RESIDUAL ECHO POWER SPECTRAL DENSITY ESTIMATION BASED ON AN OPTIMAL SMOOTHED MISALIGNMENT FOR ACOUSTIC ECHO CANCELATION

[1]*Stefan Goetze,* [2]*Markus Kallinger,* [1]*Karl-Dirk Kammeyer*

[1]goetze@ant.uni-bremen.de [2]markus.kallinger@uni-oldenburg.de
[1]University of Bremen, FB 1, Dept. of Communications Engineering,
P.O. Box 330 440, D-28334 Bremen, Germany
[2]Carl von Ossietzky-University, Oldenburg, Signal Processing Group,
D-26111 Oldenburg, Germany

ABSTRACT

For echo cancelation enhanced by a Post-Filter it is necessary to get a reliable estimate of the (residual) echo power spectral density (PSD). We describe a method to obtain the echo PSD by an estimation of the system misalignment. For room impulse responses longer than the DFT-length a partitioned processing is necessary. Our contribution addresses problems concerning the partitioned calculation of the system misalignment and proposes an algorithm for obtaining the system misalignment by an optimal frequency dependent first order recursive smoothing criteria based on a minimum mean squared error (MMSE) approach.

1. INTRODUCTION

Acoustic echo cancelation is a common problem, e.g. in hands-free speaking devices or video conferencing systems [1]. Post-Filtering is an enhancement-technique for the conventional AEC-filter [2]. For implementing an acoustic echo canceller (AEC) with a Post-Filter it is necessary to have a reliable estimate of the (residual) echo signal and its PSD respectively at the output of the echo canceller.

Figure 1 shows a typical hands-free telephony or video-conferencing system with an AEC. Without any cancelation the signal of the far speaker (denoted by $\mathbf{X}(m,l)$) would be picked up by the microphone and transmitted back. The AEC-filter $\mathbf{C}(m,l)$ estimates the room transfer function $\mathbf{H}(m,l)$. The estimated echo $\hat{\Psi}(m,l)$ is subtracted from the microphone signal. Since in general the length of the room impulse response (RIR) is infinite or at least much longer than the length of the AEC filter [3, 4], a residual echo $\Xi(m,l) = \Psi(m,l) - \hat{\Psi}(m,l)$ remains in the microphone signal $E(m,l) = S_n(m,l) + \Xi(m,l)$. All signals and filter coefficients are used in a partitioned frequency domain manner with a discrete frequency index m and a discrete block index l.

The main drawback of a conventional AEC (a time domain NLMS-algorithm, e.g.) is slow convergence during adaptation periods such as filter initialization or sudden changes in the RIR. A Post-Filter $P(m,l)$ (designed according to the Wiener-rule) permits the application of shorter AEC-filters. The residual echo $\Xi(m,l)$ - which remains after the AEC - is estimated and attenuated.

Since for the Post-Filter as well as for the AEC it is necessary that the near speaker is absent during adaptation periods, a double-talk detector has to stop the adaptation in the case of an active near speaker. This can be done with the help of minimum statistics [5, 6] or the coherence between loudspeaker- and microphone-channel [7], e.g.

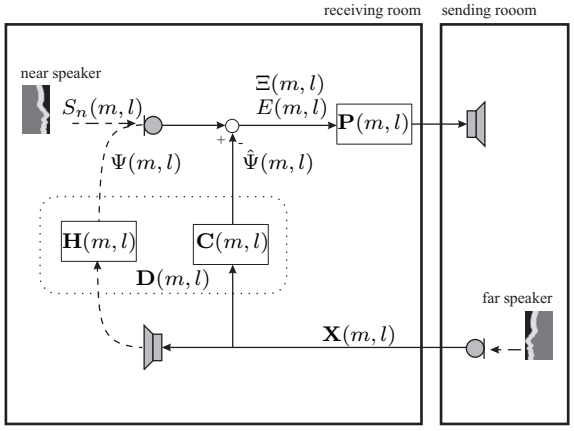

Figure 1: *Acoustic Echo Canceller (AEC) with Post-Filter.*

In section 2 we explain the problems arising from a partitioned calculation of the system misalignment. A rule for finding an optimal smoothing factor for the purpose of estimating the system misalignment is derived. In section 3 we present our simulation result and in section 4 we summarize our conclusions.

