq \mathcal{P}_x$ and $\mathcal{P}_v(H) \leq \mathcal{P}_v$ with equalities if and only if $\left|H(e^{j\omega})\right|^2 = 1,\ \forall\omega$. Contrary to $\mathcal{P}_x$ and $\mathcal{P}_v$, $\mathcal{P}_x(H)$ and $\mathcal{P}_v(H)$ can be smaller than 1.

We will see that the noise reduction (or speech enhancement) problem can be exclusively reformulated as a function of the prediction gains.

2. PERFORMANCE MEASURES

To facilitate the analysis and interpretation of the noise-reduction performance, some performance measures are presented in this section.

The most important and reliable measure in noise reduction is the signal-to-noise ratio (SNR) [3]. The fullband input SNR is defined as the ratio of the intensity of the signal of interest over the intensity of the additive noise, i.e.,

$$\text{iSNR} \triangleq \frac{\sigma^2_x}{\sigma^2_v} = \text{iSNR}_0 \cdot \frac{\mathcal{P}_x}{\mathcal{P}_v}, \quad (23)$$

where $\text{iSNR}_0 \triangleq \sigma^2_{u_x}/\sigma^2_{u_v}$. It is worth noticing how the fullband prediction gains affect the SNR. Indeed, if the desired signal is more predictable than the noise signal then iSNR $>$ iSNR_0. But if the noise signal is more predictable than the desired signal then iSNR $<$ iSNR_0. The subband input SNR is

$$\text{iSNR}(\omega) \triangleq \frac{\phi_x(\omega)}{\phi_v(\omega)} = \text{iSNR}_0 \cdot \frac{\mathcal{P}_x(\omega)}{\mathcal{P}_v(\omega)}. \quad (24)$$

After noise reduction with the frequency-domain model given in (5), the subband output SNR is

$$\text{oSNR}\left[H(e^{j\omega})\right] \triangleq \frac{\left|H\left(e^{j\omega}\right)\right|^2\phi_x(\omega)}{\left|H\left(e^{j\omega}\right)\right|^2\phi_v(\omega)} = \text{iSNR}(\omega). \quad (25)$$

So, the subband SNR is not affected by the filtering process. But the fullband output SNR is

$$\text{oSNR}\left(H\right) \triangleq \frac{\int_{-\pi}^{\pi}\left|H\left(e^{j\omega}\right)\right|^2\phi_x(\omega)d\omega}{\int_{-\pi}^{\pi}\left|H\left(e^{j\omega}\right)\right|^2\phi_v(\omega)d\omega} = \text{iSNR}_0 \cdot \frac{\mathcal{P}_x(H)}{\mathcal{P}_v(H)}. \quad (26)$$

Apparently, the prediction gains will affect the fullband output SNR. We also define the fullband SNR gain as

$$\mathcal{G}\left(H\right) \triangleq \frac{\text{oSNR}\left(H\right)}{\text{iSNR}} = \frac{\mathcal{P}_x(H)}{\mathcal{P}_x} \cdot \frac{\mathcal{P}_v}{\mathcal{P}_v(H)}. \quad (27)$$

For a constant $\alpha \neq 0$, $\mathcal{G}\left(\alpha H\right) = \mathcal{G}\left(H\right)$. Therefore, changing the gains $H\left(e^{j\omega}\right)$ by a scaling factor (same over all frequencies) will not affect the fullband SNR gain. We also observe in (27) that making the filtered desired signal more predictable than the desired signal or making the residual noise signal less predictable than the original noise signal, will increase the fullband SNR gain.

It is of great importance to design the gains $H\left(e^{j\omega}\right)$ in such a way that $\mathcal{G}\left(H\right) > 1$; this would mean that the output SNR would be improved and the estimated signal, $z(k)$, would be less noisy than the microphone signal, $y(k)$. We see from (27) that $\mathcal{G}\left(H\right)$ depends exclusively on the prediction gains, $\mathcal{P}_x$, $\mathcal{P}_x(H)$, $\mathcal{P}_v$, and $\mathcal{P}_v(H)$. Clearly, the prediction gains of the desired and noise signals play a fundamental role in noise reduction and will affect the design and performance of the gains, $H\left(e^{j\omega}\right)$.

