On the Output SNR of the Speech-Distortion Weighted Multichannel Wiener Filter

Simon Doclo, *Member, IEEE,* and Marc Moonen, *Member, IEEE*

***Abstract*—In this letter, we prove that the output signal-to-noise ratio (SNR) after noise reduction with the speech-distortion weighted multichannel Wiener filter is always larger than or equal to the input SNR, for any filter length, for any value of the tradeoff parameter between noise reduction and speech distortion, and for all possible speech and noise correlation matrices.**

***Index Terms*—Multichannel Wiener filter, output signal-to-noise ratio (SNR).**

I. Introduction

Wiener filtering in the time domain or the frequency domain is a commonly used noise reduction technique for single-channel and multichannel signals [3], e.g., in speech enhancement applications [1], [2], [4]–[8]. The standard Wiener filter minimizes the mean-square error between the filter output signal and the speech component in one of the microphone signals. Hence, the error signal typically consists of a term related to noise reduction and a term related to speech distortion. Whereas the standard Wiener filter assigns equal importance to both terms, a generalized version, the so-called speech-distortion weighted Wiener filter, provides a tradeoff between noise reduction and speech distortion [1], [2], [6].

In [8], it has been proved that the output signal-to-noise ratio (SNR) after noise reduction with the *single-channel Wiener filter* is always larger than or equal to the input SNR, for any filter length and for all possible speech and noise correlation matrices. However, the proof in [8] is quite involved, requiring the generalized eigenvalue decomposition of the speech and the noise correlation matrices and using inductive reasoning. In this letter, we provide a simpler proof of the same statement for the *speech-distortion weighted multichannel Wiener filter* (SDW-MWF) (of which the single-channel Wiener filter is a special case), for any value of the tradeoff parameter between noise reduction and speech distortion. Although this result may appear trivial, it confirms the intuition and expectation about the speech-distortion weighted multichannel Wiener filter, and, to the best of our knowledge, was not available in the literature up until now.

II. SDW-MWF

Consider a microphone array with N microphones, where the nth microphone signal $y_n(k)$ at time k consists of a speech component $x_n(k)$ and a (white or colored) noise component $v_n(k)$, i.e.,

$$y_n(k) = x_n(k) + v_n(k), \quad n = 0, \ldots, N-1.$$

The speech and the noise components in the microphone signals are assumed to be uncorrelated, i.e., $\mathcal{E}\{x_m(k)v_n(k-l)\} = 0, \forall l, m, n$, where $\mathcal{E}$ denotes the expected value operator. The NL-dimensional multichannel Wiener filter $\mathbf{w}_{\mathrm{WF}}^{m,\Delta}$ aims to estimate the delayed speech component $x_m(k-\Delta)$ in the mth microphone signal, with $0 \leq \Delta < L$, by minimizing the mean-square error between the output signal $z(k) = \mathbf{w}^T\mathbf{y}(k)$ and the delayed speech component in the mth microphone, i.e.,

$$J_{\mathrm{WF}}^{m,\Delta}(\mathbf{w}) = \mathcal{E}\left\{\left[\mathbf{w}^T\mathbf{y}(k) - x_m(k-\Delta)\right]^2\right\} \tag{1}$$

where superscript T denotes transpose of a vector or a matrix, $\mathbf{w}$ is the NL-dimensional filter vector [corresponding to N finite impulse response (FIR) filters of length L], and $\mathbf{y}(k)$ is the NL-dimensional stacked data vector defined as

$$\begin{aligned}\mathbf{y}(k) &= \begin{bmatrix}\mathbf{y}_0^T(k) & \mathbf{y}_1^T(k) & \ldots & \mathbf{y}_{N-1}^T(k)\end{bmatrix}^T \\ \mathbf{y}_n(k) &= \begin{bmatrix}y_n(k) & y_n(k-1) & \ldots & y_n(k-L+1)\end{bmatrix}^T.\end{aligned}$$