2. CALCULATION OF THE SYSTEM MISALIGNMENT

The room transfer function $\mathbf{H}(m,l)$, the AEC-filter $\mathbf{C}(m,l)$, the misalignment $\mathbf{D}(m,l)$, and the signal of the far speaker $\mathbf{X}(m,l)$ are defined as follows:

$$\mathbf{H}(m,l) = \left[H_0(m,l) \cdots H_{L'_H-1}(m,l)\right]^T \quad (1)$$

$$\mathbf{C}(m,l) = \left[C_0(m,l) \cdots C_{L'_{\mathrm{AEC}}-1}(m,l) \;\; 0 \cdots 0\right]^T \quad (2)$$

$$\mathbf{D}(m,l) = \mathbf{H}(m,l) - \mathbf{C}(m,l) \quad (3)$$

$$\mathbf{X}(m,l) = \left[X(m,l) \cdots X(m,l-L'_H+1)\right]^T \quad (4)$$

$L_H = L'_H L_{\mathrm{DFT}}$ and $L_{\mathrm{AEC}} = L'_{\mathrm{AEC}} L_{\mathrm{DFT}}$ are the lengths of the echo path impulse response and the AEC filter, respectively. Since in practical cases $L_H > L_{\mathrm{AEC}}$ only the first part of $\mathbf{H}(m,l)$ can be compensated by $\mathbf{C}(m,l)$ and the residual echo is

$$\Xi(m,l) = \mathbf{D}^T(m,l)\mathbf{X}(m,l). \quad (5)$$

For equation (5) and further on we assume an inactive near speaker ($S_n(m,l) = 0$). For this case the signal in the microphone path after the AEC $E(m,l)$ only contains the residual echo $\Xi(m,l)$.
For periods of an active near speaker ($S_n(m,l) \neq 0$) a double-talk detection algorithm has to stop the adaptation of the AEC and the Post-Filter.

2.1. Post-Filter Design

With an estimate for the system misalignment $\mathbf{D}(m,l)$, we obtain $\Xi(m,l)$ from (5). A reliable estimate of the residual echo PSD is essential for the design of the Post-Filter in order to avoid remaining echoes as well as desired speech distortions. The Wiener Post-Filter is given with

$$\begin{aligned} P(m,l) &= \frac{\hat{\Phi}_{S_nS_n}(m,l)}{\hat{\Phi}_{S_nS_n}(m,l) + \hat{\Phi}_{\Xi\Xi}(m,l)} \\ &= \frac{\hat{\Phi}_{EE}(m,l) - \hat{\Phi}_{\Xi\Xi}(m,l)}{\hat{\Phi}_{EE}(m,l)}. \end{aligned} \quad (6)$$

The PSD estimation in (6) can be calculated by the well-known Welch method [1]. Further difficulties appear when estimating $\mathbf{D}(m,l)$ at high system orders, e.g. if a video conferencing system is used in a reverberant environment, where the RIR might be very long. For this case a partitioned calculation of the system misalignment can be applied, which will be explained in the next section.

2.2. Partitioned Calculation of the System Misalignment

Figure 2 shows the corresponding system misalignment impulse response $d(k)$ in the time-domain. Although the length of the system misalignment is infinite in general, it can be assumed to be sufficiently decayed after a length of $L_H = 1024$. There are two possibilities for calculating the system misalignment. Either by using a long DFT length L_{DFT} or by investigating an increased number of partitions at short DFT lengths as illustrated in Figure 2. Note that an AEC operates at the first 512 samples. Partitioned calculation of $\mathbf{D}(m,l)$ has several advantages, e.g. a smaller delay in real-time applications or the reduction of strong loudspeaker-path correlations in multi-channel environments [2, 8].