Another important measure in noise reduction is the noise-reduction factor [1], [2], which quantifies the amount of noise being attenuated by the gains. With the frequency-domain formulation, these subband and fullband factors are defined as, respectively,

$$\xi_{\text{nr}}\left[H(e^{j\omega})\right] \triangleq \frac{\phi_v(\omega)}{\left|H\left(e^{j\omega}\right)\right|^2\phi_v(\omega)} = \frac{1}{\left|H\left(e^{j\omega}\right)\right|^2}, \quad (28)$$

$$\xi_{\text{nr}}\left(H\right) \triangleq \frac{\int_{-\pi}^{\pi}\phi_v(\omega)d\omega}{\int_{-\pi}^{\pi}\left|H\left(e^{j\omega}\right)\right|^2\phi_v(\omega)d\omega} = \frac{\mathcal{P}_v}{\mathcal{P}_v(H)}. \quad (29)$$

The larger the value of the noise-reduction factor, the more the noise is reduced. This factor should be lower bounded by 1.

The gains add distortion to the desired signal. In order to evaluate the amount of distortion to the signal of interest, we define the subband and fullband speech-distortion indices [1], [2] as, respectively,

$$\upsilon_{\rm sd}\left[H(e^{j\omega})\right] \triangleq \frac{E\left[|H(e^{j\omega})X(e^{j\omega})-X(e^{j\omega})|^2\right]}{\phi_x(\omega)} = \left|1-H(e^{j\omega})\right|^2, \quad (30)$$

$$\upsilon_{\rm sd}\left(H\right) \triangleq \frac{\int_{-\pi}^{\pi} E\left[|H(e^{j\omega})X(e^{j\omega})-X(e^{j\omega})|^2\right] d\omega}{\int_{-\pi}^{\pi}\phi_x(\omega)d\omega} = \frac{\int_{-\pi}^{\pi}\phi_x(\omega)|1-H(e^{j\omega})|^2 d\omega}{\int_{-\pi}^{\pi}\phi_x(\omega)d\omega} = \frac{\int_{-\pi}^{\pi}\mathcal{P}_x(\omega)|1-H(e^{j\omega})|^2 d\omega}{2\pi\mathcal{P}_x}. \quad (31)$$

The speech-distortion index is lower bounded by 0 and expected to be upper bounded by 1 for optimal gains. The higher the value of this index, the more the desired signal is distorted.

3. WIENER FILTER

In this section, we derive the Wiener filter and explain its relationship with AR processes.

We define the error signal between the estimated and desired signals at frequency ω as

$$\mathcal{E}(e^{j\omega}) \triangleq Z(e^{j\omega}) - X(e^{j\omega}) = H(e^{j\omega})Y(e^{j\omega}) - X(e^{j\omega}). \quad (32)$$

This error can also be put into the form:

$$\mathcal{E}(e^{j\omega}) = \mathcal{E}_x(e^{j\omega}) + \mathcal{E}_v(e^{j\omega}), \quad (33)$$

where

$$\mathcal{E}_x(e^{j\omega}) \triangleq \left[H(e^{j\omega})-1\right]X(e^{j\omega}) \quad (34)$$

is the speech distortion due to the complex gain, and

$$\mathcal{E}_v(e^{j\omega}) \triangleq H(e^{j\omega})V(e^{j\omega}) \quad (35)$$

represents the residual noise.

