7. Stereophonic Acoustic Echo Cancellation Using NLMS-OCF

Echo in stereo-teleconferencing systems is undesirable but inevitable. Echo cancelers are used to mitigate the echo. Stereophonic echo cancelers suffer from two problems that are generally absent in mono echo cancellers [50]. Firstly, the adaptive filter might misconverge due to the correlation between the left and right channel signals [51, 52]. Secondly, again due to the correlation between the stereo signals, simple adaptation algorithms such as LMS and NLMS will converge slowly. Several solutions have been proposed to ameliorate the performance of stereo echo cancelers. Most of these solutions attempt to decorrelate the left and right channel signals by using some pre-processing technique [53, 54, 55]. These techniques essentially distort the signals, which is undesirable. Furthermore, the achievable performance improvement might be limited by the distortion inaudibility restrictions.

In this paper, we propose using NLMS-OCF to adapt the echo canceler, which is modeled as a two-input single-output FIR filter. The proposed strategy does not introduce any distortion and it mitigates some of the difficulties encountered in typical stereophonic echo cancelers. This flexibility provided by NLMS-OCF in choosing the vectors used for adaptation can be exploited to reduce the misconvergence of echo cancelers and to improve their convergence speed.

7.1 Stereo Echo Canceler Adaptation Equations

The stereo echo canceler configuration is shown in Figure 7.1. To avoid clutter, in Figure 7.1, we show the echo paths corresponding to only one of the two stereo channels. In reality, similar paths exist in the other channel as well.

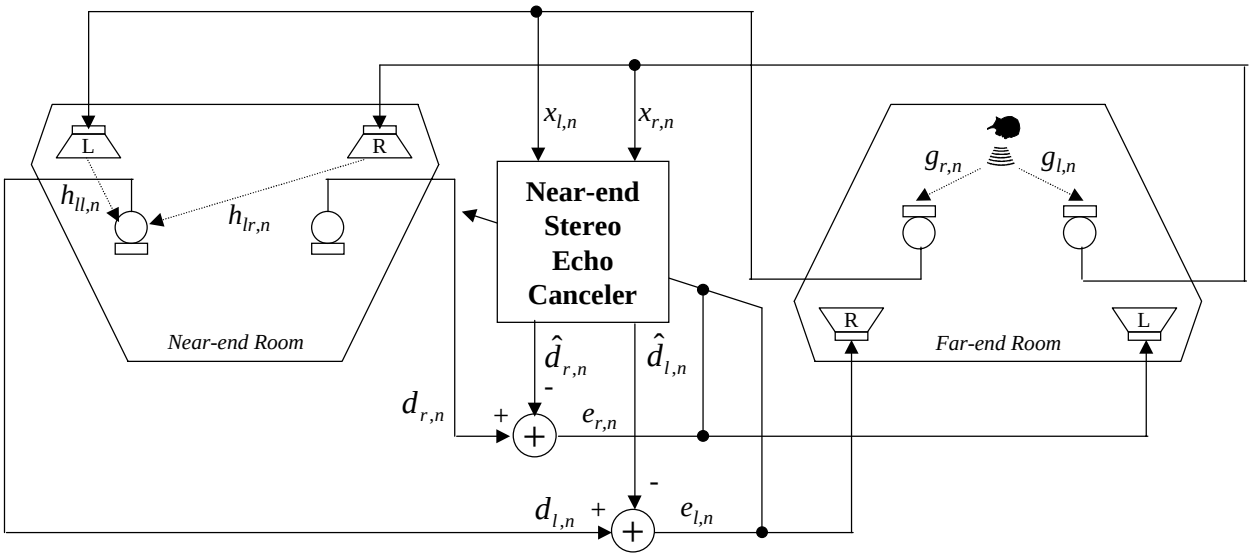


Figure 7.1 Stereophonic Echo Canceler Configuration.

Let us denote the impulse responses from the speech source in the far-end room to the right and left microphones as $g_{r,n}$ and $g_{l,n}$, respectively. The impulse responses of the echo-paths from the left and right speakers to the left microphone in the near-end room are assumed to be $h_{ll,n}$ and $h_{lr,n}$, respectively. Let $d_{l,n}$ be the echo received by the left microphone. The adaptive echo canceler (AEC), which is modeled using FIR filters, generates an estimate $\hat{d}_{l,n}$ for the echo, which is subtracted from the true echo to form the error signal $e_{l,n}$. The left-channel echo canceler weight vector $\hat{\mathbf{h}}_{l,n} = \left[\hat{\mathbf{h}}_{ll,n}^T \quad \hat{\mathbf{h}}_{lr,n}^T\right]^T$ is adapted with the objective of minimizing the mean squared left-channel residual echo, $E\left(e_{l,n}^2\right)$, in the absence of near-end speech. The dimensions of $\hat{\mathbf{h}}_{ll,n}$ and $\hat{\mathbf{h}}_{lr,n}$ are assumed to be N_l and N_r, respectively. That is, N_l and N_r are the estimated lengths of the left and right echo paths, respectively.