The speech and the noise vectors $\mathbf{x}(k)$ and $\mathbf{v}(k)$ are defined similarly as $\mathbf{y}(k)$. Solving (1) and using the assumption that the speech and the noise components are uncorrelated results in the well-known multichannel Wiener solution, i.e.,

$$\begin{aligned}\mathbf{w}_{\mathrm{WF}}^{m,\Delta} &= \mathbf{R}_y^{-1}\mathbf{R}_x\mathbf{u}_{(m-1)L+\Delta+1} \\ &= (\mathbf{R}_x + \mathbf{R}_v)^{-1}\mathbf{R}_x\mathbf{u}_{(m-1)L+\Delta+1},\end{aligned}$$

with

$$\begin{aligned}\mathbf{R}_y &= \mathcal{E}\{\mathbf{y}(k)\mathbf{y}^T(k)\} \\ \mathbf{R}_x &= \mathcal{E}\{\mathbf{x}(k)\mathbf{x}^T(k)\} \\ \mathbf{R}_v &= \mathcal{E}\{\mathbf{v}(k)\mathbf{v}^T(k)\}\end{aligned}$$

the $NL \times NL$-dimensional correlation matrices of the noisy microphone signal, the speech component and the noise component, respectively, and with $\mathbf{u}_i$ the ith canonical NL-dimensional vector, i.e., a vector of which the ith element is equal to

Manuscript received June 2, 2005; revised July 9, 2005. This work was carried out at the SCD Research Group of the Katholieke Universiteit Leuven, Belgium, and was supported in part by the F.W.O. Project G.0233.01. Signal processing and automatic patient fitting for advanced auditory prostheses were supported by the I.W.T. Project 020540, Performance improvement of cochlear implants by innovative speech processing algorithms, by the I.W.T. Project 040803, Sound Management System for Public Address systems (SMS4PA-II), by the Concerted Research Action GOA-AMBIORICS, Algorithms for medical and biological research, integration, computation and software, by the Interuniversity Attraction Pole IUAP P5–22, Dynamical Systems and Control: Computation, Identification and Modeling, and was supported in part by Cochlear. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vesa Valimaki.

The authors are with the Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, B-3001 Leuven, Belgium (e-mail: simon.doclo@esat.kuleuven.be; marc.moonen@esat.kuleuven.be).

Digital Object Identifier 10.1109/LSP.2005.859530

1070-9908/$20.00 © 2005 IEEE

1 and all other elements are equal to 0. Since the speech and the noise components in the different microphone signals are assumed to be (short-time) jointly stationary processes, the correlation matrices are symmetric $N \times N$ block matrices consisting of $L \times L$-dimensional Toeplitz matrices. Note that in a typical practical speech enhancement application, the correlation matrix $\mathbf{R}_y$ is estimated during speech periods, and the noise correlation matrix $\mathbf{R}_v$ is estimated during noise-only periods (speech pauses), such that the filter $\mathbf{w}_{\mathrm{WF}}^{m,\Delta}$ is calculated as

$$\mathbf{w}_{\mathrm{WF}}^{m,\Delta} = \mathbf{R}_y^{-1}(\mathbf{R}_y - \mathbf{R}_v)\mathbf{u}_{(m-1)L+\Delta+1}.$$

Since $x(k)$ and $v(k)$ are uncorrelated, the cost function $J_{\mathrm{WF}}^{m,\Delta}(\mathbf{w})$ in (1) can be decomposed as

$$J_{\mathrm{WF}}^{m,\Delta}(\mathbf{w}) = \underbrace{\mathcal{E}\left\{\left[\mathbf{w}^T\mathbf{x}(k) - x_m(k-\Delta)\right]^2\right\}}_{\epsilon_x^2(k)} + \underbrace{\mathcal{E}\{[\mathbf{w}^T\mathbf{v}(k)]^2\}}_{\epsilon_v^2(k)}.$$