The frequency-domain (or subband) mean-squared error (MSE) is then

$$J\left[H(e^{j\omega})\right] = E\left[|\mathcal{E}(e^{j\omega})|^2\right]. \quad (36)$$

Minimizing the subband MSE with respect to $H(e^{j\omega})$, we easily find the Wiener gain:

$$H_{\rm W}(e^{j\omega}) = \frac{\phi_x(\omega)}{\phi_y(\omega)} = 1 - \frac{\phi_v(\omega)}{\phi_y(\omega)} = \frac{\rm iSNR(\omega)}{1+\rm iSNR(\omega)}. \quad (37)$$

With this filter, we have the following property.

Property: With the optimal noncausal Wiener filter given in (37), the fullband SNR gain is always greater than or equal to 1, i.e., $\mathcal{G}(H_{\rm W}) \geq 1$.

The proof of this can be found in [4]. This fundamental property shows that the optimal gain can never amplify the noise. Nevertheless, it is essential to understand when the Wiener filter improves the SNR when it can.

It is more informative to rewrite the optimal gain (37) as a function of the prediction gains:

$$H_{\rm W}(e^{j\omega}) = \frac{\rm iSNR_0\cdot\mathcal{P}_x(\omega)}{\mathcal{P}_v(\omega)+\rm iSNR_0\cdot\mathcal{P}_x(\omega)}. \quad (38)$$

From the definition of the fullband SNR gain, we observe that the worst-case scenario for the Wiener filter is when $\mathcal{P}_x(\omega) = \mathcal{P}_v(\omega),\ \forall\omega$. Indeed, in this situation

$$H_{\rm W}(e^{j\omega}) = \frac{\rm iSNR_0}{1+\rm iSNR_0} = \text{Constant},\ \forall\omega \quad (39)$$

and it is easy to verify that $\mathcal{G}(H_{\rm W}) = 1$, so the SNR cannot be improved. The case $\mathcal{P}_x(\omega) = \mathcal{P}_v(\omega),\ \forall\omega$, happens when either the speech and noise signals are completely unpredictable (white random signals) so that $\mathcal{P}_x = \mathcal{P}_v = 1$ [and $\mathcal{P}_x(\omega) = \mathcal{P}_v(\omega) = 1,\ \forall\omega$] or the speech and noise signals are basically the same signals (having the same AR parameters) but they may have different volumes. It is now easy to understand why with babbling noise, for example, the Wiener filter may not perform at its best, since the speech and noise signals may have similar spectra [3]. This simple analysis of the Wiener filter is very insightful; it explains some practical situations where this filter may have limited performances.

To further analyze (38), we assume that $\rm iSNR_0 = 1$ (for simplicity) and the desired and noise signals are neither white nor they have the same AR parameters. In this case

$$H_{\rm W}(e^{j\omega}) = \frac{\mathcal{P}_x(\omega)}{\mathcal{P}_v(\omega)+\mathcal{P}_x(\omega)}. \quad (40)$$

If the desired signal is more predictable than the noise signal at all frequencies [i.e., $\mathcal{P}_x(\omega) > \mathcal{P}_v(\omega),\ \forall\omega$], we have $H_{\rm W}(e^{j\omega}) > 0.5,\ \forall\omega$. But if the noise signal is more predictable than the desired signal at all frequencies [i.e., $\mathcal{P}_v(\omega) > \mathcal{P}_x(\omega),\ \forall\omega$] then $H_{\rm W}(e^{j\omega}) < 0.5,\ \forall\omega$. We deduce, by looking at (27), that the fullband SNR gain is greater in the second case than in the first one but at a price of more speech distortion by inspection of (31). These two cases have opposite behaviors. These results may come as a small surprise. The fact that the noise is predictable may not be, after all, a bad thing as long as its spectrum is different from the speech spectrum. It is often intuitively thought that the non-whiteness of the noise may limit the performance of the Wiener filter but we understand now that this limitation that may occur in practical situations may mainly be due to the fact that noise and speech may have similar spectra.

