Multiframe Echo Suppression Based on Orthogonal Signal Decompositions

Hai Huang[1,2], *Christian Hofmann*[2], *Walter Kellermann*[2], *Jingdong Chen*[1], *and Jacob Benesty*[3]

[1] IAIC Research Center, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[1] Email: huanghai@mail.nwpu.edu.cn, jchen@nwpu.edu.cn
[2] Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, 91058 Erlangen, Germany
[2] Email: {hai.huang, christian.hofmann, walter.kellermann}@fau.de
[3] INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada
[3] Email: benesty@emt.inrs.ca

Abstract

Acoustic echo arises due to the acoustic coupling between a loudspeaker and a microphone in a full-duplex voice communication device. Recently, the use of only acoustic echo suppression (AES) has been proposed without precise echo path estimation. In this paper, we propose an extension to the scheme of the multiframe Wiener filter. A non-parametric variable step-size (NPVSS) normalized-least-mean-square (NLMS) algorithm is used to estimate the echo signal in every STFT bin and time frame. Based on the orthogonal signal decompositions, we discuss how to form different echo suppression filters by optimizing different cost functions. We cover the design of the Wiener, MVDR, tradeoff, and LCMV filters. From a theoretical point of view, the Wiener and tradeoff filters are identical to the MVDR filter up to a scaling factor. Experimental results demonstrate that the developed filters can efficiently attenuate the undesired echo and the estimate near-end signal with little distortion.

1 Introduction

Controlling the detrimental effect of acoustic echoes, resulting from the acoustic coupling between a loudspeaker and a microphone, for applications like teleconferencing and hands-free communication systems has been investigated for several decades and is still an active research topic [1–3]. In the literature, two fundamental techniques, i.e., acoustic echo cancellation (AEC) and acoustic echo suppression (AES), have been developed to eliminate or reduce the undesired echoes.

Since the successful design of the first adaptive echo canceler [4], plenty of adaptation algorithms have been proposed [5], and AEC has reached a mature state nowadays. Ideally, AEC can remove the echoes from the microphone signal without mutilating the desired near-end speech. Usually, AEC is combined with a post-filter (subband suppressor) to eliminate the residual echo [6] which occurs whenever the filter length is not long enough [2], the echo path changes [7], or when there is nonlinearity in the echo path [8].

Alternatively, acoustic echoes can be attenuated by using AES, which acts similarly to the above-mentioned post-filter with completely discarding the AEC part. Generally, AES achieves echo attenuation using parametric spectral modification algorithms on the microphone signal in the frequency domain [9–16]. The major advantages of AES over AEC is robustness against echo path changes and double-talk [10]. Furthermore, the computational complexity of AES is usually also lower than that of AEC. Recently, we proposed a new framework of AES by considering the interframe correlation in STFT domain and derived a parametric Wiener subband filter based on that [17]. By adjusting two parameters in the parametric Wiener filter, we can control the level of the residual and identify a proper tradeoff between the amount of echo attenuation and the degree of near-end speech distortion.

In this paper, we extend the idea in [17] by employing orthogonal signal decompositions to develop different echo suppression filters, such as the Wiener, MVDR, tradeoff, and LCMV filters, that can exploit the interframe information. Section 2 describes the signal model and the formulation of the echo suppression problem in the STFT domain. Then, in Section 3, different echo suppression filters are derived by optimizing different cost functions based on orthogonal signal decompositions. The remaining sections give simulations results of the echo suppression filters and conclusions.

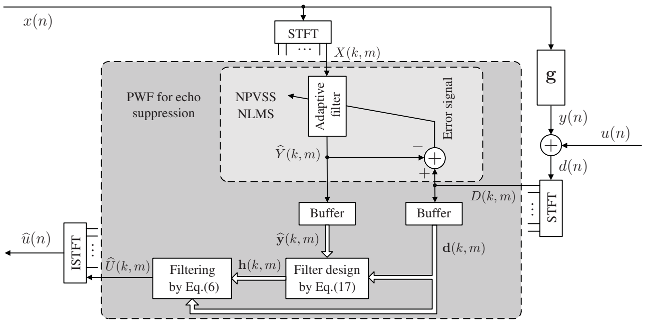

Figure 1: Block diagram of echo suppression by applying an FIR filter in the STFT domain (taken from [17]).

2 Problem Formulation

Let us consider the conventional signal model shown in Fig. 1, where acoustic echoes are generated from the linear coupling between a loudspeaker and a microphone [2]. Therein, the discrete-time microphone signal at time index n can be written as

$$d(n) = u(n) + g(n) * x(n) = u(n) + y(n), \tag{1}$$

where $u(n)$ is the near-end signal, $x(n)$ is the loudspeaker (or far-end) signal, $g(n)$ is the impulse response from the loudspeaker to the microphone, and $y(n)$ is the echo signal. All signals in (1) are considered to be real-valued, zero mean, and broadband and we assume that $u(n)$ and $y(n)$ are uncorrelated.

Using the STFT, the signal model given in (1) can be expressed in the time-frequency domain as

$$D(k,m) = U(k,m) + Y(k,m), \tag{2}$$

where $D(k,m)$, $U(k,m)$, and $Y(k,m)$ are the STFTs of $d(n)$, $u(n)$, and $y(n)$, respectively, at time frame m and frequency bin $k \in \{0, 1, \ldots, K-1\}$.

