TWO STAGE ESTIMATION FOR THE ECHO PATHS IN STEREOPHONIC ACOUSTIC ECHO CANCELLATION

[1]*Christof Faller,* [2]*Tomas Gänsler, and* [1]*Martin Vetterli*

[1]christof.faller@epfl.ch
[1] Audiovisual Communications Laboratory, EPFL Lausanne, Switzerland
[2]mh acoustics LLC, Summit, NJ, USA

ABSTRACT

A novel algorithm is introduced for estimating the echo path responses for stereophonic acoustic echo cancellation. When the Wiener solution for the echo path responses is non-unique, the echo paths are estimated in two stages. In a first stage, one of the non-unique Wiener solutions is estimated. Then in a second stage, a disturbance is introduced and the left and right misalignment vectors can be formulated as the Wiener solution with respect to the residual signals appearing after adding the disturbance. The aim is to take advantage of the knowledge about the specific disturbance which is introduced. Adaptive algorithms based on this two stage idea are used. It is shown with a series of numerical simulations that the proposed algorithm converges much faster than a corresponding conventional one stage adaptive algorithm.

1. INTRODUCTION

The *acoustic echo canceler* (AEC) [1] is an essential part of full-duplex tele-communication systems. Echoes arise from the coupling between a loudspeaker and a microphone. The AEC removes the echo components present in the microphone signal. A *stereophonic acoustic echo canceler* (SAEC) [2] is shown in Fig. 1. The left and right loudspeaker signals, x_1 and x_2, propagate through the echo paths to the microphone. In the static case, it is assumed that the acoustic echo paths from the left and right loudspeakers to the microphone are accurately modeled by the linear filters h_1 and h_2 with finite impulse responses. The microphone signal y is composed of the left and right echoes, the near-end talker signal v, and ambient noise w,

$$y(n) = \mathbf{h}_1^T \mathbf{x}_1(n) + \mathbf{h}_2^T \mathbf{x}_2(n) + v(n) + w(n)\,, \qquad (1)$$

where

$$\begin{aligned} \mathbf{x}_1(n) &= [x_1(n)\, x_1(n-1) \ldots x_1(n-M+1)]^T \\ \mathbf{x}_2(n) &= [x_2(n)\, x_2(n-1) \ldots x_2(n-M+1)]^T \\ \mathbf{h}_1 &= [h_{1,0}\, h_{1,1} \ldots h_{1,M-1}]^T \\ \mathbf{h}_2 &= [h_{2,0}\, h_{2,1} \ldots h_{2,M-1}]^T\,, \qquad (2) \end{aligned}$$

and M is the length of the echo path responses. The SAEC uses adaptive filters, $\hat{h}_1$ and $\hat{h}_2$, to estimate the echo path responses, h_1 and h_2. The error signal is defined as

$$e(n) = y(n) - \hat{\mathbf{h}}_1^T \mathbf{x}_1(n) - \hat{\mathbf{h}}_2^T \mathbf{x}_2(n)\,, \qquad (3)$$

where

$$\begin{aligned} \hat{\mathbf{h}}_1 &= [\hat{h}_{1,0}\, \hat{h}_{1,1} \ldots \hat{h}_{1,M-1}]^T \\ \hat{\mathbf{h}}_2 &= [\hat{h}_{2,0}\, \hat{h}_{2,1} \ldots \hat{h}_{2,M-1}]^T \qquad (4) \end{aligned}$$

are the adaptive filter coefficient vectors. Note that in this paper we are assuming that the lengths of the echo path impulse responses and adaptive filters are the same, i.e. we are assuming that the adaptive filters are large enough such that the modeling error due to the "tail effect" is negligible. The top of Fig. 1 and (1) is denoted *acoustic system* in this paper.

Minimization of $\mathrm{E}\{e^2(n)\}$ with respect to the modeling filters leads to the normal equation [3],

$$\mathbf{R}\hat{\mathbf{h}} = \mathbf{r}\,, \qquad (5)$$

where $\hat{\mathbf{h}} = [\hat{\mathbf{h}}_1\ \hat{\mathbf{h}}_2]^T$, $\mathbf{R}$ is the covariance matrix of two concatenated processes x_1 and x_2, and and $\mathbf{r}$ is the cross-correlation vector between the loudspeaker signals and the microphone signal.

