Proc. 3rd IEEE Benelux Signal Processing Symposium (SPS-2002), Leuven, Belgium, March 21–22, 2002

RESIDUAL ECHO ESTIMATION WITH THE HELP OF MINIMUM STATISTICS

Markus Kallinger, Karl-Dirk Kammeyer

University of Bremen, FB 1
Dept. of Communications Engineering
P.O. Box 330 440
D–28334 Bremen, Germany
kallinger@ant.uni-bremen.de

Jörg Bitzer

Houpert Digital Audio
Anne-Conway-Str. 1
D–28359 Bremen, Germany
j.bitzer@hda.de

ABSTRACT

This contribution deals with a new technique to estimate the residual echo at the output of an acoustic echo canceller (AEC). It is known that the compensation filter of the AEC has to be very long in reverberant environments in order to provide sufficient echo attenuation. However, high filter orders involve long convergence periods during initialisation or after modifications of the echo path impulse response. In these periods poor echo attenuation leads to undesired artifacts. One possible solution to this problem is to use a low-order and quickly converging AEC in combination with a Wiener post-filter after the AEC. Therefore, the power spectral density (PSD) of the residual echo has to be estimated properly. However, the problem is to achieve reliable estimates in periods of double-talk or during modifications of the echo path impulse response. This paper suggests a new method for estimating the residual echo with the help of minimum statistics to suppress interferences by near-end speech during double-talk situations.

1. INTRODUCTION

The ideal solution to suppress acoustic echoes is the AEC. Depending on the acoustic environment, the adaptive filter must be chosen to be very long. This results into slow convergence of the AEC [1]. In order to support the AEC during initial convergence and after changes of the echo path impulse response, an adaptive post-filter can be used [2]. With a speech signal $S(n,s)$, the residual echo $B(n,s)$ and a noise signal $N(n,s)$ the Wiener design rule leads to an optimal filter with the transfer function

$$P(n,s) = \frac{\hat{\Phi}_{SS}(n,s)}{\hat{\Phi}_{SS}(n,s) + \hat{\Phi}_{NN}(n,s) + \hat{\Phi}_{BB}(n,s)}, \quad (1)$$

where s is the frame index and n denotes the discrete frequency index. Each $\hat{\Phi}(n,s)$ is an estimate of the actual block by block power spectral density $\Phi(n,s)$. The residual echo signal $B(n,s) = D(n,s) - \hat{D}(n,s)$ contributes to the mixed signal $E(n,s) = S(n,s) + N(n,s) + B(n,s)$ after the AEC. All involved signals and systems are illustrated in figure 1. The design rule in equation (1) was deduced under the assumption of statistically independent signals. The isolated PSD $\Phi_{NN}(n,s)$ is usually estimated under the assumption of stationarity for the background noise. It can be estimated by the use of a voice activity detector (VAD) or with the help of the well-known *minimum statistics estimation method* [3].

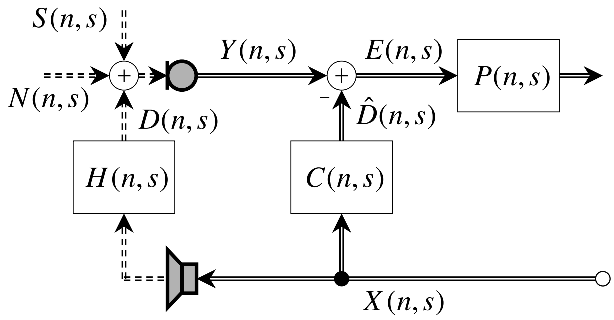

Figure 1: Frequency domain signal model of an acoustic echo canceller with a succeeding post-filter.