Let us further exploit (38) by rewriting it as follows:

$$H_{\rm W}^2(e^{j\omega})\left[\mathcal{P}_v(\omega)+\rm iSNR_0\cdot\mathcal{P}_x(\omega)\right] = \rm iSNR_0\cdot\mathcal{P}_x(\omega)H_{\rm W}(e^{j\omega}). \quad (41)$$

Integrating both sides of (41) and multiplying by $1/(2\pi)$, we obtain

$$\mathcal{P}_v(H_{\rm W})+\rm iSNR_0\cdot\mathcal{P}_x(H_{\rm W}) = \frac{\rm iSNR_0}{2\pi}\int_{-\pi}^{\pi} H_{\rm W}(e^{j\omega})\mathcal{P}_x(\omega)d\omega \leq \rm iSNR_0\cdot\mathcal{P}_x. \quad (42)$$

Therefore, we deduce the bounds for the fullband prediction gain of the filtered desired signal with the Wiener filter:

$$0 \leq \mathcal{P}_x(H_{\rm W}) \leq \mathcal{P}_x - \frac{\mathcal{P}_v(H_{\rm W})}{\rm iSNR_0} \leq \mathcal{P}_x, \quad (43)$$

or

$$0 \leq \mathcal{P}_x(H_{\rm W}) \leq \mathcal{P}_x\frac{\rm oSNR}(H_{\rm W})}{1+\rm oSNR}(H_{\rm W})} \leq \mathcal{P}_x, \quad (44)$$

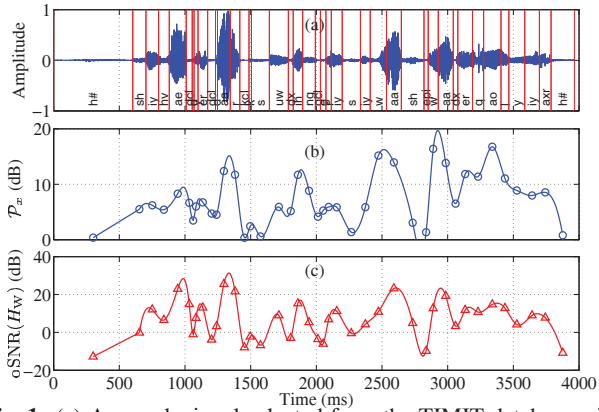

Fig. 1. (a) A speech signal selected from the TIMIT database with phoneme labeling and phoneme boundaries. (b) The fullband prediction gain $\mathcal{P}_x$ for each phoneme. (c) The output SNR for each phoneme with the Wiener filter.

and the bounds for the fullband prediction gain of the residual noise signal:

$$0 \leq \mathcal{P}_v(H_{\rm W}) \leq \mathcal{P}_v \cdot \min(1, \text{iSNR}). \tag{45}$$

We can also deduce from (42) that

$$\text{oSNR}(H_{\rm W}) \leq \text{iSNR} \cdot \xi_{\rm nr}(H_{\rm W}) - 1. \tag{46}$$

As a result,

$$1 \leq \mathcal{G}(H_{\rm W}) \leq \xi_{\rm nr}(H_{\rm W}) - \frac{1}{\text{iSNR}}. \tag{47}$$

Now assume that for a particular frequency ω_0, the noise signal is completely unpredictable [i.e., $\mathcal{P}_v(\omega_0) = 1$]. We have

$$\lim_{\text{iSNR}_0 \to 0} H_{\rm W}(e^{j\omega_0}) = \text{iSNR}_0 \cdot \mathcal{P}_x(\omega_0) = \frac{\text{iSNR}_0}{\left|A_x(e^{j\omega_0})\right|^2}, \tag{48}$$

$$\lim_{\text{iSNR}_0 \to \infty} H_{\rm W}(e^{j\omega_0}) = 1. \tag{49}$$

This shows more explicitly how the Wiener filter, in certain conditions, is mostly related to the AR parameters of the desired signal.