We propose using the NLMS-OCF algorithm to adapt the echo canceler weight-vector. The adaptation equation is as follows:

$$\hat{\mathbf{h}}_{l,n+1} = \hat{\mathbf{h}}_{l,n} + \mu_{l,0}\mathbf{x}_n + \mu_{l,1}\mathbf{x}_n^1 + \ldots + \mu_{l,M}\mathbf{x}_n^M \tag{7.1}$$

where $\mathbf{x}_n$ is the input vector at the *n*th instant given by

$$x_n = \left[x_{l,n} \quad x_{l,n-1} \quad \cdots \quad x_{l,n-N_l+1} \quad \vdots \quad x_{r,n} \quad x_{r,n-1} \quad \cdots \quad x_{r,n-N_r+1}\right]^T \tag{7.2}$$

and $\mu_{l,k}$, for $k = 0,1,...,M$ are chosen as in (3).

$$\mu_{l,k} = \begin{cases} \dfrac{\bar{\mu} e_{l,n}}{\mathbf{x}_n^H \mathbf{x}_n} & \text{for } k = 0, \text{ if } \|\mathbf{x}_n\| \neq 0 \\ \dfrac{\bar{\mu} e_{l,n}^k}{\mathbf{x}_n^{k\,H} \mathbf{x}_n^k} & \text{for } k = 1,2,...,M, \text{ if } \|\mathbf{x}_n^k\| \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{7.3}$$

where

$$\begin{aligned} e_{l,n} &= d_{l,n} - \hat{\mathbf{h}}_{l,n}^H \mathbf{x}_n \,, \\ e_{l,n}^k &= d_{l,n-kD} - \left(\hat{\mathbf{h}}_{l,n}^k\right)^H \mathbf{x}_{n-kD}, \quad \text{for } k = 1,2,...,M, \text{ and} \\ \hat{\mathbf{h}}_{l,n}^k &= \hat{\mathbf{h}}_{l,n} + \mu_{l,0}\mathbf{x}_n + \mu_{l,1}\mathbf{x}_n^1 + ... + \mu_{l,k-1}\mathbf{x}_n^{k-1}. \end{aligned} \tag{7.4}$$

Similarly the right-channel echo canceler weights $\hat{\mathbf{h}}_{r,n} = \left[\hat{\mathbf{h}}_{rl,n}^T \quad \hat{\mathbf{h}}_{rr,n}^T\right]^T$ are estimated so that the corresponding mean squared residual echo, $E\left(e_{r,n}^2\right)$, is minimized. The weights are adapted as shown below:

$$\hat{\mathbf{h}}_{r,n+1} = \hat{\mathbf{h}}_{r,n} + \mu_{r,0}\mathbf{x}_n + \mu_{r,1}\mathbf{x}_n^1 + ... + \mu_{r,M}\mathbf{x}_n^M \tag{7.5}$$

where $\mu_{r,k}$ for $k \in \{0,1,\ldots,M\}$ are computed similarly to $\mu_{l,k}$ by replacing all the left-channel parameters and signals with right-channel parameters and signals. Note from (7.1) and (7.5) that the orthogonal correction factors used by both left- and right-channel echo cancelers are identical. Hence, they need to be evaluated only once.

7.2 Non-Uniqueness Problem

A serious problem encountered in stereophonic echo cancelers is that the echo canceler coefficients do not converge to the true impulse response of the echo path. Due to the cross-correlation between the left and right channel signals, the weight estimate that minimizes the error between echo d_n and estimated echo $\hat{d}_n$ is not unique. In fact there is an affine space of valid minimizing solutions [52], which is determined by the cross-correlation between the left and right channel signals, and the weight estimate converges to the point in this affine space that

is nearest to the initial guess. Hence, the estimated weights need not necessarily match the true weights. If there is any change in the far-end room, the correlation between the left and right signal changes, which in turn leads to a change in the minimizing-solution space. Consequently, the adaptation algorithm needs to readapt the parameters when there is a change not only in the near-end room but also in the far-end room. However, the estimation error between the true and estimated weights reduces with every variation in the cross-correlation between the stereo signals. The stereo projection algorithm, which is equivalent to the affine projection algorithm (APA), is used to emphasize the variations in the cross-correlation by using a block of input vectors for adaptation, thereby accelerating the convergence of the echo canceler weights. When there is a change in the cross-correlation, the input vector block used for adaptation consists of input signals with different cross-correlations. This accelerates the convergence of the estimated weights to the true weights. The variations in cross-correlation can be further emphasized by using the NLMS-OCF algorithm as explained below.