This paper addresses the problem of estimating the residual echo's PSD $\Phi_{BB}(n,s)$. We have to deal with a partly compensated speech signal, which is still non-stationary. However, $B(n,s)$ is related to the known far-end speaker's signal $X(n,s)$ via a linear convolution with the system misalignment transfer function $G(n,s) = H(n,s) - C(n,s)$. A recently proposed method [4] suggests the estimation of the magnitude squared coherence (MSC) between the signals $X(n,s)$ and $E(n,s)$ to gather the residual echo's PSD $\Phi_{BB}(n,s)$. Other approaches [2] employ the estimation of a virtual transfer function $F(n,s)$ to compute the residual echo by $B(n,s) \approx F(n,s)D(n,s)$. Our focus lies on the computation of the system misalignment transfer function $G(n,s)$, which will be robust against additive disturbances such as the near-end speech signal $S(n,s)$ or the background noise $N(n,s)$. Thus, we can get $B(n,s) = G(n,s)X(n,s)$.

We introduce the new method in section 2. First, we explain, how we incorporate the minimum statistics estimation method especially to suppress additive influences of the near-end speaker. Then, we illustrate, how a comparably long impulse response can be gathered, although we only employ short lengths for the fast Fourier transform (FFT) with short block sizes of signal line feed. Section 3 presents some simulation results. Apart from the single-channel Wiener filter we propose two other possible applications for the new estimation method in section 4. In section 5 we conclude the paper.

2. RESIDUAL ECHO ESTIMATION

To estimate the misalignment transfer function $G(n,s)$ we employ two separate procedures: The first is intended for the fast detection of disturbances by near-end speech signals. In a second step, we apply an accurate method to get the transfer function $\hat{G}(n,s)$, which fulfills the relation $E(n,s) = \hat{G}(n,s)X(n,s)$. A detailed description of this method is given in section 2.2. In section 2.1 we introduce a technique to find frequency indices n_u, at which we can suppose that $G(n_u,s) \approx \hat{G}(n_u,s)$. Due to short term correlations of the statistically independent signals $s(k)$ and $b(k)$, direct estimations of $G(n,s)$ would be strongly corrupted by near-end speech interferences.

2.1. Integration of minimum statistics

At first, we try to extract the squared magnitude of the echo path transfer function

$$|H(n,s)|^2 \approx |\hat{H}(n,s)|^2 = \frac{\Phi_{YY}(n,s)}{\Phi_{XX}(n,s)}. \qquad (2)$$

Let us assume that the current acoustic scenario is characterised by low background noise and near-end speech signal powers. For a slowly changing echo path, we expect $|\hat{H}(n,s)|^2$ to vary hardly. On the other hand, sudden rising peaks in this estimation result from uncorrelated additive disturbances, e.g. a near-end speech signal. Therefore, we apply the minimum statistics estimation method to $|\hat{H}(n,s)|^2$ to suppress the influence of near-end speech in corrupted subbands. Associated frequency indices n_u with a low level of this kind of interference fulfill the condition

$$\frac{|\hat{H}(n_u,s)|^2 - \text{MinStat}\{|\hat{H}(n_u,s)|^2\}}{\text{MinStat}\{|\hat{H}(n_u,s)|^2\}} < c_{Thr}. \qquad (3)$$

The *minimum statistics operator* denoted as $\text{MinStat}\{\cdot\}$ symbolises the search for the minimum in each subband in the time direction within a sliding time window. c_{Thr} is a frequency independent constant, which represents a threshold for relative distances between the squared magnitude of the echo path transfer function and the minimum statistics processed version of it. In our simulations we used a threshold of $c_{Thr} = 4.0$.

The computations of $|\hat{H}(n,s)|^2$ and $\hat{G}(n,s)$ run in parallel. Now that we can distinguish between subbands with near-end speech interferences n_k and undisturbed subbands n_u we can update the actual estimation of $G(n,s)$ at indices $n = n_u$. At all other indices we keep the former estimates in the way as illustrated in figure 2. As an alternative, we could apply $\hat{G}(n,s)$ instead of $\hat{H}(n,s)$ in condition (3). However, experiments with suddenly modifying echo path impulse responses have shown that the minimum statistics estimation method cannot distinguish near-end speech from impulse response changes in this case. Even short and fast adaptive AEC-filters cannot follow such modifications instantly. The residual echo's signal power quickly rises, which is the same effect interfering speech signals cause. When $\hat{H}(n,s)$ comes into operation, this problem does not occur. In our simulations an adaptive filter with 128 coefficients was updated by a simple NLMS-algorithm. A non-subband method incorporating minimum statistics has been suggested in [5].