During tele-conferencing most times only one talker is active. The left and right stereo signals with one active

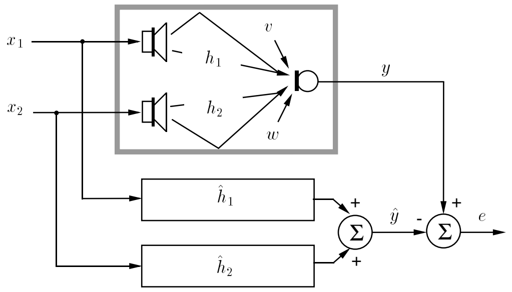

Figure 1: *Stereo acoustic echo canceler (SAEC).*

virtual source (talker) are highly correlated. This results in that the covariance matrix $\mathbf{R}$ is not full rank or ill-conditioned. If $\mathbf{R}$ is not full rank, the Wiener solution is non-unique and the SAEC does in general not converge to the echo path but one of the other Wiener solutions.

To prevent that the Wiener solution of the system shown in Fig. 1 is non-unique or highly ill-conditioned, the loudspeaker signals are pre-processed in order to reduce their coherence. One of the most successful approaches for reducing the coherence is to add to the loudspeaker signals, x_1 and x_2, non-linear disturbances [4],

$$\begin{aligned} \Delta x_1(n) &= \alpha \frac{x_1(n) + |x_1(n)|}{2} \\ \Delta x_2(n) &= \alpha \frac{x_2(n) - |x_2(n)|}{2} \,. \end{aligned} \tag{6}$$

For speech signals, the constant α can be as large as 0.3 before the distortions become annoying. Furthermore, this type of disturbance does hardly alter (e.g. widen) the stereo image.

Since the disturbance that is added to the loudspeaker signals has to be rather weak, i.e. it has to be hardly perceptible, the loudspeaker signals are only partially decorrelated after pre-processing. Thus, the associated covariance matrix is often still ill-conditioned. Therefore, for SAEC a simple algorithm such as the *normalized least mean squares* (NLMS) [3] will not or only very slowly converge. This paper is about accelerating adaptive filter algorithms by explicitly taking into consideration the knowledge about the disturbances which are introduced into the loudspeaker signals.

2. THE PROPOSED ALGORITHM

The acoustic echo path responses are estimated in two stages. In a first stage, a (possibly non-unique) Wiener solution is estimated for the echo path responses. In a second stage, disturbances (e.g. (6)) are added to the loudspeaker signals and a new Wiener problem is formulated for the left and right misalignment vectors, ϵ_1 and ϵ_2.

During the first stage, the acoustic system is driven by the unmodified loudspeaker signals and a possibly non-unique Wiener solution is estimated

$$\begin{aligned} \hat{\mathbf{h}}_1 &= \mathbf{h}_1 - \boldsymbol{\epsilon}_1 \\ \hat{\mathbf{h}}_2 &= \mathbf{h}_2 - \boldsymbol{\epsilon}_2 \,, \end{aligned} \tag{7}$$

where $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are the misalignment coefficient vectors

$$\begin{aligned} \boldsymbol{\epsilon}_1 &= [\epsilon_{1,0}\, \epsilon_{1,1} \ldots \epsilon_{1,M-1}]^T \\ \boldsymbol{\epsilon}_2 &= [\epsilon_{2,0}\, \epsilon_{2,1} \ldots \epsilon_{2,M-1}]^T \,. \end{aligned} \tag{8}$$

Assuming that x_1 and x_2 have non-singular covariance, $v = 0$, and that w is orthogonal to x_1 and x_2, the condition for the set of Wiener solutions is that the sum of the