For the purpose of considering interframe signal correlation later on, we concatenate L consecutive frames of the microphone signal at the frequency bin k to a vector

$$\begin{aligned}\mathbf{d}(k,m) &= [D(k,m)\ D(k,m-1)\ \cdots\ D(k,m-L+1)]^T \\ &= \mathbf{u}(k,m) + \mathbf{y}(k,m),\end{aligned} \tag{3}$$

where the superscript ^{T} denotes transposition, and $\mathbf{u}(k,m)$ and $\mathbf{y}(k,m)$ are defined in an analogous way to $\mathbf{d}(k,m)$.

Since $\mathbf{u}(k,m)$ and $\mathbf{y}(k,m)$ are uncorrelated and of zero mean by assumption, the correlation matrix (of size $L \times L$) of $\mathbf{d}(k,m)$ is

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{d}}(k,m) &\overset{\triangle}{=} \mathscr{E}\left\{\mathbf{d}(k,m)\mathbf{d}^H(k,m)\right\} \\ &= \mathbf{\Phi}_{\mathbf{u}}(k,m) + \mathbf{\Phi}_{\mathbf{y}}(k,m),\end{aligned} \tag{4}$$

where $\mathscr{E}\{\cdot\}$ denotes mathematical expectation, the superscript ^{H} is the transpose-conjugation operator, and $\mathbf{\Phi}_{\mathbf{u}}(k,m)$ and $\mathbf{\Phi}_{\mathbf{y}}(k,m)$ are correlation matrices of $\mathbf{u}(k,m)$ and $\mathbf{y}(k,m)$, respectively, and are defined analogously.

We aim at obtaining an estimate $\widehat{U}(k,m)$ of the desired near-end signal by applying a finite-impulse-response (FIR) filter to the microphone signal in each subband as illustrated in Fig. 1 [17, 18], i.e.,

$$\begin{aligned}\widehat{U}(k,m) &= \mathbf{h}^H(k,m)\mathbf{d}(k,m) \\ &= U_{\mathrm{f}}(k,m) + Y_{\mathrm{re}}(k,m),\end{aligned} \tag{5}$$

[1] This work is supported in part by the NSFC "Distinguished Young Scientists Fund" under Grant No. 61425005. The work of H. Huang is supported in part by the China Scholarship Council.

ISBN 978-3-8007-4275-2

where $\mathbf{h}(k,m)$ is a complex-valued vector of length L, which contains the coefficients of the FIR filter.

$$U_{\mathrm{f}}(k,m) = \mathbf{h}^H(k,m)\mathbf{u}(k,m) \tag{6}$$

is a filtered version of the desired near-end subband signal, and

$$Y_{\mathrm{re}}(k,m) = \mathbf{h}^H(k,m)\mathbf{y}(k,m) \tag{7}$$

is the residual echo, which is uncorrelated with $U_{\mathrm{f}}(k,m)$.

To further analyze the filtering operation and derive statistically optimal filters, consider the normalized (subband) interframe correlation vector

$$\begin{aligned}\boldsymbol{\gamma}_U(k,m) &= \frac{\mathscr{E}\left\{\mathbf{u}(k,m)U^*(k,m)\right\}}{\phi_U(k,m)} &(8)\\ &= \left[\gamma_U(k,m,0),\ldots,\gamma_U(k,m,L-1)\right]^T, &(9)\end{aligned}$$

where the superscript * stands for the complex conjugation and $\phi_U(k,m) \triangleq \mathscr{E}\left\{|U(k,m)|^2\right\}$ is the variance of $U(k,m)$. As done in [19], exploiting the stationarity of the signal $U(k,m)$, a backward prediction of $U(k,m-l)$ decomposes it into two orthogonal components

$$\begin{aligned}U_{\mathrm{c}}(k,m,l) &= U(k,m)\gamma_U(k,m,l), &(10)\\ U_{\mathrm{i}}(k,m,l) &= U(k,m) - U(k,m)\gamma_U(k,m,l), &(11)\end{aligned}$$

where $U_{\mathrm{c}}(k,m)$ is the prediction component and $U_{\mathrm{i}}(k,m,l)$ is the prediction error component. Stacking this prediction and prediction errors into the vectors

$$\begin{aligned}\mathbf{u}_{\mathrm{c}}(k,m) &= U(k,m)\boldsymbol{\gamma}_U(k,m), &(12)\\ \mathbf{u}_{\mathrm{i}}(k,m) &= \mathbf{u}(k,m) - U(k,m)\boldsymbol{\gamma}_U(k,m), &(13)\end{aligned}$$

repectively, $\mathbf{u}(k,m)$ can be decomposed according to

$$\mathbf{u}(k,m) = \mathbf{u}_{\mathrm{c}}(k,m) + \mathbf{u}_{\mathrm{i}}(k,m), \tag{14}$$

as done in [19]. In the following, $\mathbf{u}_{\mathrm{c}}(k,m)$ and $\mathbf{u}_{\mathrm{i}}(k,m)$ will be referred to as prediction and prediction error components of the near-end signal, respectively. This decomposition will allow for an explicit protection the prediction component of the near-end signal later on.