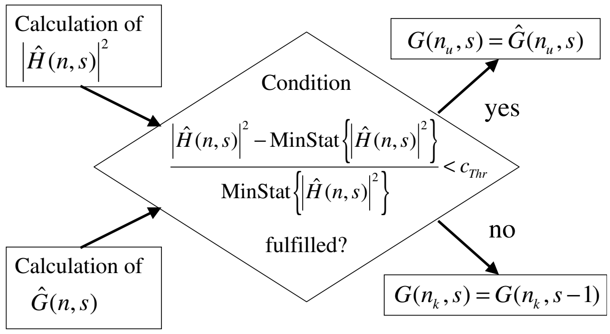

Figure 2: Flow chart for the application of the proposed update condition.

2.2. Transfer function estimation

As a first step, we only have to solve the equivalent to the Wiener-Hopf equation in the frequency domain

$$\hat{G}(n,s) = \frac{\Phi_{XE}(n,s)}{\Phi_{XX}(n,s)}. \qquad (4)$$

However, we must consider that by this we can only estimate the first M taps of the associated system misalignment impulse response $\hat{g}(\kappa)$. M is the applied FFT-resolution. In addition, only M samples of the signals $x(k)$ and $e(k)$ are used for estimating the cross spectral density $\Phi_{XE}(n,s)$ (CSD) and the PSD $\Phi_{XX}(n,s)$. The estimates will be biased too strongly. Therefore, we apply a first-order IIR filter

$$\hat{\Phi}_{XE}(n,s) = (1-\alpha)X^*(n,s)E(n,s) + \alpha\hat{\Phi}_{XE}(n,s-1). \quad (5)$$

to smooth the estimates. However, the smoothing constant α must be set to a comparably small value, since we must not smear the influence of additive disturbances to upcoming estimates. We will use former signal blocks to get more accurate spectral density estimates using Rader's method [6]. Since this technique is based on summing up small partial correlations, we will briefly illustrate the calculation of one partial correlation in the following.

$X_{2M}(n,s)$ is the $2M$-point Fourier transformed version of the signal $x_s(k)$ with 50% zero-padding, i.e.

$$x_s(k) = \begin{cases} x(k+s\cdot M) & \text{for } 0 \le k \le M-1 \text{ and} \\ 0 & \text{for } M \le k \le 2M-1. \end{cases} \quad (6)$$

In order to get unbiased estimates of the block by block time-variant cross correlation $r_{XE}(\kappa,s) = \mathrm{E}\{x(k+sM)\, e(k+sM+\kappa)\}$ we must form a signal $e_s(k)$ by

$$e_s(k) = e(k+s\cdot M) \quad \text{for } 0 \le k \le 2M-1. \quad (7)$$

$\mathrm{E}\{\cdot\}$ represents the expectation operator. A concatenation to gather $e_s(k)$ can be done efficiently in the frequency domain. Let us take two $2M$-point Fourier transformed versions $E_{2M}(n,s)$ and $E_{2M}(n,s+1)$, which were generated just as $X_{2M}(n,s)$ with 50% zero-padding. The transformed version of $e_s(k)$ can be calculated by

$$\mathrm{DFT}\{e_s(k)\} = E_{2M}(n,s) + e^{\frac{-j2\pi Mn}{2M}} E_{2M}(n,s+1) \quad (8)$$