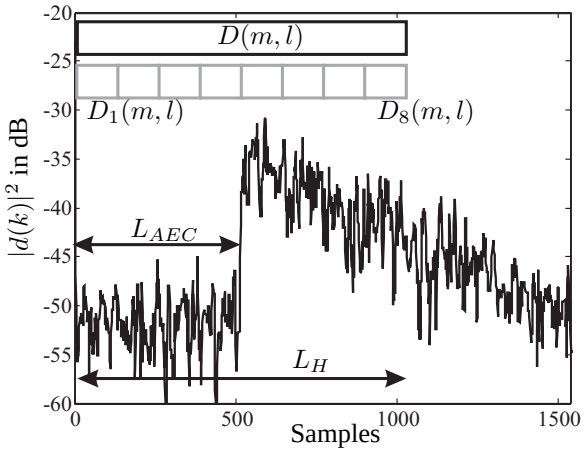

Figure 2: *System misalignment in time domain. Estimation of system misalignment can be processed in one part (black) or in blocks (gray).*

For obtaining $\mathbf{D}(m,l)$ we define an error $Q(m,l)$ which has to be minimized by an MMSE approach. For the reason of readability the frequency- and block-time-dependance is omitted for the next lines.

$$\begin{aligned} Q &= \Xi - \mathbf{X}^T\mathbf{D} \\ \mathrm{E}\{Q^*Q\} &= \mathrm{E}\{|\Xi|^2\} - \mathrm{E}\{\mathbf{D}^H\mathbf{X}^*\Xi\} \\ &\quad -\mathrm{E}\{\Xi^*\mathbf{X}^T\mathbf{D}\} + \mathrm{E}\{\mathbf{D}^H\mathbf{X}^*\mathbf{X}^T\mathbf{D}\} \\ \frac{\partial \mathrm{E}\{Q^*Q\}}{\partial \mathbf{D}} &= -2\mathrm{E}\{\mathbf{X}^*\Xi\} + 2\mathrm{E}\{\mathbf{X}^*\mathbf{X}^T\mathbf{D}\} \\ &\stackrel{!}{=} 0 \\ \mathbf{D} &= \mathrm{E}\{\mathbf{X}^*\mathbf{X}^T\}^{-1}\mathrm{E}\{\mathbf{X}^*\Xi\} \\ \mathbf{D}(m,l) &= \mathbf{R}_{XX}^{-1}(m,l)\mathbf{\Phi}_{X\Xi}(m,l). \end{aligned} \quad (7)$$

$\mathrm{E}\{\cdot\}$ is the expectation operator. $(\cdot)^*$ is the conjugate complex, $(\cdot)^T$ the transpose and $(\cdot)^H$ the hermitian (the conjugate transpose).
For the moment let us assume the loudspeaker signal $\mathbf{X}(m,l)$ to be uncorrelated in temporal direction:

$$\mathrm{E}\{X^*(m,l-i)X(m,l-k)\} = 0, \quad \forall\, i \neq k. \quad (8)$$

With (8) each partition of $\mathbf{D}(m,l)$ can be calculated sep-

arately.

$$D_i(m,l) = \frac{\Phi_{i,X\Xi}(m,l)}{\Phi_{XX}(m,l-i)} \quad (9)$$

$$\Phi_{i,X\Xi}(m,l) = \mathrm{E}\{X^*(m,l)\Xi(m,l)\} \quad (10)$$

We rewrite equation (5) as a sum of the vectors' elements. L'_D is the number of partitions of $\mathbf{D}(m,l)$.