This cost function consists of a term $\epsilon_x^2(k)$ related to speech distortion and a term $\epsilon_v^2(k)$ related to noise reduction. Whereas the standard multichannel Wiener filter assigns equal importance to both terms, a generalized version, the so-called SDW-MWF, provides a tradeoff between the noise reduction and the speech distortion term [1], [2], [6]. This generalized cost function can be written as

$$\begin{aligned} J_{\mathrm{SDW}}^{m,\Delta}(\mathbf{w}) &= \epsilon_x^2(k) + \mu\epsilon_v^2(k) \\ &= \mathcal{E}\left\{\left[\mathbf{w}^T\mathbf{x}(k) - x_m(k-\Delta)\right]^2\right\} + \mu\mathcal{E}\left\{\left[\mathbf{w}^T\mathbf{v}(k)\right]^2\right\} \end{aligned} \tag{2}$$

where $\mu \geq 0$ is the tradeoff parameter between noise reduction and speech distortion. The SDW-MWF $\mathbf{w}_{\mathrm{SDW}}^{m,\Delta}$ minimizing this generalized cost function is equal to

$$\mathbf{w}_{\mathrm{SDW}}^{m,\Delta} = (\mathbf{R}_x + \mu\mathbf{R}_v)^{-1}\mathbf{R}_x\mathbf{u}_{(m-1)L+\Delta+1}.$$

If $\mu > 1$, the residual noise level is reduced at the expense of increased signal distortion. On the contrary, if $\mu < 1$, signal distortion is reduced at the expense of decreased noise reduction [1], [2], [6].

III. Input and Output SNR

The input SNR of the mth microphone signal $y_m(k)$ is equal to

$$\mathrm{SNR}_{\mathrm{in}} = \frac{\mathcal{E}\{x_m^2(k)\}}{\mathcal{E}\{v_m^2(k)\}}.$$

Since the speech and the noise components are assumed to be (short-term) stationary processes, the input SNR can also be written as

$$\mathrm{SNR}_{\mathrm{in}} = \frac{\mathcal{E}\{x_m^2(k-\Delta)\}}{\mathcal{E}\{v_m^2(k-\Delta)\}} = \frac{\sigma_x^2}{\sigma_v^2}.$$

The output SNR of the signal $z(k) = (\mathbf{w}_{\mathrm{SDW}}^{m,\Delta})^T\mathbf{y}(k)$, i.e., after noise reduction with the SDW-MWF, is equal to

$$\begin{aligned} \mathrm{SNR}_{\mathrm{out}} &= \frac{\mathcal{E}\left\{\left[\left(\mathbf{w}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{x}(k)\right]^2\right\}}{\mathcal{E}\left\{\left[\left(\mathbf{w}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{v}(k)\right]^2\right\}} \\ &= \frac{\left(\mathbf{w}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{R}_x\mathbf{w}_{\mathrm{SDW}}^{m,\Delta}}{\left(\mathbf{w}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{R}_v\mathbf{w}_{\mathrm{SDW}}^{m,\Delta}}. \end{aligned} \tag{3}$$

In [8], it has been proved that the output SNR after noise reduction with the single-channel Wiener filter is always larger than or equal to the input SNR, for any filter length L and for all possible speech and noise correlation matrices. However, the proof in [8] is quite involved, requiring the generalized eigenvalue decomposition of the speech and the noise correlation matrices and using an inductive reasoning. In the sequel, we will provide a simpler proof of the same statement for the SDW-MWF (of which the single-channel Wiener filter obviously is a special case for $N = 1$ and $\mu = 1$), for any value of the tradeoff parameter μ.

It is always possible to decompose the speech vector $\mathbf{x}(k)$ as

$$\mathbf{x}(k) = \mathbf{a}x_m(k-\Delta) + \mathbf{b}(k) \tag{4}$$

with $\mathbf{a}$ and $\mathbf{b}(k)$ NL-dimensional vectors equal to

$$\begin{aligned} \mathbf{a} &= \frac{\mathcal{E}\{\mathbf{x}(k)x_m(k-\Delta)\}}{\mathcal{E}\{x_m^2(k-\Delta)\}} \\ \mathbf{b}(k) &= \mathbf{x}(k) - \mathbf{a}x_m(k-\Delta). \end{aligned}$$