Substituting (14) into (5), we get

$$\begin{aligned}\widehat{U}(k,m) &= \mathbf{h}^H(k,m)[U(k,m)\boldsymbol{\gamma}_U(k,m) +\\ &\qquad \mathbf{u}_{\mathrm{i}}(k,m) + \mathbf{y}(k,m)]\\ &= U_{\mathrm{fd}}(k,m) + U_{\mathrm{ri}}(k,m) + Y_{\mathrm{re}}(k,m), &(15)\end{aligned}$$

where $U_{\mathrm{fd}}(k,m) \triangleq U(k,m)\mathbf{h}^H(k,m)\boldsymbol{\gamma}_U(k,m)$ is a filtered version of the prediction component and $U_{\mathrm{ri}}(k,m) \triangleq \mathbf{h}^H(k,m)\mathbf{u}_{\mathrm{i}}(k,m)$ is the filtered version of the prediction error of the near-end signal. Comparing (15) to (5), one can see that the individual contributions of the prediction and the prediction error near-end signal component to its estimate can be assessed individually now. This will be beneficial for analyzing filters exploiting subband correlation to protect speech later on.

Since the three terms on the right-hand side of (15) are mutually incoherent, the variance of $\widehat{U}(k,m)$ is

$$\phi_{\widehat{U}}(k,m) = \phi_{U_{\mathrm{fd}}}(k,m) + \phi_{U_{\mathrm{ri}}}(k,m) + \phi_{Y_{\mathrm{re}}}(k,m), \tag{16}$$

where

$$\begin{aligned}\phi_{U_{\mathrm{fd}}}(k,m) &= \phi_U(k,m)\left|\mathbf{h}^H(k,m)\boldsymbol{\gamma}_U(k,m)\right|^2\\ &= \mathbf{h}^H(k,m)\boldsymbol{\Phi}_{\mathbf{u}_{\mathrm{c}}}(k,m)\mathbf{h}(k,m), &(17a)\\ \phi_{U_{\mathrm{ri}}}(k,m) &= \mathbf{h}^H(k,m)\boldsymbol{\Phi}_{\mathbf{u}_{\mathrm{i}}}(k,m)\mathbf{h}(k,m)\\ &= \mathbf{h}^H(k,m)\boldsymbol{\Phi}_{\mathbf{u}}(k,m)\mathbf{h}(k,m) -\\ &\qquad \phi_U(k,m)\left|\mathbf{h}^H(k,m)\boldsymbol{\gamma}_U(k,m)\right|^2, &(17b)\\ \phi_{Y_{\mathrm{re}}}(k,m) &= \mathbf{h}^H(k,m)\boldsymbol{\Phi}_{\mathbf{y}}(k,m)\mathbf{h}(k,m), &(17c)\end{aligned}$$

where $\boldsymbol{\Phi}_{\mathbf{u}_{\mathrm{c}}}(k,m) = \phi_U(k,m)\boldsymbol{\gamma}_U(k,m)\boldsymbol{\gamma}_U^H(k,m)$ is the instantaneous correlation matrix of $\mathbf{u}_{\mathrm{c}}(k,m)$ and $\boldsymbol{\Phi}_{\mathbf{u}_{\mathrm{i}}}(k,m) = \mathscr{E}\left\{\mathbf{u}_{\mathrm{i}}(k,m)\mathbf{u}_{\mathrm{i}}^H(k,m)\right\}$ is the correlation matrix of $\mathbf{u}_{\mathrm{i}}(k,m)$.

Furthermore, the subband error signal between the estimated and desired near-end signals can be written as

$$\begin{aligned}\varepsilon(k,m) &= \widehat{U}(k,m) - U(k,m)\\ &= \varepsilon_{\mathrm{d}}(k,m) + \varepsilon_{\mathrm{r}}(k,m), &(18)\end{aligned}$$

where

$$\begin{aligned}\varepsilon_{\mathrm{d}}(k,m) &\triangleq U_{\mathrm{fd}}(k,m) - U(k,m)\\ &= U(k,m)\left[\mathbf{h}^H(k,m)\boldsymbol{\gamma}_U(k,m) - 1\right] &(19)\end{aligned}$$

is the signal distortion due to the FIR filter and

$$\varepsilon_{\mathrm{r}}(k,m) \triangleq U_{\mathrm{ri}}(k,m) + Y_{\mathrm{re}}(k,m) \tag{20}$$

represents the prediction-error-plus-echo residual. The MSE cost function is then

$$\begin{aligned}J\left[\mathbf{h}(k,m)\right] &\triangleq \mathscr{E}\left\{|\varepsilon(k,m)|^2\right\}\\ &= J_{\mathrm{d}}\left[\mathbf{h}(k,m)\right] + J_{\mathrm{r}}\left[\mathbf{h}(k,m)\right], &(21)\end{aligned}$$

where

$$\begin{aligned}J_{\mathrm{d}}\left[\mathbf{h}(k,m)\right] &= \mathscr{E}\left\{|\varepsilon_{\mathrm{d}}(k,m)|^2\right\}\\ &= \phi_U(k,m)\left|\mathbf{h}^H(k,m)\boldsymbol{\gamma}_U(k,m) - 1\right|^2 &(22)\\ J_{\mathrm{r}}\left[\mathbf{h}(k,m)\right] &= \mathscr{E}\left\{|\varepsilon_{\mathrm{r}}(k,m)|^2\right\}\\ &= \mathbf{h}^H(k,m)\boldsymbol{\Phi}_{\mathrm{in}}(k,m)\mathbf{h}(k,m) &(23)\end{aligned}$$

and

$$\boldsymbol{\Phi}_{\mathrm{in}}(k,m) = \boldsymbol{\Phi}_{\mathbf{u}_{\mathrm{i}}}(k,m) + \boldsymbol{\Phi}_{\mathbf{y}}(k,m) \tag{24}$$

is the prediction-error-plus-echo residual correlation matrix.