Consider the scenario where one talker in the far-end room stops talking and another starts talking. This results in an abrupt (step) change in the far-end room impulse responses $g_{r,n}$ and $g_{l,n}$, say at time instant k. Suppose NLMS-OCF is used for adaptation, with M orthogonal correction factors based on input vectors that are spaced D apart, then the input vector block used for adaptation from $n = k$ to $n = k + MD$ consists of input signals that are correlated differently. Thus a larger D helps to emphasize the variations. The stereo projection algorithm, which restricts D to unity, does not provide as much flexibility.

Consider another scenario where a time-varying all-pass filter is used to artificially introduce variations in the cross-correlation between the left and right signals [54]. The time variation of the all-pass filter has to be slow so that the changes are inaudible to the listener. However, for the stereo projection algorithm to successfully exploit the variations in cross-correlation between the left and right channels, there should be a significant change in the all-pass filter characteristics within the time window of data used for adaptation. By using NLMS-OCF with a D larger than unity, the time window of data used for adaptation is expanded. Hence, by using NLMS-OCF with a larger D the variations of the all-pass filter can be slowed down without compromising on the convergence speed of the echo canceler.

In practice, the cross-correlation between the stereo signals varies slightly even when the talker does not move the head or body while speaking [55]. Even though such variations are

small, they can be effectively emphasized, by using NLMS-OCF, since by choosing a large *D* we adapt using input signals that are farther apart in time (hence with larger variation in cross-correlation).

7.3 Convergence Rate

The rate of convergence of the echo canceler estimates depends on the eigenvalue spread of the autocorrelation matrix of the input vector (also referred to as the information vector). An important distinction between the mono echo canceler and stereo echo canceler is that the eigenvalue spread depends only on the input signal in the mono case while the spread is significantly larger in the stereo case, due to the correlation between the left and right signals, irrespective of the input signal. This results in slow convergence of the stereo echo canceler coefficients. Using NLMS-OCF can mitigate this problem. It has been observed that, unless the step size μ is "too small," a larger value of D results in faster convergence. Thus, the flexibility provided by NLMS-OCF in selecting the input vectors used for adaptation can be exploited to accelerate the convergence of the echo canceler.

7.4 Simulation Results

A stereo-echo canceler with its coefficients adapted using the NLMS-OCF algorithm is simulated in MATLAB. The near-end and far-end room left-channel impulse responses are modeled using a 255^{th} order FIR filter estimated based on measurements from two actual rooms. Slight perturbations of the left-channel impulse responses are used to model the right-channel impulse responses. The impulse responses are shown in Figure 7.2. A real speech signal, shown in Figure 7.3, is used as the far-end input. All simulations use soft initialization; that is, the delay line of the adaptive filter is filled with true data values before the simulation is started. All the adaptive filter coefficients are initialized to zero. We assume far-end only talk (with only one far-end talker at any time) for our simulations, since the adaptive filters are usually adapted only under this condition.

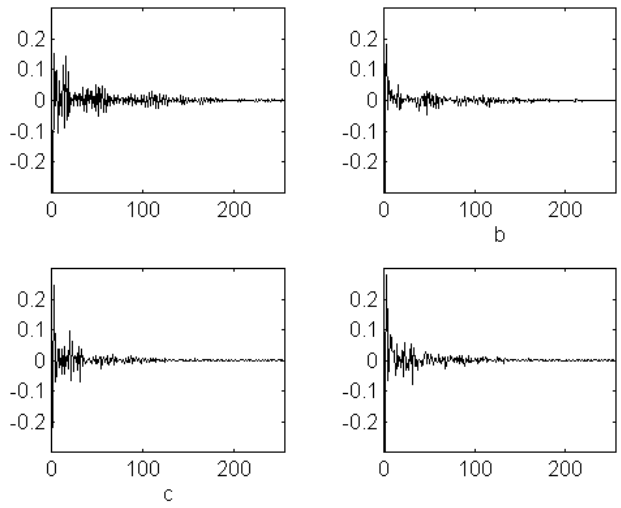


Figure 7.2 Channel Impulse Responses Used for Simulation:
(a) Far-end Left, (b) Far-end Right, (c) Near-end Left, and (d) Near-end Right.