$$\Xi(m,l) = \sum_{i=0}^{L'_D-1} D_i(m,l)\cdot X(m,l-i) \quad (11)$$

With (11) we can exemplaryly write for the first partition of $\mathbf{D}(m,l)$

$$\frac{\Phi_{0,X\Xi}(m,l)}{\Phi_{XX}(m,l)} = \frac{\mathrm{E}\left\{X^*(m,l)\sum_{i=0}^{L'_D-1} D_i(m,l)X(m,l-i)\right\}}{\mathrm{E}\{X^*(m,l)X(m,l)\}}$$
$$= D_0(m,l) + \underbrace{\frac{\sum_{i=1}^{L'_D-1} D_i(m,l)\mathrm{E}\{X^*(m,l)X(m,l-i)\}}{\mathrm{E}\{X^*(m,l)X(m,l)\}}}_{=0}. \quad (12)$$

$D_0(m,l)$ will be calculated correctly under the assumption of uncorrelated partitions of the loudspeaker signal $\mathbf{X}(m,l)$ (8).
For practical realization the expectation operators $\mathrm{E}\{\cdot\}$ have to be replaced by an estimation method $\hat{\mathrm{E}}\{\cdot\}$ which causes an additive disturbance.

$$\frac{\sum_{i=1}^{L'_D-1} D_i(m,l)\hat{\mathrm{E}}\{X^*(m,l)X(m,l-i)\}}{\hat{\mathrm{E}}\{X^*(m,l)X(m,l)\}} \neq 0 \quad (13)$$

To reduce the variance of the system misalignment estimation a first order smoothing can be applied.

$$\left|\hat{D}_0(m,l)\right|^2 = \alpha\left|\hat{D}_0(m,l-1)\right|^2 + (1-\alpha)\left|\frac{\hat{\Phi}_{0,X\Xi}(m,l)}{\hat{\Phi}_{XX}(m,l)}\right|^2 \quad (14)$$

Now we try to find an optimal smoothing factor α_{opt} for a stochastical system $\tilde{D}_0(m,l)$. With an optimal α the expectation $\mathrm{E}\{|\tilde{D}_0(m,l)|^2\}$ should converge to the real system misalignment $|D_0(m,l)|^2$.
The stochastical system misalignment is assumed to be zero-mean with mutual uncorrelated partitions

$$\mathrm{E}\left\{\tilde{D}_i^*(m,l)\tilde{D}_k^*(m,l)\right\} = 0, \quad \forall\, i \neq k. \quad (15)$$

Following [5] we define the MMSE minimization criterion as

$$\left(\left|\hat{D}_0(m,l)\right|^2 - \mathrm{E}\left\{\left|\tilde{D}_0(m,l)\right|^2\right\}\right)^2 \stackrel{!}{=} \min_{\alpha}. \quad (16)$$

Putting (14), (12) and (15) in (16) we get

$$\left(\left|\hat{D}_0(m,l)\right|^2 - \mathrm{E}\left\{\left|\tilde{D}_0(m,l)\right|^2\right\}\right)^2$$
$$= \left[\alpha\left|\hat{D}_0(m,l-1)\right|^2 + (1-\alpha)\left|\mathrm{E}\left\{\tilde{D}_0(m,l)\right\}\right.\right.$$
$$\left.+\frac{\sum_{i=1}^{L'_D-1}\mathrm{E}\left\{\tilde{D}_i(m,l)\right\}\hat{\mathrm{E}}\{X^*(m,l)X(m,l-i)\}}{\hat{\mathrm{E}}\{X^*(m,l)X(m,l)\}}\right|^2$$
$$\left.-\mathrm{E}\left\{\left|\tilde{D}_0(m,l)\right|^2\right\}\right]^2$$
$$= \left[\alpha\left(\left|\hat{D}_0(m,l-1)\right|^2 - \mathrm{E}\left\{\left|\tilde{D}_0(m,l)\right|^2\right\}\right.\right.$$
$$\left.-\sum_{i=1}^{L'_D-1}\mathrm{E}\left\{\left|\tilde{D}_i(m,l)\right|^2\right\}\frac{\hat{\Phi}_{XX}(m,l-i)}{\hat{\Phi}_{XX}(m,l)}\right)$$
$$\left.+\sum_{i=1}^{L'_D-1}\mathrm{E}\left\{\left|\tilde{D}_i(m,l)\right|^2\right\}\frac{\hat{\Phi}_{XX}(m,l-i)}{\hat{\Phi}_{XX}(m,l)}\right]^2. \quad (17)$$