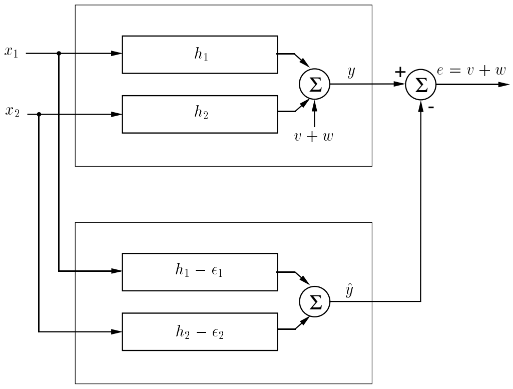

Figure 2: *The acoustic system and the non-unique Wiener model.*

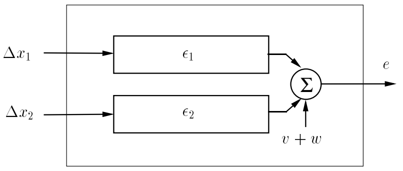

Figure 3: *The residual system.*

output of the two filters, $\hat{\mathbf{h}}_1$ and $\hat{\mathbf{h}}_2$, is equal to the sum of the outputs of the two echo path responses, $\mathbf{h}_1$ and $\mathbf{h}_2$. Therefore,

$$\boldsymbol{\epsilon}_1^T \mathbf{x}_1(n) + \boldsymbol{\epsilon}_2^T \mathbf{x}_2(n) = 0 \tag{9}$$

must hold. The estimation error of a Wiener solution is equal to $e = v + w$ as indicated in Fig. 2.

Given the acoustic system and its (possibly non-unique) Wiener estimate from the first stage, for the second stage, disturbances Δx_1 and Δx_2 are introduced into the loudspeaker signals. The combined system comprising the acoustic system and its Wiener estimate can be simplified and is equivalent to the *residual system*, shown in Fig. 3, which computes the error signal e as a function of the disturbances,

$$e(n) = \boldsymbol{\epsilon}_1^T \Delta\mathbf{x}_1(n) + \boldsymbol{\epsilon}_2^T \Delta\mathbf{x}_2(n) + v(n) + w(n) \,. \tag{10}$$

This residual system is similar in structure to the original acoustic system of the stereophonic echo cancellation problem (1). However its SNR (SNR of $e(n)$) is smaller than the SNR of the acoustic system since the disturbances need to be virtually imperceptible and thus are weaker than the unmodified loudspeaker signals.

Adaptive filters, $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$, are used to estimate the misalignments of the left and right modeling filters. The error signal is defined as

$$e'(n) = e(n) - \hat{\boldsymbol{\epsilon}}_1^T \Delta\mathbf{x}_1(n) - \hat{\boldsymbol{\epsilon}}_2^T \Delta\mathbf{x}_2(n) \,, \tag{11}$$

where the disturbance signal and coefficient vectors are

$$\begin{aligned}\Delta\mathbf{x}_1(n) &= [\Delta x_1(n)\,\Delta x_1(n-1)\ldots\Delta x_1(n-M+1)]^T\\ \Delta\mathbf{x}_2(n) &= [\Delta x_2(n)\,\Delta x_2(n-1)\ldots\Delta x_2(n-M+1)]^T\\ \hat{\boldsymbol{\epsilon}}_1 &= [\hat{\epsilon}_{1,0}\,\hat{\epsilon}_{1,1}\ldots\hat{\epsilon}_{1,M-1}]^T\\ \hat{\boldsymbol{\epsilon}}_2 &= [\hat{\epsilon}_{2,0}\,\hat{\epsilon}_{2,1}\ldots\hat{\epsilon}_{2,M-1}]^T\,. \qquad (12)\end{aligned}$$

The condition of the covariance matrix corresponding to (10) is determined by the properties of the disturbances and thus can be controlled by the choice of the disturbances. The SAEC output signal for the second stage is $e'(n)$, i.e. the sum of the estimates of the filters of stage 1 and 2 are used as the estimated echo path impulse response.

3. SIMULATIONS

In this section, we analyze the convergence properties of the proposed two stage adaptive algorithm. For all simulations in this section the loudspeaker signals without disturbances are the same white Gaussian noise signals, such that $x_1 = x_2$. The near-end talker signal v is not present ($v = 0$) and white Gaussian noise is used for w for an SNR of 30 dB. Two measured room impulse responses, truncated to $M = 100$ taps, are used for h_1 and h_2. The l_2 norms of $\mathbf{h}_1$ and $\mathbf{h}_2$ are approximately the same ($\|\mathbf{h}_1\| = 0.098$, $\|\mathbf{h}_2\| = 0.094$). For stage 1 and 2 NLMS algorithms with a step size of 0.05 are used.