Note that $\mathbf{a}$ is a constant vector, i.e., independent of k, since the speech components in the different microphone signals are assumed to be jointly stationary processes. Since $\mathbf{b}(k)$ and $x_m(k-\Delta)$ are uncorrelated, i.e.,

$$\begin{aligned} &\mathcal{E}\{\mathbf{b}(k)x_m(k-\Delta)\} \\ &\quad = \mathcal{E}\{\mathbf{x}(k)x_m(k-\Delta)\} - \mathbf{a}\mathcal{E}\{x_m^2(k-\Delta)\} = 0 \end{aligned}$$

the correlation matrix $\mathbf{R}_x$ is equal to

$$\mathbf{R}_x = \sigma_x^2\mathbf{a}\mathbf{a}^T + \mathbf{R}_b \tag{5}$$

with

$$\mathbf{R}_b = \mathcal{E}\{\mathbf{b}(k)\mathbf{b}^T(k)\}.$$

Using (4), the cost function $J_{\mathrm{SDW}}^{m,\Delta}(\mathbf{w})$ in (2) can now be written as

$$\begin{aligned} J_{\mathrm{SDW}}^{m,\Delta}(\mathbf{w}) &= \mathcal{E}\left\{\left[(\mathbf{w}^T\mathbf{a} - 1)x_m(k-\Delta)\right]^2\right\} \\ &\quad + \mathcal{E}\{[\mathbf{w}^T\mathbf{b}(k)]^2\} + \mu\mathcal{E}\{[\mathbf{w}^T\mathbf{v}(k)]^2\} \\ &= (\mathbf{w}^T\mathbf{a} - 1)\sigma_x^2(\mathbf{a}^T\mathbf{w} - 1) + \mathbf{w}^T\mathbf{R}_b\mathbf{w} + \mu\mathbf{w}^T\mathbf{R}_v\mathbf{w}. \end{aligned} \tag{6}$$

In the Appendix, it is shown that the filter $\mathbf{w}_{\mathrm{SDW}}^{m,\Delta}$ minimizing (6) is equal to a scaled version of the filter $\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}$ minimizing the cost function

$$\tilde{J}_{\mathrm{SDW}}^{m,\Delta}(\mathbf{w}) = \mathbf{w}^T\mathbf{R}_b\mathbf{w} + \mu\mathbf{w}^T\mathbf{R}_v\mathbf{w} \tag{7}$$

subject to the constraint $\mathbf{w}^T\mathbf{a} = 1$. Since $\mathbf{u}_{(m-1)L+\Delta+1}$ satisfies the constraint $\mathbf{u}_{(m-1)L+\Delta+1}^T\mathbf{a} = 1$, the inequality

$$\tilde{J}_{\mathrm{SDW}}^{m,\Delta}(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}) \le \tilde{J}_{\mathrm{SDW}}^{m,\Delta}(\mathbf{u}_{(m-1)L+\Delta+1})$$

holds. Therefore, since $\mathbf{u}_{(m-1)L+\Delta+1}^T\mathbf{b}(k) = 0$ and hence $\mathbf{u}_{(m-1)L+\Delta+1}^T\mathbf{R}_b\mathbf{u}_{(m-1)L+\Delta+1} = 0$

$$\left(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{R}_b\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta} + \mu\left(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{R}_v\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta} \le \mu\mathbf{u}_{(m-1)L+\Delta+1}^T\mathbf{R}_v\mathbf{u}_{(m-1)L+\Delta+1}$$

such that

$$\left(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{R}_v\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta} \le \mathbf{u}_{(m-1)L+\Delta+1}^T \cdot \mathbf{R}_v\mathbf{u}_{(m-1)L+\Delta+1} = \sigma_v^2. \tag{8}$$

Using the fact that the filter $\mathbf{w}_{\mathrm{SDW}}^{m,\Delta}$ is a scaled version of the filter $\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}$, the output SNR in (3) is also equal to

$$\mathrm{SNR}_{\mathrm{out}} = \frac{\left(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{R}_x\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}}{\left(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{R}_v\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}}.$$

Hence, using (5) and the fact that $(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta})^T\mathbf{a} = 1$, the output SNR can be written as

$$\mathrm{SNR}_{\mathrm{out}} = \frac{\left(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}\right)^T\left(\sigma_x^2\mathbf{a}\mathbf{a}^T + \mathbf{R}_b\right)\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}}{\left(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{R}_v\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}} = \frac{\sigma_x^2 + \left(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{R}_b\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}}{\left(\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}\right)^T\mathbf{R}_v\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}}.$$