Deriving (17) with respect to α and setting the result equal to zero we get the optimal smoothing factor

$$\alpha_{opt}(m,l) = \frac{1}{1+\frac{\left|\mathrm{E}\left\{\left|\tilde{D}_0(m,l)\right|^2\right\} - \left|\hat{D}_0(m,l-1)\right|^2\right|\hat{\Phi}_{XX}(m,l)}{\sum_{i=1}^{L'_D-1}\mathrm{E}\left\{\left|\tilde{D}_i(m,l)\right|^2\right\}\hat{\Phi}_{XX}(m,l-i)}}. \quad (18)$$

The factor $\left|\mathrm{E}\left\{\left|\tilde{D}_0(m,l)\right|^2\right\} - \left|\hat{D}_0(m,l-1)\right|^2\right|$ can be interpreted as a permitted adaptation speed, because if it equals zero α becomes one and the adaptation will be stopped.
Since $D_0(m,l)$ is not available in practical environments we define a first suboptimal $\tilde{\alpha}$

$$\tilde{\alpha}(m,l) = \frac{1}{1+\frac{\tilde{C}\cdot\hat{\Phi}_{XX}(m,l)}{\sum_{i=1}^{L'_D-1}\left|\hat{D}_i(m,l-1)\right|^2\hat{\Phi}_{XX}(m,l-i)}} \quad (19)$$

with a constant $\tilde{C}$ for $\left|\mathrm{E}\{|\tilde{D}_0(m,l)|^2\} - |\hat{D}_0(m,l-1)|^2\right|$ and the system misalignment of the recent block $|\hat{D}_i(m,l-1)|^2$ for $\mathrm{E}\{|\tilde{D}_i(m,l)|^2\}$. As simulations have shown, typical values for $\tilde{C}$ are in the range of 0.08 to 0.12 and thus we choose $\tilde{C} = 0.1$. As a further simplification every dependence of the system misalignment can be neglected in (18) and we define

$$\alpha'(m,l) = \frac{1}{1+C'\frac{\hat{\Phi}_{XX}(m,l)}{\sum_{i=1}^{L'_D-1}\hat{\Phi}_{XX}(m,l-i)}}. \quad (20)$$

As a reference we take a fixed non frequency dependent smoothing factor α_{fixed} which we obtain in the from

$$\alpha_{fixed} = e^{-\frac{F_B}{\tau\cdot f_s}} \quad (21)$$

with the block feed F_B, the forgetting time τ and the sampling frequency f_s.

3. SIMULATION RESULTS

In Figure 3 we see the estimation of the residual echo PSD which is necessary for the Post-Filter design in (6), e.g. The solid line is the true PSD.