It is clear that the objective of echo suppression by considering the interframe correlation is to find optimal filters $\mathbf{h}(k,m)$ at each frequency-bin k and time-frame m that would either minimize $J\left[\mathbf{h}(k,m)\right]$ or minimize $J_{\mathrm{d}}\left[\mathbf{h}(k,m)\right]$ or $J_{\mathrm{r}}\left[\mathbf{h}(k,m)\right]$ with some constraint.

3 Optimal Filters

In this section, we will employ the orthogonal decomposition of the near-end signal in the STFT domain to derive commonly used filters for extracting the desired near-end speech and attenuate the echo component. In particular, we will consider the Wiener, MVDR, and tradeoff filters. Furthermore, by decomposing the echo term in a similar way, we can deduce an LCMV filter, which can handle more than one linear constraint.

3.1 Wiener

The Wiener filter is easily derived by taking the gradient of the MSE $J\left[\mathbf{h}(k,m)\right]$, defined in (21), with respect to $\mathbf{h}(k,m)$ and equating the result to zero [17]:

$$\begin{aligned}\mathbf{h}_{\mathrm{W}}(k,m) &= \boldsymbol{\Phi}_{\mathbf{d}}^{-1}(k,m)\boldsymbol{\Phi}_{\mathbf{u}}(k,m)\mathbf{i}_0\\ &= \left[\mathbf{I} - \boldsymbol{\Phi}_{\mathbf{d}}^{-1}(k,m)\boldsymbol{\Phi}_{\mathbf{y}}(k,m)\right]\mathbf{i}_0, &(25)\end{aligned}$$

where $\mathbf{I}$ is the identity matrix of size $L \times L$, and $\mathbf{i}_0$ is the first column of $\mathbf{I}$. Since

$$\boldsymbol{\Phi}_{\mathbf{u}}(k,m)\mathbf{i}_0 = \phi_U(k,m)\boldsymbol{\gamma}_U(k,m), \tag{26}$$

we can rewrite (25) as

$$\mathbf{h}_{\mathrm{W}}(k,m) = \phi_U(k,m)\boldsymbol{\Phi}_{\mathbf{d}}^{-1}(k,m)\boldsymbol{\gamma}_U(k,m). \tag{27}$$

From Section 2, it is easy to verify that

$$\boldsymbol{\Phi}_{\mathbf{d}}(k,m) = \phi_U(k,m)\boldsymbol{\gamma}_U(k,m)\boldsymbol{\gamma}_U^H(k,m) + \boldsymbol{\Phi}_{\mathrm{in}}(k,m). \tag{28}$$

Determining the inverse of $\boldsymbol{\Phi}_{\mathbf{d}}(k,m)$ from (28) with the Woodbury's identity [20], and substituting the result into (27), we get another interesting formulation of the Wiener filter:

$$\mathbf{h}_{\mathrm{W}}(k,m) = \frac{\phi_U(k,m)\boldsymbol{\Phi}_{\mathrm{in}}^{-1}(k,m)\boldsymbol{\gamma}_U(k,m)}{1 + \phi_U(k,m)\boldsymbol{\gamma}_U^H(k,m)\boldsymbol{\Phi}_{\mathrm{in}}^{-1}(k,m)\boldsymbol{\gamma}_U(k,m)}. \tag{29}$$

If we set $L = 1$, the Wiener filter reduces to a simple Wiener gain as in [9–17].

ISBN 978-3-8007-4275-2

3.2 MVDR

The frequently used minimum variance distortionless response (MVDR) filter, originally proposed by Capon [21], is usually derived in the context of microphone arrays. Remarkably, we can derive such an MVDR filter with a single sensor only with our framework considering the interframe signal correlations. We minimize the MSE of the prediction-error-plus-echo residual, $J_\text{r}\left[\mathbf{h}(k,m)\right]$ in (23), with the constraint that the desired signal is not distorted. Mathematically, this is equivalent to

$$\begin{aligned}&\min_{\mathbf{h}(k,m)} \mathbf{h}^H(k,m)\mathbf{\Phi}_\text{in}(k,m)\mathbf{h}(k,m)\\ &\text{subject to } \mathbf{h}^H(k,m)\boldsymbol{\gamma}_U(k,m)=1,\end{aligned} \tag{30}$$

for which the solution is

$$\mathbf{h}_\text{MVDR}(k,m)=\frac{\mathbf{\Phi}_\text{in}^{-1}(k,m)\boldsymbol{\gamma}_U(k,m)}{\boldsymbol{\gamma}_U^H(k,m)\mathbf{\Phi}_\text{in}^{-1}(k,m)\boldsymbol{\gamma}_U(k,m)}. \tag{31}$$