Only short echo path impulse responses are used so that we can show the effectiveness in terms of a simple NLMS algorithm. If much longer echo path impulse responses are used, then there is a need for more powerful adaptive algorithms for all cases compared here.

3.1. Convergence of stage 2

In practice, stage 1 estimates $\hat{h}_1$ and $\hat{h}_2$ only to be a Wiener solution with limited accuracy. The accuracy of these estimations influences the convergence of stage 2.

To assess the effect of limited accuracy of the non-unique Wiener solution estimate on convergence of stage 2, we ran a number of simulations using the proposed adaptive algorithm. Independent white Gaussian noise, 20 dB below the unmodified loudspeaker signal level, was used as disturbances. Figure 4 shows the normalized misalignment for stage 2. Non-unique Wiener solution estimates with different precision were used. The normalized misalignments (errors) of the non-unique Wiener estimates are indicated in the graphs in dB. White Gaussian noise was used as misalignment vectors to simulate non-precise non-unique Wiener estimates. As expected, the performance of stage 2 decreases as the non-unique Wiener estimates become less precise. However, stage 2 seems to be quite robust and performs still fairly well for stage 1 errors as high as -10 dB.

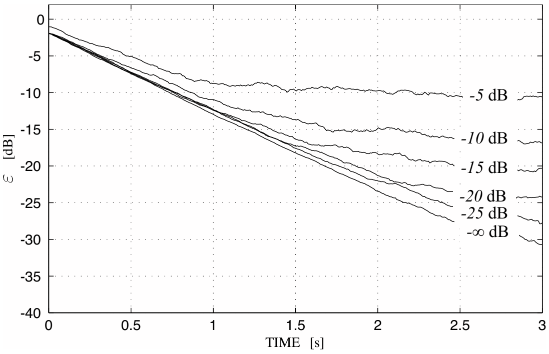

Figure 4: *The performance of stage 2 for different precisions of the non-unique Wiener solution estimate of stage 1 (The misalignment between the non-unique Wiener solution estimate before operating stage 2 is indicated in the graphs).*

3.2. Comparison of NLMS and two stage NLMS

Figure 5 shows various graphs for comparing the performance of the NLMS algorithm to NLMS-based two stage algorithms. The graphs show the normalized misalignment (solid) and the *normalized mean square error* (NMSE) (dotted). The normalized misalignment indicates how close the adaptive filters are to the echo path responses. The echo cancellation performance is better the lower the NMSE is. A non-linear disturbance, (6) with $\alpha = 0.2$, is used.

The results for a conventional NLMS-based SAEC are shown in the top graph of Fig. 5. It converges very slowly as implied by the normalized misalignment which is about -2.5 dB after 4 s. However, note that the echo is effectively suppressed as indicated by the NMSE. The slow convergence can be explained by the weakness of the disturbance and the associated ill-conditioned covariance matrix.

The NLMS-based two stage algorithm, switching from stage 1 to 2 at 1 s is shown in the middle graph. The graph indicates that it converges much faster than the standard NLMS algorithm. Note that the disturbance is only introduced for stage 2 which explains the slightly increasing NMSE at 1 s.

In order to prevent the necessity of explicit switching between stage 1 and 2, we run a simulation where both stages run simultaneously and the disturbance is always introduced. Every 100 ms the used estimate is either updated with stage 1 or stage 2, depending on which results in better average NMSE. The result is shown in the bottom graph of the figure. This algorithm with “continuous switching” converges slightly faster than the two stage algorithm with explicit switching. This is so because the non-unique Wiener solution estimate with continuous