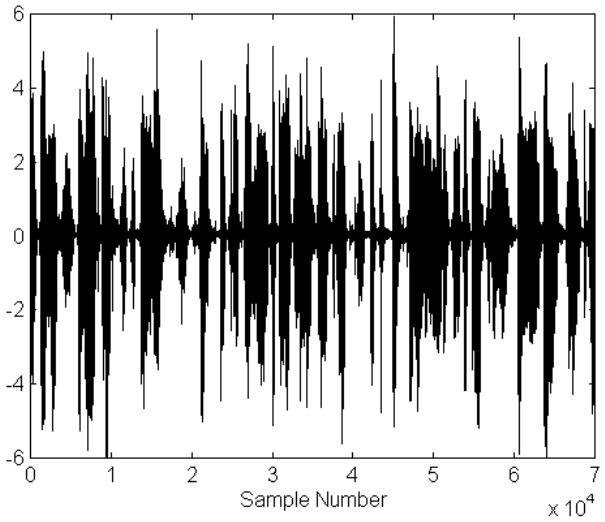
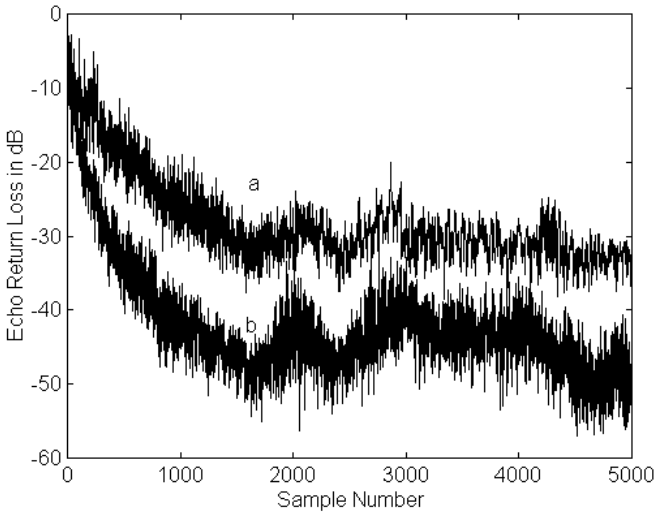


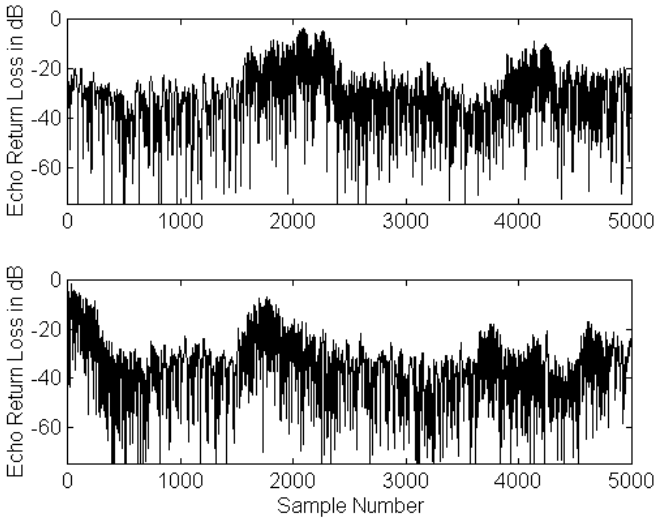
Figure 7.3 Test Speech Signal (Sampling Rate 8 kHz).