By the no distortion constraint in (30), $U_\text{fd}(k,m)=U(k,m)$ becomes the desired signal and $U_\text{ri}(k,m)$ takes the role of an interfering signal. Hence, the ratio between the desired signal and all interfering components (the prediction-error-plus-echo residual) at the filter output will be referred to as output signal-to-undesired ratio (SUR)

$$\begin{aligned}\text{oSUR}\left[\mathbf{h}(k,m)\right]&=\frac{\phi_{U_\text{fd}}(k,m)}{\phi_{U_\text{ri}}(k,m)+\phi_{Y_\text{re}}(k,m)}\\ &=\frac{\mathbf{h}^H(k,m)\mathbf{\Phi}_{\mathbf{u}_\text{c}}(k,m)\mathbf{h}(k,m)}{\mathbf{h}^H(k,m)\mathbf{\Phi}_\text{in}(k,m)\mathbf{h}(k,m)},\end{aligned} \tag{32}$$

using (31), we deduce that the output SUR is results in

$$\begin{aligned}\text{oSUR}\left[\mathbf{h}_\text{MVDR}(k,m)\right]&=\phi_U(k,m)\boldsymbol{\gamma}_U^H(k,m)\cdot\\ \mathbf{\Phi}_\text{in}^{-1}(k,m)\boldsymbol{\gamma}_U(k,m)&\triangleq\lambda_\text{max}(k,m).\end{aligned} \tag{33}$$

Note that the right-hand side of (32) is known as the generalized Rayleigh quotient [20], and $\lambda_\text{max}(k,m)$ is the maximum eigenvalue of the matrix $\mathbf{\Phi}_\text{in}^{-1}(k,m)\mathbf{\Phi}_{\mathbf{u}_\text{c}}(k,m)$; meanwhile $\mathbf{h}_\text{MVDR}(k,m)$ achieves the maximum output SUR.

Comparing (29) and (31) reveals that the Wiener and MVDR filters are simply related by

$$\mathbf{h}_\text{W}(k,m)=\alpha_{\mathbf{h}_\text{w}}(k,m)\mathbf{h}_\text{MVDR}(k,m), \tag{34}$$

where $\alpha_{\mathbf{h}_\text{w}}(k,m)=\lambda_\text{max}(k,m)/\left(1+\lambda_\text{max}(k,m)\right)$. Obviously, the MVDR filter does not distort the desired near-end signal in theory.

3.3 Tradeoff

Similar to the tradeoff filter for noise reduction in [19], we will perform a tradeoff between echo attenuation and speech distortion. To this end, we consider a cost function

$$J_{\text{T},\mu}\left[\mathbf{h}(k,m)\right]\triangleq J_\text{d}\left[\mathbf{h}(k,m)\right]+\mu J_\text{r}\left[\mathbf{h}(k,m)\right], \tag{35}$$

which is the weighted superposition of the speech distortion index $J_\text{d}\left[\mathbf{h}(k,m)\right]$ in (22) and the energy of the prediction-error-plus-echo residual $J_\text{r}\left[\mathbf{h}(k,m)\right]$ in (23), and where the tradeoff parameter $\mu>0$ controls the proportion between the echo attenuation and speech distortion. By minimizing $J_{\text{T},\mu}\left[\mathbf{h}(k,m)\right]$ with respect to $\mathbf{h}(k,m)$, we obtain the echo-suppression tradeoff filter as

$$\begin{aligned}\mathbf{h}_{\text{T},\mu}(k,m)&=\phi_U(k,m)[\phi_U(k,m)\boldsymbol{\gamma}_U(k,m)\boldsymbol{\gamma}_U^H(k,m)\\ &\quad+\mu\mathbf{\Phi}_\text{in}(k,m)]^{-1}\boldsymbol{\gamma}_U(k,m)\\ &=\frac{\phi_U(k,m)\mathbf{\Phi}_\text{in}^{-1}(k,m)\boldsymbol{\gamma}_U(k,m)}{\mu+\lambda_\text{max}(k,m)}.\end{aligned} \tag{36}$$

Again, we observe here as well that the tradeoff and MVDR filters are equivalent up to a scaling factor, i.e.,

$$\mathbf{h}_{\text{T},\mu}(k,m)=\alpha_{\text{T},\mu}(k,m)\mathbf{h}_\text{MVDR}(k,m), \tag{37}$$

where $\alpha_{\text{T},\mu}(k,m)=\lambda_\text{max}(k,m)/\left(\mu+\lambda_\text{max}(k,m)\right)$.

Note that the tradeoff filter does not provide an analytical expression the optimum tradeoff parameter μ — it has to be determined heuristically. Interestingly, there are two special cases:

- $\mu=1$, $\mathbf{h}_{\text{T},1}(k,m)=\mathbf{h}_\text{W}(k,m)$, which is the Wiener filter derived in Section 3.1, and
- $\mu=0$, $\mathbf{h}_{\text{T},0}(k,m)=\mathbf{h}_\text{MVDR}(k,m)$, which is the MVDR filter derived in Section 3.2.

In general, increasing μ results in lower residual echo at the expense of higher speech distortion and vice versa for a decreasing μ.

3.4 LCMV

We can derive a linearly constrained minimum variance (LCMV) filter [22] which can handle more than one linear constraint, by exploiting the structure of the echo signal as well as the desired near-end speech.

In Section 2, we decomposed $\mathbf{u}(k,m)$ into two orthogonal components. Additionally, we can also decompose the echo signal vector $\mathbf{y}(k,m)$ into orthogonal terms according to

$$\mathbf{y}(k,m)=Y(k,m)\boldsymbol{\gamma}_Y(k,m)+\mathbf{y}_\text{i}(k,m), \tag{38}$$

where $\boldsymbol{\gamma}_Y(k,m)$ and $\mathbf{y}_\text{i}(k,m)$ are defined in a similar way as $\boldsymbol{\gamma}_U(k,m)$ and $\mathbf{u}_\text{i}(k,m)$. The objective of the LCMV filter is to perfectly recover the desired near-end speech component $U(k,m)$, and to completely attenuate the predictable echo component $Y(k,m)\boldsymbol{\gamma}_Y(k,m)$. Putting the two constraints together in a matrix form as

$$\mathbf{\Gamma}^H(k,m)\mathbf{h}(k,m)=\mathbf{i}, \tag{39}$$

where $\mathbf{\Gamma}(k,m)=\left[\boldsymbol{\gamma}_U(k,m)\ \ \boldsymbol{\gamma}_Y(k,m)\right]$ is our constraint matrix of size $L\times 2$, and $\mathbf{i}=\left[1\ \ 0\right]^T$.