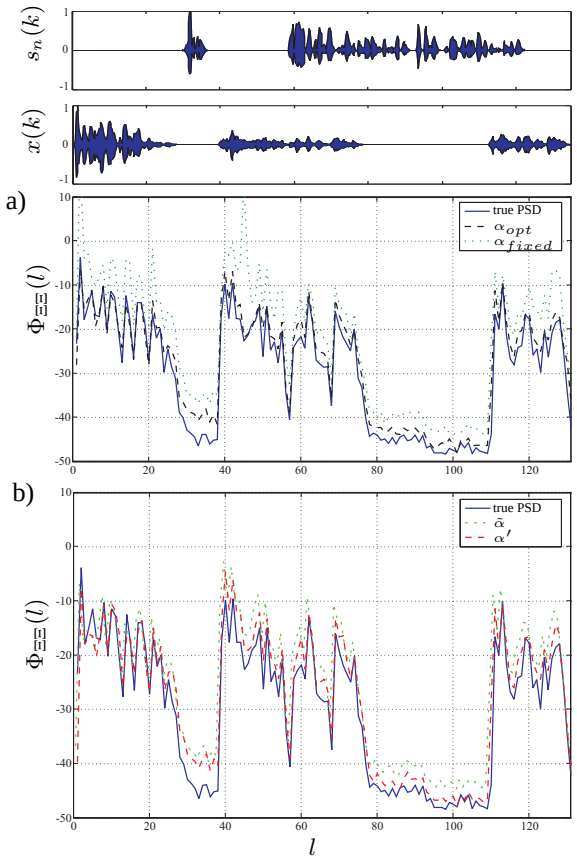

Figure 3: *Residual echo PSDs with different methods for smoothing the system misalignment* $\mathbf{D}(m,l)$

In the upper part we see the loudspeaker signal $x(k)$ containing the far speaker and the near speaker $s_n(k)$ in the time domain. During periods of double-talk the adaptation of the filters was stopped [6].

As Figure 3a) shows, we meet the real PSD best when using the optimal smoothing factor $\alpha_{opt}(m,l)$ (18). Using α_{fixed} from equation (21) for the calculation of the system misalignment produces an over-estimation of the residual echo PSD.

Figure 3b) shows the results for the suboptimal smoothing factors $\alpha'(m,l)$ (dashed line, eq. (20)) and $\tilde{\alpha}(m,l)$ (dotted line, eq. (19)). The residual-echo-PSD is met for both of them. $\alpha'(m,l)$ shows somewhat better results than $\tilde{\alpha}(m,l)$. The reason for this is in our opinion the substitution of the real system misalignment by its estimation from the last block. With (20) we present a new smoothing method, which is robust and straight forward to implement.

4. CONCLUSIONS

We presented an optimal frequency dependent first order recursive smoothing factor $\alpha(m,l)$ for the calculation of the system misalignment $\mathbf{D}(m,l)$, designed according to an MMSE approach. The system misalignment obtained from the theoretically optimal smoothing factor led to an exact estimate of the residual echo PSD. We derived two approximations of the smoothing factor for the available signals in a practical environment. One of the solutions led to significantly improved results compared to the case of using a fixed smoothing factor.

5. REFERENCES

[1] G. Enzner, R. Martin, and P. Vary, "Unbiased Residual Echo Estimation for Hands-Free Telephony," *Intern. Conf. on Acoustics Speech and Signal Processing (ICASSP'02)*, 2002.

[2] M. Kallinger, J. Bitzer, and K.D. Kammeyer, "Post-Filtering for Stereo Acoustic Echo Cancellation," *Int. Workshop on Acoustic Echo and Noise Control (IWAENC-2003)*, 2003.

[3] D.W.E. Schobben and P.C.W. Sommen, "On the Performance of too Short Adaptive FIR Filters," in *Proc. of the ProRISC Workshop on Circuits, Systems and Signal Processing*, 1997, pp. 473–476.

[4] R.D. Poltmann, "Stochastic Gradient Algorithm for System Identification Using Adaptive FIR-Filters with too Low Number of Coefficients," *IEEE Trans. on Circuits and Systems*, vol. 35, no. 2, pp. 247–250, Feb 1988.

[5] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[6] Markus Kallinger, Karl-Dirk Kammeyer, and Jörg Bitzer, "Multi-Microphone Residual Echo Estimation," *Intern. Conf. on Acoustics Speech and Signal Processing (ICASSP'03), Hong Kong, China*, 2003.

[7] J. Benesty, D.R. Morgan, and J.H. Cho, "A New Class of Doubletalk Detectors Based on Cross-Correlation," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 2, pp. 168–171, March 2002.

[8] M. Mohan Sondhi and D.R. Morgan, "Stereophonic Acoustic Echo Cancellation - An Overview of the Fundamental Problem," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 148–151, August 1995.