The simulated echo canceler uses a 255th order FIR filter to model each of the left and right channels of the near-end room. The echo canceler coefficients are adapted using the NLMS-OCF algorithm. The step-size $\bar{\mu}$ is chosen to be unity. The number of vectors used for adaptation is fixed at 20 (i.e. *M* equals 19). Different delay values are chosen, viz. *D* = 1, or 64, to select the input vectors used for adaptation. Figure 7.4(i) shows the learning curves corresponding to these delays obtained by ensemble averaging 25 different realizations, after removing 2 outliers. The

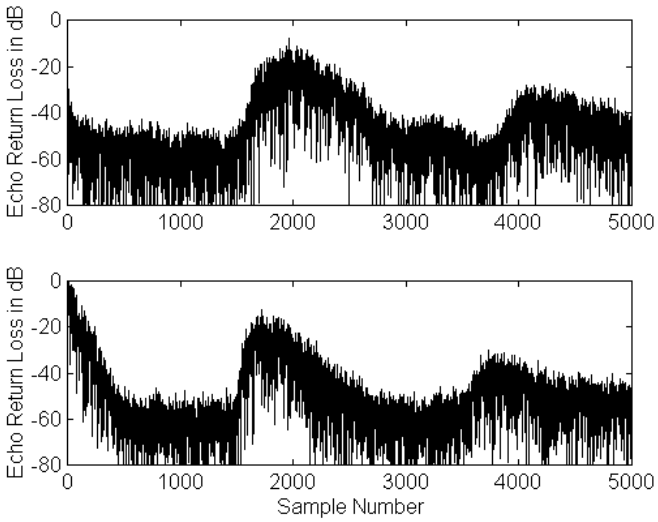
ignored outliers are shown in Figures 7.4 (ii) and (iii). The echo return loss is defined as the ratio of residual echo energy to original echo energy. The 25 input realizations are derived as 8000-point-long segments from the speech signal shown in Figure 7.3, with 7000-point overlap between successive realizations. These simulations assume presence of measurement noise at 60 dB below the echo signal. Hence, the theoretical limit on achievable ERL is –60 dB. We observe a two-fold advantage in using D larger than unity. Firstly, a larger value of D results in faster convergence. Secondly, a larger value of D results in better echo rejection (better by around 12 dB) because of less misadjustment.



(i)



(ii)



(iii)

Figure 7.4 (i) Learning Curves of NLMS-OCF Corresponding to Speech Input for Different Values of Delay *D*: (a) *D* = 1 and (b) *D* = 64. (ii) Outliers Ignored Corresponding to *D* = 1, and (iii) Outliers Ignored Corresponding to *D* = 64.

Less misadjustment with larger values of D can be explained as follows. The weight increment of NLMS-OCF, during each iteration, can be written in matrix notation, as in the case of APA [13], as

$$\Delta\hat{\mathbf{w}}_{n+1} = \bar{\mu}\mathbf{X}_n\left(\mathbf{X}_n^t\mathbf{X}_n\right)^{-1}\mathbf{e}_n^* \tag{7.6}$$

where $\mathbf{X}_n = \begin{bmatrix}\mathbf{x}_n & \mathbf{x}_{n-D} & \cdots & \mathbf{x}_{n-MD}\end{bmatrix}$ is the matrix comprising the vectors used for adaptation as its columns and $\mathbf{e}_n = \begin{bmatrix}e_{n/n} & e_{n-D/n} & \cdots & e_{n-MD/n}\end{bmatrix}^T$ is the error vector. Here, $e_{k/n}$ is the error in estimating the output at time instant k based on the weight estimate at n. That is, $e_{k/n} = d_k - \hat{\mathbf{w}}_n^H\mathbf{x}_k$. The error vector consists of two components, one due to the true output estimation error and the other due to the measurement noise. The measurement noise component contributes to the misadjustment. The amount of the contribution depends on $\left(\mathbf{X}_n^t\mathbf{X}_n\right)^{-1}$, which in turn depends on the condition number of $\mathbf{X}_n^t\mathbf{X}_n$. If $\mathbf{X}_n^t\mathbf{X}_n$ is ill conditioned, then $\left(\mathbf{X}_n^t\mathbf{X}_n\right)^{-1}$ will amplify the measurement noise more than when $\mathbf{X}_n^t\mathbf{X}_n$ is well conditioned. For smaller values of delay the vectors used for adaptation, viz. $\mathbf{x}_n, \mathbf{x}_{n-D}, \ldots, \mathbf{x}_{n-MD}$, are likely to be more correlated

than when the delay is large. Hence, $\mathbf{X}_n^t \mathbf{X}_n$ is more likely to be better conditioned for larger values of D than for smaller values. This explains the better echo rejection with larger delay.

The non-uniqueness problem is evident from the misalignment curve shown in Figure 7.5. We define misalignment ξ_n as

$$\xi_n = \frac{\left\| h_n - \hat{h}_n \right\|}{\left\| h_n \right\|} \tag{7.7}$$

where $\|\cdot\|$ is the standard Euclidean norm. The steady-state misalignment is large (nearly 0.45), even though the steady-state ERL is better than -30 dB, due to the non-uniqueness of the minimizing solution. Consequently, if there is any change in the far-end room impulse responses, the residual echo level temporarily increases, even if there is no change in the near-end room.