Since the numerator of $\mathrm{SNR}_{\mathrm{out}}$ is larger than or equal to σ_x^2 and, using (8), the denominator of $\mathrm{SNR}_{\mathrm{out}}$ is smaller than or equal to σ_v^2, *the output SNR is always larger than or equal to the input SNR*, i.e.,

$$\mathrm{SNR}_{\mathrm{out}} \ge \mathrm{SNR}_{\mathrm{in}}.$$

It should be noted that the output SNR after noise reduction with the (speech-distortion weighted) Wiener filter in single- and multi-microphone speech enhancement applications generally can be much larger than the input SNR [1], [2], [6], [8]. In fact, the worst-case scenario, where the output SNR is equal to the input SNR, only occurs when the speech components are a scaled version of the noise components, i.e., $x_n(k) = \beta v_n(k), n = 0, \ldots, N-1$. For this scenario, the output SNR is equal to the input SNR for any filtering operation.

IV. Conclusion

In this letter, we have proved that the output SNR after noise reduction with the SDW-MWF is always larger than or equal to the input SNR of the microphone signal of which the speech component is estimated. This theoretical result is a generalization of a similar result for the single-channel Wiener filter and confirms the intuition and expectation about the SDW-MWF.

Appendix

The filter $\mathbf{w}_{\mathrm{SDW}}^{m,\Delta}$ minimizing (6) is equal to

$$\mathbf{w}_{\mathrm{SDW}}^{m,\Delta} = \left(\sigma_x^2\mathbf{a}\mathbf{a}^T + \mathbf{R}_b + \mu\mathbf{R}_v\right)^{-1}\sigma_x^2\mathbf{a}.$$

Using the matrix inversion lemma, this filter can be written as

$$\mathbf{w}_{\mathrm{SDW}}^{m,\Delta} = \left[(\mathbf{R}_b + \mu\mathbf{R}_v)^{-1} - \frac{\sigma_x^2(\mathbf{R}_b + \mu\mathbf{R}_v)^{-1}\mathbf{a}\mathbf{a}^T(\mathbf{R}_b + \mu\mathbf{R}_v)^{-1}}{1 + \sigma_x^2\mathbf{a}^T(\mathbf{R}_b + \mu\mathbf{R}_v)^{-1}\mathbf{a}}\right]\sigma_x^2\mathbf{a} = \frac{\sigma_x^2(\mathbf{R}_b + \mu\mathbf{R}_v)^{-1}\mathbf{a}}{1 + \sigma_x^2\mathbf{a}^T(\mathbf{R}_b + \mu\mathbf{R}_v)^{-1}\mathbf{a}}.$$

The filter $\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}$ minimizing (7), subject to the constraint $\mathbf{w}^T\mathbf{a} = 1$, is equal to

$$\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta} = \frac{(\mathbf{R}_b + \mu\mathbf{R}_v)^{-1}\mathbf{a}}{\mathbf{a}^T(\mathbf{R}_b + \mu\mathbf{R}_v)^{-1}\mathbf{a}}$$

such that

$$\mathbf{w}_{\mathrm{SDW}}^{m,\Delta} = \alpha\tilde{\mathbf{w}}_{\mathrm{SDW}}^{m,\Delta}$$

with

$$\alpha = \frac{\sigma_x^2\mathbf{a}^T(\mathbf{R}_b + \mu\mathbf{R}_v)^{-1}\mathbf{a}}{1 + \sigma_x^2\mathbf{a}^T(\mathbf{R}_b + \mu\mathbf{R}_v)^{-1}\mathbf{a}}.$$

References

[1] 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.

[2] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Process.*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[3] L. L. Scharf, *Statistical Signal Processing : Detection, Estimation and Time Series Analysis*, 1st ed. Reading, MA: Addison-Wesley, Jul. 1991.

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

[5] 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. 32, no. 6, pp. 1109–1121, Dec. 1984.

[6] 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.

[7] E. J. Diethorn, "Subband noise reduction methods for speech enhancement," in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds. Boston, MA: Kluwer, 2000, ch. 9, pp. 155–178.

[8] J. Benesty, J. Chen, A. Huang, and S. Doclo, "Study of the Wiener filter for noise reduction," in *Speech Enhancement*, J. Benesty, J. Chen, and S. Makino, Eds. New York: Springer-Verlag, 2005, ch. 2, pp. 9–42.