DFT denotes the discrete Fourier transform. If we want to consider a whole of L samples for the time-variant correlation estimate $\hat{r}_{XE}(\kappa,s)$, we get

$$\hat{r}_{XE}(\kappa,s) = \quad (9)$$
$$\frac{1}{L}\mathrm{DFT}^{-1}\left\{ \begin{array}{l} \displaystyle\sum_{i=s-L/M}^{s-1} X^*_{2M}(n,i)[E_{2M}(n,i) \\ \qquad\qquad + (-1)^n E_{2M}(n,i+1)] \end{array} \right\}$$
$$\text{for} \quad 0 \le \kappa \le M.$$

The estimate $\hat{r}_{XX}(\kappa,s)$ of the auto correlation can be calculated in the same way. Values of $\hat{r}_{XE}(\kappa)$ with $M < \kappa$ contain parts of an undesired cyclic convolution. If we now transform the correlation estimates back to the frequency domain, we can calculate the transfer function $\hat{G}(n,s)$. But the associated impulse response $\hat{g}(\kappa)$ will only be valid for a limited range of κ. For larger values of κ, e.g. $M+1 \le \kappa \le 2M$ we must insert preceding frames of the reference signal $x(k)$. In this example, we would have to use $X^*_{2M}(n,s-2)$ instead of $X^*_{2M}(n,s-1)$ in equation (9). The segments $E_{2M}(n,s-1)$ and $E_{2M}(n,s)$ are maintained.

During the simulations, we used a block size of $M = 128$ and a span for calculating the correlations of $L = 512$. κ covered an overall range of $0 \le \kappa \le 512$. The advantage of this method is the fact that we obtain updated transfer functions at each 128 samples. Each update only demands FFT-operations at a resolution of 256. We can use the partitioned sections of $g(\kappa)$ later on to calculate $B(n,s)$ efficiently with low processing delay [7]. $|\hat{H}(n,s)|^2$ is estimated in the same way with a scope of $L = 256$ and a range of $0 \le \kappa \le 128$. The indices n_u and n_k, which we retrieve from condition (3), are applied to the partitioned parts of $\hat{G}(n,s)$.

3. SIMULATION RESULTS

Figure 3 shows the residual echo's broadband signal power as a function of time (we used a white noise excitation signal $x(k)$). The reverberation time τ_{60} of the used echo path impulse response accounted to $100ms$. We have employed simulated impulse responses using the well-known image method [8]. We can see that the proposed method delivers almost bias-free estimates. After 50,000 samples the echo path impulse response was modified. The estimate can follow the actual value rather quickly.

A near-end speaker became active between sample 80,000 and 100,000. Here, the estimate is affected and we get a bias of about $4dB$. However, without the minimum statistics estimation method the results would be much worse. We have performed the simulation at a signal-to-noise ratio (SNR) of $40dB$ of the 'reference noise' against the background noise. At an SNR of $15dB$ the estimates are biased at about $1.5dB$; at $20dB$ we get a bias of only $0.5dB$. Finally, we show that the new method works with speech signals as an reference signal $x(k)$ as well (figure 4). The scenario has been the same as during the simulations for figure 3. We can see that neither the change of the echo path impulse response nor the near-end speaker disrupt the estimates drastically. However, there is a certain variance in the residual echo's estimated signal power against its actual version.

4. APPLICATIONS

The design of a single-channel Wiener filter according to equation (1) is only one possible application of the proposed residual echo estimation method. We could also calculate a minimum variance distortionless response (MVDR) beamformer, which is especially suited to suppress the residual echo without being sensitive against steering errors of the associated microphone array. The knowledge of the isolated signal $b(k)$ can also be used for a reliable step-size control in the AEC.