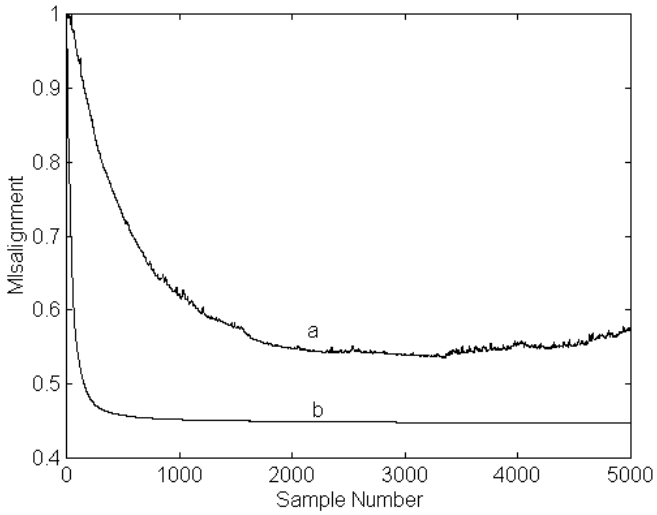


Figure 7.5 Misalignment of Echo Canceler Coefficients Corresponding to Speech Input for Different Values of Delay *D*: (a) *D* = 1 and (b) *D* = 64.

We see that both residual echo and misalignment decrease faster for larger values of *D*. NLMS-OCF with *D* = 1 corresponds to the stereo projection algorithm. Hence, the flexibility of choosing input vector delays in the NLMS-OCF adaptation was exploited here to achieve faster convergence as well as better echo return loss than that obtained with the stereo projection algorithm (corresponding to the *D*=1 case).

Now, we present results that show that the NLMS-OCF algorithm can emphasize variations in cross-correlations better than the stereo projection algorithm, thereby converging faster to the true solution for echo channel weights. All the parameters are chosen as before. Measurement noise is assumed to be absent so that the misalignment is solely due to the non-uniqueness problem. We simulate the condition where one far-end talker stops talking and another far-end talker starts talking. This results in an abrupt step change in the channel impulse responses in the far-end room, thereby introducing variations in the near-end received signal cross-correlation. This change is assumed to occur at n = 7000, 15500, and 24000. The simulation results are shown in Figure 7.6. We observe that NLMS-OCF with larger delays converges faster, to lower misalignment values, than the stereo projection algorithm, which uses unit delay by default.