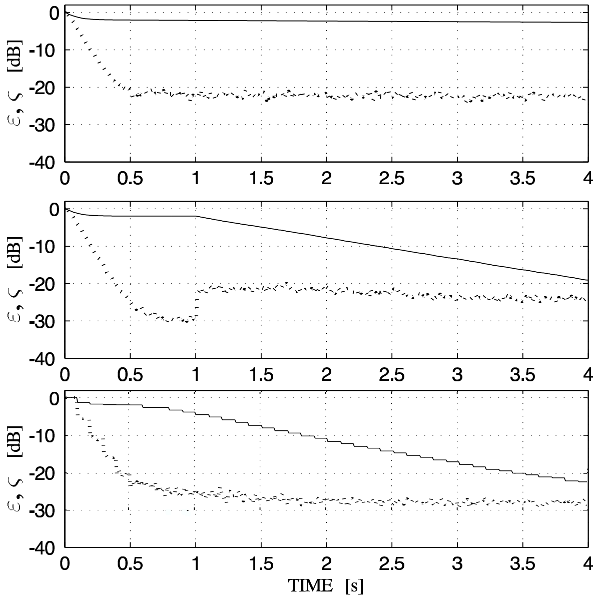

Figure 5: *Misalignment (solid) and NMSE (dotted): NLMS (top); NLMS-based two stage algorithm switching between stage 1 and 2 at 1 s (middle); NLMS-based two stage algorithm with continuous switching.*

switching is periodically refined.

3.3. Simulations with a stereo rendering change

Simulations with a stereo rendering change were carried out. The initial stereo signal $x_2(n) = x_1(n)$ is changed to $x_2(n) = 3x_1(n - D)$ after 3 s, where D is chosen such that it corresponds to 1.5 ms. The same non-linear disturbances as previously were used.

The top graph of Fig. 6 shows a simulation using the NLMS algorithm. The NMSE before 3 s indicates that the echo is effectively suppressed before the stereo rendering change. However, the increase to about -2.5 dB of the NMSE after 3 s indicates that the echo is not anymore effectively suppressed after the stereo rendering change until the adaptive filter re-converges. The misalignment indicates that the echo path responses are not accurately estimated which also explains the increase in NMSE after the stereo rendering change.

The bottom graph of Fig. 6 shows the same simulation as previously but using an NLMS-based two stage algorithm with continuous switching. Since the misalignment at the time of the stereo rendering change is already rather small the increase in NMSE is much less severe than for the NLMS algorithm, indicating that the echo is still canceled (about -14 dB) after the stereo rendering change.

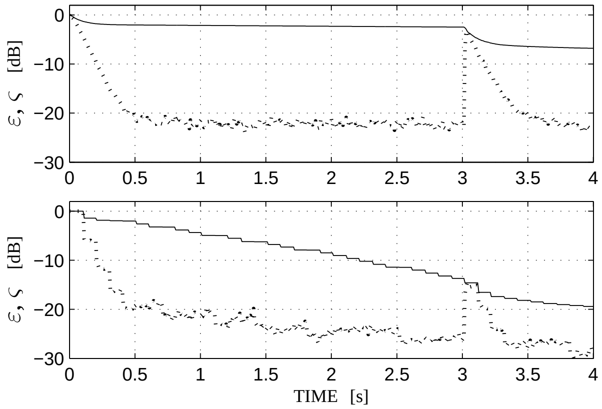

Figure 6: *Misalignment (solid) and NMSE (dotted) for a simulation with a stereo rendering change at* 3 *s: NLMS (top); NLMS-based two stage algorithm with continuous switching (bottom).*

4. CONCLUSIONS

We proposed a novel algorithm for addressing the problem of non-uniqueness in stereophonic acoustic echo cancellation. While previous approaches introduce disturbances into the loudspeaker signals for reducing the coherence, they do not take advantage of the fact that the disturbances are known. The algorithm proposed here explicitly takes into consideration the disturbances and thus has the potential for faster convergence.

The proposed algorithm operates in two stages. In a first stage a Wiener solution is estimated. A second stage estimates the misalignment vectors between the previously estimated Wiener solution and the true echo path responses.

In order to avoid explicit switching between the two stages continuous switching was investigated, i.e. operating both stages simultaneously and selecting periodically the stage yielding better performance.

5. REFERENCES

[1] M. M. Sondhi, "An adaptive echo canceler," *Bell Syst. Tech. J.*, vol. 46, pp. 497–510, Mar. 1967.

[2] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation - an overview of the fundamental problem," *IEEE Signal Processing Lett.*, vol. 2, pp. 148–151, Aug. 1995.

[3] S. Haykin, *Adaptive Filter Theory (third edition)*, Prentice Hall, 1996.

[4] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 156–165, Mar. 1998.