Then, the optimal filter is obtained by minimizing the energy at the filter output with the constraints in (39), i.e.,

$$\begin{aligned}&\min_{\mathbf{h}(k,m)} \mathbf{h}^H(k,m)\mathbf{\Phi}_\mathbf{d}(k,m)\mathbf{h}(k,m)\\ &\text{subject to } \mathbf{\Gamma}^H(k,m)\mathbf{h}(k,m)=\mathbf{i}.\end{aligned} \tag{40}$$

The solution to (40) is given by

$$\begin{aligned}\mathbf{h}_\text{LCMV}(k,m)&=\mathbf{\Phi}_\mathbf{d}^{-1}(k,m)\mathbf{\Gamma}(k,m)\cdot\\ &\quad\left[\mathbf{\Gamma}^H(k,m)\mathbf{\Phi}_\mathbf{d}^{-1}(k,m)\mathbf{\Gamma}(k,m)\right]^{-1}\mathbf{i},\end{aligned} \tag{41}$$

assuming that $\mathbf{\Gamma}^H(k,m)\mathbf{\Phi}_\mathbf{d}^{-1}(k,m)\mathbf{\Gamma}(k,m)$ is invertible.

4 Experimental Results

4.1 Experimental Setup

In our experimental setup, the far-end and near-end speech signals are recorded in an anechoic room from male and female talkers with a sampling rate of 8 kHz, separately. The echo signal is obtained by convolving the far-end signal with a measured impulse response from an acoustic chamber of size 6.7 m × 6.1 m × 2.9 m with a reverberation time $T_{60}\approx 380$ ms and a loudspeaker-microphone distance of 1.5 m. The microphone signal is then synthesized by superimposing this echo with near-end speech at a near-end-signal-to-echo ratio (NER) of 5 dB and with white Gaussian noise at a signal-to-noise ratio (SNR) of 30 dB.

The subband processing is done with a DFT-modulated filterbank with a DFT length of $K=256$ and 75% overlap between neighboring frames and a Kaiser window both in the analysis and synthesis stage.

The implementation of the echo suppression filters derived in Section 3 requires the estimation of the correlation matrices $\mathbf{\Phi}_\mathbf{d}(k,m)$, $\mathbf{\Phi}_\mathbf{y}(k,m)$, and $\mathbf{\Phi}_\mathbf{u}(k,m)$, the normalized interframe correlation vectors $\boldsymbol{\gamma}_U(k,m)$ and $\boldsymbol{\gamma}_Y(k,m)$, and the signal variance $\phi_U(k,m)$. In all experiments, an NPVSS NLMS algorithm is used to estimate the echo component $\widehat{Y}(k,m)$ in every frequency bin k and time frame m in the STFT domain [6, 23]. The $\mathbf{\Phi}_\mathbf{d}(k,m)$ and $\mathbf{\Phi}_{\widehat{\mathbf{y}}}(k,m)$ matrices are then computed with a rank-1 update approach [24]:

$$\mathbf{\Phi}_\mathbf{z}(k,m)=\lambda\mathbf{\Phi}_\mathbf{z}(k,m-1)+(1-\lambda)\mathbf{z}(k,m)\mathbf{z}^H(k,m), \tag{42}$$

where $\lambda\in(0,1)$ is a forgetting factor, and $\mathbf{z}(k,m)\in\{\mathbf{d}(k,m),\widehat{\mathbf{y}}(k,m)\}$. Note that in our simulations, we use the first 100 frames

ISBN 978-3-8007-4275-2

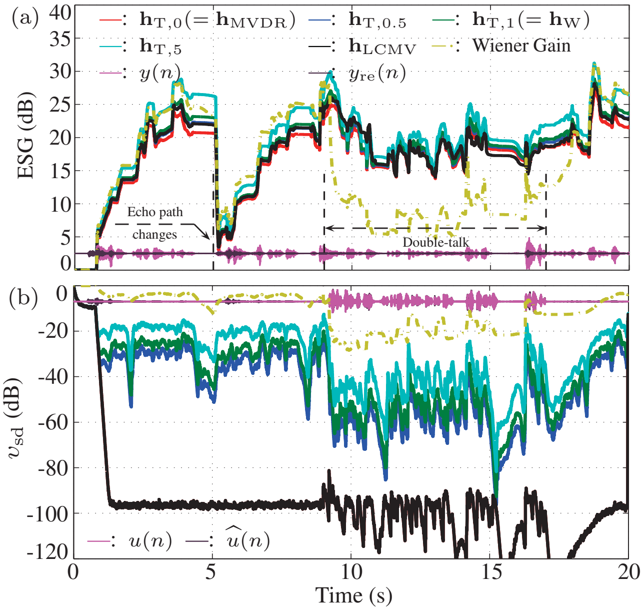

Figure 2: Performance of different echo suppression filters, i.e., the tradeoff filter with the vary value of μ from 0 to 5, and the LCMV filter: (a) ESG and (b) speech distortion index $\upsilon_{\rm sd}$. $\lambda = 0.35$ and $L = 4$. The echo path changes at 5 seconds, and the situation of double-talk is considered between 9 and 17 seconds.