Finally, to finish this part, we assume this time that for a particular frequency ω_0, the desired signal is completely unpredictable [i.e., $\mathcal{P}_x(\omega_0) = 1$]. We find that

$$\lim_{\text{iSNR}_0 \to 0} H_{\rm W}(e^{j\omega_0}) = \frac{\text{iSNR}_0}{\mathcal{P}_v(\omega_0)} = \text{iSNR}_0 \cdot \left|A_v(e^{j\omega_0})\right|^2, \tag{50}$$

$$\lim_{\text{iSNR}_0 \to \infty} H_{\rm W}(e^{j\omega_0}) = 1. \tag{51}$$

We see now how the Wiener filter is mostly related to the AR parameters of the noise signal.

It is instructive to observe how the Wiener filter is clearly dependent on the AR parameters of the desired signal when the noise signal is not predictable, while it is clearly dependent on the AR parameters of the noise signal when the desired signal is not predictable.

4. EXPERIMENTAL RESULTS

The clean speech used in this experiment is from the TIMIT database [7], which was designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition (ASR) systems. Each speech signal in this database is recorded using a 16-kHz sampling rate and is accompanied by manually segmented phonetic (based on 61 phonemes) transcripts. In this experiment, we took one signal from the speaker FAKS0 and downsampled it into 8 kHz. This signal is then used as the clean speech. Figure 1(a) plots this signal and also visualizes both the phonetic transcription and phoneme boundaries. The corresponding noisy signal is generated by adding a white Gaussian noise into the clean speech with $\text{iSNR} = 10$ (dB).

To make the experiment simple, we model each phoneme with a 12-order AR model, and $a_x(k)$ and $a_y(k)$ are directly computed from the clean and noisy speech signals. The fullband prediction gain of each phoneme of the clean speech is shown in Fig. 1(b). Then, a Wiener filter for each phoneme is constructed according to (37) with the computed AR parameters.

To perform noise reduction in the frequency domain, the input speech signal is partitioned into overlapping frames with a frame width of 8 ms and an overlapping factor of 75%. A Kaiser window is then applied to each frame and the windowed frame signal is subsequently transformed into the frequency domain using a 64-point FFT. The noisy speech spectra are then passed through the constructed phoneme-dependent Wiener filter. The inverse FFT (with overlap add) is used to obtain the time-domain speech estimate. To evaluate the performance, the output SNR for each phoneme is computed and shown in Fig. 1(c). Since the white Gaussian noise is used in this experiment, which is completely unpredictable, the SNR improvement of the Wiener filter should depend on the prediction gain of the clean speech, which is clearly verified by Comparing Fig. 1(b) and (c).

5. CONCLUSIONS

In this paper, we presented a theoretical analysis about the performance of the frequency-domain noise-reduction Wiener filter. Specifically, based on the use of the AR model to model both the desired speech and noise signals, we showed the relationship between the frequency-domain Wiener filter and the AR parameters of the desired speech and noise signals. We then developed the connection between the noise-reduction performance (in terms of SNR gain, noise-reduction factor, and speech-distortion index) and the prediction gain. We demonstrated that the Wiener filter can improve the fullband SNR, but the SNR gain would depend on the prediction gains of the clean speech and noise signals. In case that the noise is unpredictable, the Wiener filter is mostly related to the AR parameters of the desired speech signal. On the contrary, if the desired speech signal is unpredictable, the Wiener filter is mostly related to the noise AR parameters. These new insights would be helpful to better understand the performance behavior of the Wiener filter, and may also help develop new ideas on noise reduction.

6. REFERENCES

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

[2] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 1218–1234, July 2006.

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

[4] J. Chen, J. Benesty, and Y. Huang, "On the optimal linear filtering techniques for noise reduction," *Speech Communication*, vol. 49, pp. 305–316, 2007.

[5] S. Haykin, *Adaptive Filter Theory*. Fourth Edition, Upper Saddle River, NJ: Prentice-Hall, 2002.

[6] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[7] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 1641–1648, Nov. 1989.