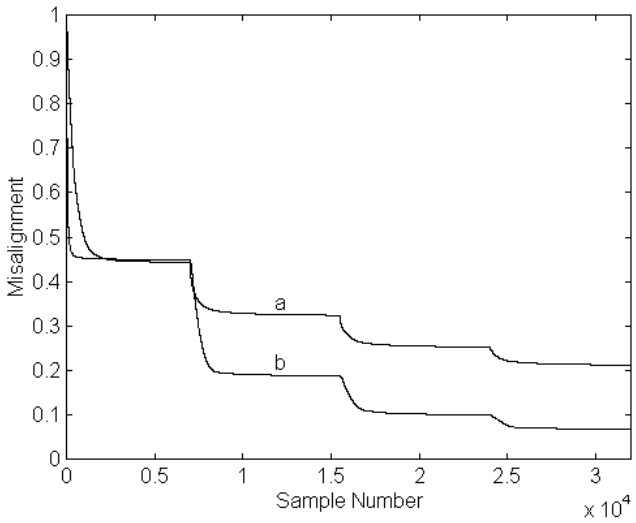
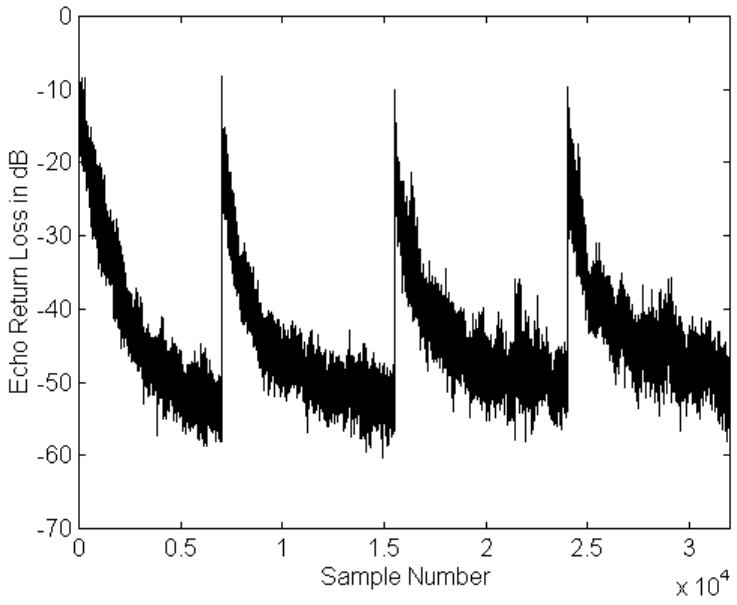
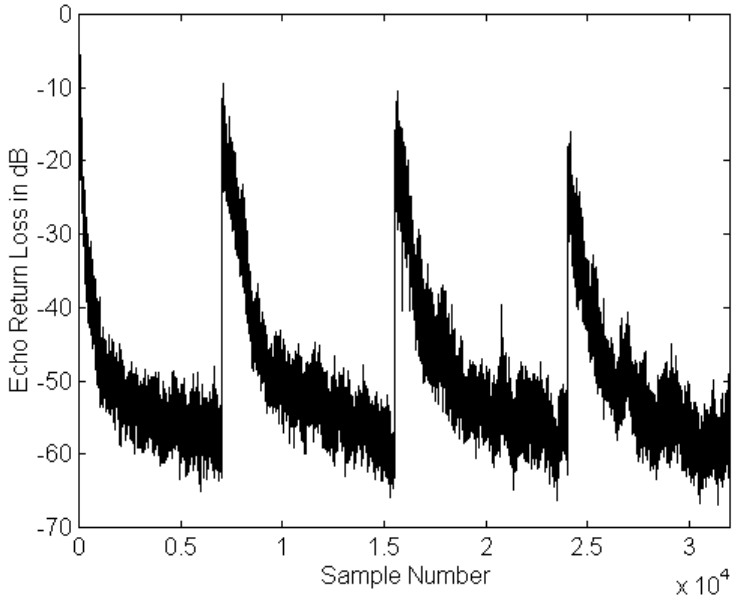


Figure 7.6 Misalignment of Echo Canceler with Step Changes in Cross-Correlation of Stereo Signals Corresponding to Different Delays: (a) *D* = 1 and (b) *D* = 64.

Figure 7.7 shows the learning curve corresponding to the changing far-end room scenario for $D = 1$ and for $D = 64$. Observe that the echo level increases whenever there is a change in the far-end room, which is undesirable. This is due to the misalignment of the echo canceler coefficients. We also note that the maximum temporary increase in the echo level decreases as the misalignment is reduced. It is also seen that the maximum temporary increase decreases faster with $D = 64$ than with $D = 1$.



(a)



(b)

Figure 7.7 Learning Curves of Echo Canceler with Step Changes in Cross-Correlation corresponding to (a) *D* = 1 and (b) *D* = 64.

Comment 1: While increasing *D* results in performance improvement, there are some drawbacks in using a larger *D*. Firstly, a larger D necessitates storing more of the past data and hence requires more memory. Secondly, using a very large *D* can affect the near-end tracking performance of the echo canceler. If the time window of data used for adaptation is so long that

the near-end room characteristics change significantly within that time, then the weight estimates can never track the weights correctly. This problem can be mitigated by using an exponential weighting on the orthogonal correction factors. That is, the weight adaptation equations are modified to

$$\hat{\mathbf{h}}_{l,n+1} = \hat{\mathbf{h}}_{l,n} + \mu_{l,0}\mathbf{x}_n + \mu_{l,1}\lambda\mathbf{x}_n^1 + \mu_{l,2}\lambda^2\mathbf{x}_n^2 + \ldots + \mu_{l,M}\lambda^M\mathbf{x}_n^M \tag{7.8a}$$

$$\hat{\mathbf{h}}_{r,n+1} = \hat{\mathbf{h}}_{r,n} + \mu_{r,0}\mathbf{x}_n + \mu_{r,1}\lambda\mathbf{x}_n^1 + \mu_{r,2}\lambda^2\mathbf{x}_n^2 + \ldots + \mu_{r,M}\lambda^M\mathbf{x}_n^M \tag{7.8b}$$

where $\lambda \in (0,1)$ is the weighting factor. Of course, lower values of λ result in better tracking at the cost of a reduction in convergence rate.