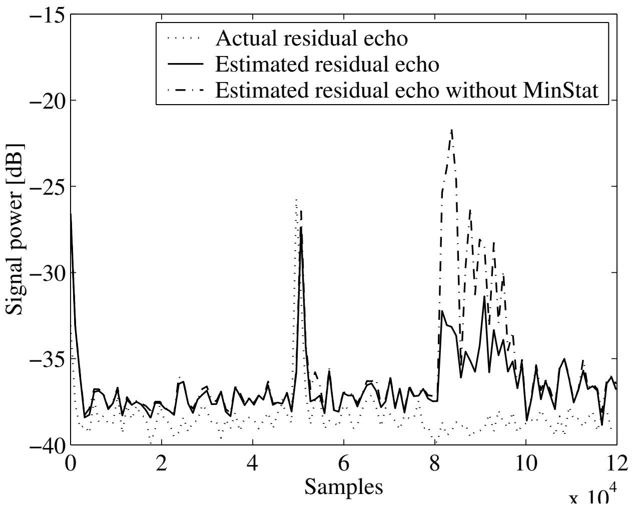

Figure 3: Residual echo's signal power as a function of time (estimated with and without minimum statistics and actually measured within our simulation environment) using white noise for the excitation signal $x(k)$.

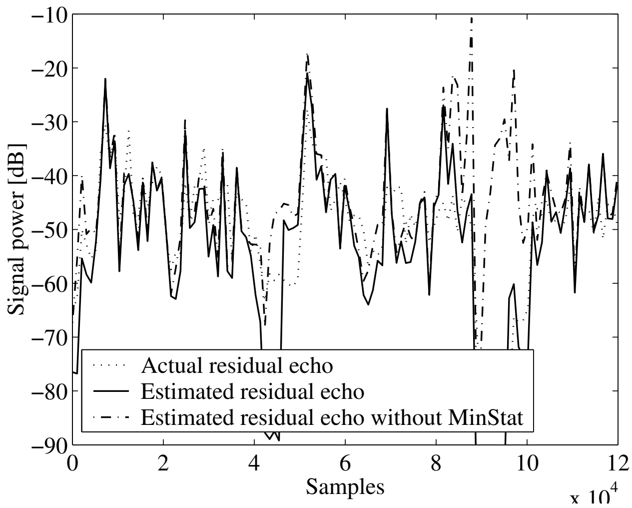

Figure 4: Residual echo's signal power as a function of time (estimated with and without minimum statistics and actually measured within our simulation environment) using a speech signal for the excitation signal $x(k)$.

5. CONCLUSIONS

In this contribution we have introduced a new method to estimate the residual echo signal at the output of an AEC. The simulations have shown that the results are hardly biased and rather robust against interferences introduced by near-end speech signals. Additionally, the estimates are not disrupted by additive noise up to an SNR of $20dB$. Therefore, our method can come into operation in office environments without additional measures. By using the residual echo quite a number of applications are possible. Experiments with a single-channel Wiener filter have shown that the estimation method can quickly react to sudden changes of the echo path impulse response and thus support the AEC efficiently. Informal listening test have confirmed the robustness during double-talk situations: The near-end speech signal was hardly distorted.

6. REFERENCES

[1] C. Breining, P. Dreiseitel, E. Hänsler, 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 Processing Magazine*, pp. 42–69, July 1999.

[2] S. Gustafsson, R. Martin, and P. Vary, "Combined Acoustic Echo Control and Noise Reduction for Hands–Free Telephony," *Signal Processing*, vol. 64, pp. 21–32, 1998.

[3] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *European Signal Processing Conference (EUSIPCO–94)*, (Edinburgh, UK), pp. 1182–1185, September 1994.

[4] G. Enzner, R. Martin, and P. Vary, "On Spectral Estimation of Residual Echo in Hands–Free Telephony," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, (Darmstadt, Germany), pp. 211–214, Sep 2001.

[5] M. Kallinger, J. Bitzer, and K. D. Kammeyer, "Interpolation of MVDR Beamformer Coefficients for Joint Echo Cancellation and Noise Reduction," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, (Darmstadt, Germany), pp. 39–42, Sep 2001.

[6] C. M. Rader, "An Improved Algorithm for High Speed Autocorrelation with Application to Spectral Estimation," *IEEE Trans. on Audio and Electroacoustics*, vol. 18, Dec 1979.

[7] J.-S. Soo and K. Pang, "Multidelay Block Frequency Domain Adaptive Filter," *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. 38, pp. 373–376, Feb 1990.

[8] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small–Room Acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.