to compute the initial estimates of the $\mathbf{\Phi}_{\mathbf{d}}(k,m)$ and $\mathbf{\Phi}_{\widehat{\mathbf{y}}}(k,m)$ matrices with a short-time average. Finally, all the other parameters are estimated in the following way: $\mathbf{\Phi}_{\mathbf{u}}(k,m) = \mathbf{\Phi}_{\mathbf{d}}(k,m) - \mathbf{\Phi}_{\widehat{\mathbf{y}}}(k,m)$, $\phi_U(k,m)$ is equal to the first element of $\mathbf{\Phi}_{\mathbf{u}}(k,m)$, and $\boldsymbol{\gamma}_U(k,m)$ and $\boldsymbol{\gamma}_Y(k,m)$ are taken as the first column of the corresponding correlation matrices $\mathbf{\Phi}_{\mathbf{u}}(k,m)$ and $\mathbf{\Phi}_{\widehat{\mathbf{y}}}(k,m)$ normalized by their first elements, respectively.

To evaluate the echo suppression performance of the designed filters, we use three performance metrics: echo suppression gain (ESG), near-end speech distortion index, and and the perceptual evaluation of speech quality (PESQ) measure [25]. The ESG, which is also called echo-return loss enhancement (ERLE) in AEC [1], is defined as

$$\mathrm{ESG}(n) = 10\log_{10}\frac{\mathscr{E}\left\{y^2(n)\right\}}{\mathscr{E}\left\{y_{\rm re}^2(n)\right\}}, \tag{43}$$

and the near-end speech distortion index [18] is given by

$$\upsilon_{\rm sd}(n) = 10\log_{10}\frac{\mathscr{E}\left\{[u(n)-u_{\rm fd}(n)]^2\right\}}{\mathscr{E}\left\{u^2(n)\right\}}, \tag{44}$$

where $y_{\rm re}(n)$ and $u_{\rm fd}(n)$ are the time-domain signals reconstructed from $Y_{\rm re}(k,m)$ and $U_{\rm fd}(k,m)$, respectively.

Through repeated experiments, we find that the best performance is achieved with the forgetting factor $\lambda = 0.35$ and the length of the FIR filters $L = 4$, so they are set as the basic experiment conditions in our simulation setup. In order to evaluate the tracking capabilities of all the designed filters, an echo path change is simulated in the experiments by shifting the impulse responses in the near-end location to the right by 100 samples at 5 seconds. The problem of double-talk detection has been widely studied in the literature but is out of the scope of this paper. Therefore, an "ideal" detector is used to stop the filter adaptation during double-talk periods between 9 and 17 seconds.

4.2 Simulation Results

In the following, we investigate the performance of the tradeoff filter for four conditions: $\mu = 0$, 0.5, 1, and 5. Recall that, according to Section 3.3, a tradeoff filter with $\mu = 0$ corresponds to the MVDR filter and $\mu = 1$ leads to a Wiener filter. The performance of the LCMV filter is also considered here. The experimental result of the traditional Wiener filter (with $L = 1$, as mentioned in Section 3.1) is plotted for comparison.

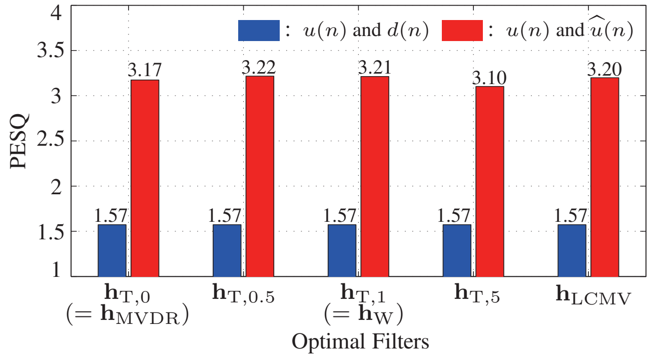

Figure 3: The PESQ score of different echo suppression filters with $\lambda = 0.35$ and $L = 4$. The blue bars are the PESQ score calculated between $u(n)$ and the raw microphone signal $d(n)$, whereas the red bars are the PESQ score calculated between $u(n)$ and its estimates $\widehat{u}(n)$ from the developed echo suppression filters.

Fig. 2 depicts ESG and speech distortion index for these filters. To visualize the echo suppression performance, the residual echo and estimated near-end signals obtained with the Wiener filter are also added in the subfigures (a) and (b) of Fig. 2, respectively. As can be seen, as the tradeoff parameter μ increases from 0 to 5, higher ESG is obtained at the output of the tradeoff filter, but the speech distortion index increases as well. This provides a nice way to make a compromise between echo attenuation and near-end speech distortion. Comparing the performance of the Wiener, MVDR and LCMV filters, reveals that the MVDR achieves less ESG than the Wiener filter, but preserves the desired near-end speech without much distortion, as expected. The LCMV does not distort the desired near-end speech either, while still achieving higher ESG than the MVDR, due to the additional constraint on the unwanted echo component. Fig. 2 also shows the behavior after an abrupt system changes at 5 seconds. In this case, the tradeoff filter with larger value of μ tracks faster than its smaller value counterpart. For all of the designed filters, we notice that they can achieve more than 25 dB echo attenuation and the near-end speech distortion is less than -40 dB during the double-talk periods, which can satisfy the requests of many applications. Comparing the results of the designed filters with $L = 4$ with those of the traditional Wiener gain with $L = 1$, one can see that the reconvergence rate of the traditional Wiener gain (the dashed yellow line) is a little bit faster than the designed filters, but it yields a much lower ESG and much more speech distortion to the desired near-end speech than the designed filters. This shows the advantages of considering the interframe correlation information in echo suppression.

Furthermore, we evaluate the quality of the desired near-end speech estimated by different filters through the PESQ measure and the result are presented in Fig. 3. For the tradeoff filter, one can see that the PESQ score first increases and then decreases as the value of μ increases. This is because both ESG and speech distortion index increase with μ: For a smaller μ, the speech distortion index is very low and increasing μ can help obtaining higher ESG, thereby improving the PESQ score. Nonetheless, when increasing the value of μ continuously, the near-end speech distortion becomes higher as well and becomes the principal factor degrading the PESQ performance. The highest PESQ is obtained when μ is varied from 0.5 to 1, that is why we are interested in the tradeoff filter. Meanwhile, the PESQ performance of the LCMV is located between the MVDR and Wiener filters.

5 Conclusions

In this paper, we considered the problem of AES by considering the interframe correlation in the STFT domain. Based on the orthogonal signal decompositions, we reformulated the Wiener, MVDR, tradeoff, and LCMV filters by solving different optimization problems. Through theoretical analysis, we demonstrated that the Wiener and tradeoff filters are equivalent to the MVDR filter up to a scaling factor. By adjusting the tradeoff parameter μ, a reasonable compromise between the amount of echo attenuation and the degree of near-end speech distortion is achievable. The theoretical studies were finally corroborated by the simulation of actual AES systems.

ISBN 978-3-8007-4275-2

References

[1] C. Breining, P. Dreiscitel, E. Gänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control–an application of very-high-order adaptive filters," *IEEE Signal Process. Mag.*, vol. 16, pp. 42–69, July 1999.

[2] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. Berlin, Germany: Springer-Verlag, 2001.

[3] M. M. Sohdhi, "Adaptive echo cancelation for voice signals," in *Springer Handbook of Speech Processing* (J. Benesty, M. M. Sondhi, and Y. Huang, eds.), pp. 903–927, Berlin, Germany: Springer-Verlag, 2007.

[4] M. M. Sondhi, "An adaptive echo canceller," *Bell Syst. Tech. J.*, vol. 46, pp. 497–511, Mar. 1967.

[5] S. Haykin, *Adaptive Filter Theory,* 4th ed. Upper Saddle River, NJ: Prentice-Hall, 2008.

[6] E. A. P. Habets, S. Gannot, and I. Cohen, "Robust early echo cancellation and late echo suppression in the STFT domain," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 14–17, 2008.

[7] G. W. Elko, E. Diethorn, and T. Gänsler, "Room impulse response variation due to thermal fluctuation and its impact on acoustic echo cancellation," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 67–70, 2003.

[8] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral feature-based nonlinear residual echo suppression," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–4, 2013.

[9] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 175–178, 2001.

[10] F. Wallin and C. Faller, "Perceptual quality of hybrid echo canceler/suppressor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 157–160, 2004.

[11] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Trans. Speech, Audio Process.*, vol. 13, pp. 1048–1062, Sep. 2005.

[12] C. Faller and C. Tournery, "Estimating the delay and coloration effect of the acoustic echo path for low complexity echo suppression," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 53–56, 2005.

[13] C. Faller and C. Tournery, "Robust acoustic echo control using a simple echo path model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 281–284, 2006.

[14] A. Favrot, C. Faller, M. Kallinger, and M. Schmidt, "Modeling late reverberation in acoustic echo suppression," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 1–4, 2008.

[15] A. Favrot, C. Faller, and F. Kuech, "Modeling late reverberation in acoustic echo suppression," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 1–4, 2012.

[16] Y. S. Park and J. H. Chang, "Frequency domain acoustic echo suppression based on soft decision," *IEEE Signal Process. Lett.*, vol. 16, pp. 53–56, Jan. 2009.

[17] H. Huang, C. Hofmann, W. Kellermann, J. Chen, and J. Benesty, "A multiframe parametric wiener filter for acoustic echo suppression," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)* 2016, accepted.

[18] H. Huang, J. Benesty, J. Chen, K. Helwani, and H. Buchner, "A study of the MVDR filter for acoustic echo suppression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 615–619, 2013.

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

[20] G. H. Golub and C. F. Van Loan, *Matrix Computations, Fourth Edition*. Johns Hopkins University Press, 2013.

[21] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.

[22] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926–935, Aug. 1972.

[23] J. Benesty, H. Rey, L. R. Vega, and S. Tressens, "A nonparametric VSS NLMS algorithm," *IEEE Signal Process. Lett.*, vol. 13, pp. 581–684, Oct. 2006.

[24] K. Yu, "Recursive updating the eigenvalue decomposition of a covariance matrix," *IEEE Trans. Signal Process.*, vol. 39, pp. 1136–1145, May 1991.

[25] ITU-T P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recommendation P.862*, 2001.

ISBN 978-3-8007-4275-2