Comment 2: The presence of independent noise in the stereo channels, which could be caused by quantization or by independent electronic circuitry in the two channels, can decorrelate the two channels and thereby mitigate the non-uniqueness problem. We investigated this using simulation by adding independent white noise at various levels in the stereo channels. The models used for far-end and near-end rooms are the same as in our earlier simulations and we assume that no change in echo path occurs in either room. The misadjustment characteristics corresponding to different noise variances are shown in Figure 7.8. We observe that for significant mitigation of the non-uniqueness problem the SNR in each channel has to be as low as 20 dB.

We repeated our simulation using a real stereo signal, recorded using “relatively noisy” electronics as far-end stereo signals. The estimated SNR of the recorded signal is approximately 23 dB. The near-end room effects are simulated using the same room impulse responses as above. The resulting misalignment characteristic is shown in Figure 7.9. We observe that the misalignment characteristic shown in Figure 7.9 lies somewhere between the misalignment characteristics (c) and (d), corresponding to SNRs of 30 dB and 40 dB respectively, shown in Figure 7.8. This makes sense because there is some correlation between the left and right channel noise in the recorded signal while in the simulation we added truly independent noise to the left and right channel signals. The noise in the recorded signal dominates the effects due to quantization. If there is any effect due to head movement it seems to be small compared to the effect of the noise that is present.

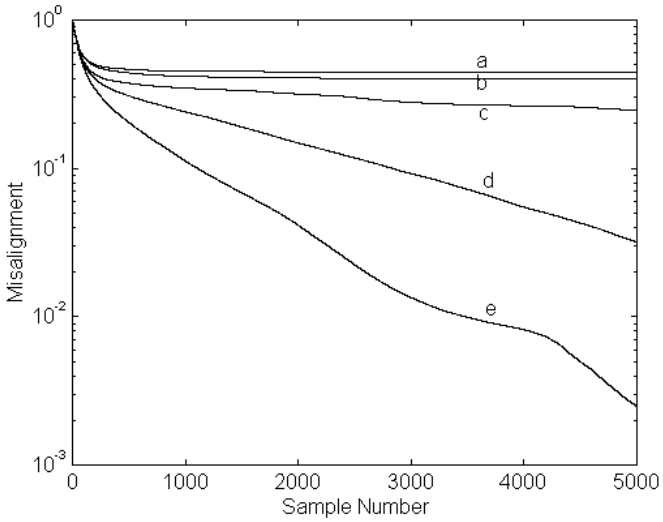


Figure 7.8 Misalignment Characteristic of Echo Canceler Corresponding to Different Levels of Noise in the Stereo Channels (a) SNR = 80 dB, (b) SNR = 60 dB, (c) SNR = 40 dB (d) SNR = 30 dB and (e) SNR = 20 dB.

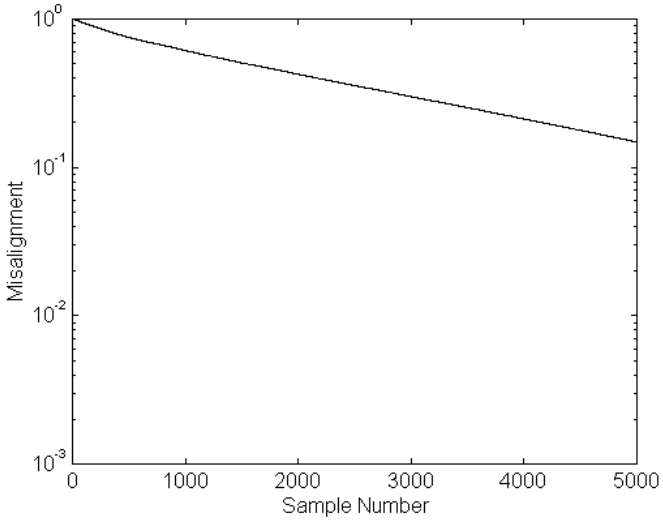


Figure 7.9 Misalignment Characteristic of Echo Canceler with Recorded Stereo Signal as Far-end Input.

Comment 3: The convergence rate of APA, which by default uses an input vector delay of unity, can be improved only by increasing the number of orthogonal correction factors. However, NLMS-OCF provides flexibility in choosing the input vector delay and this also can be exploited

to improve the convergence rate. This is evident from the simulation result shown in Figure 7.10. The learning curves shown are the ensemble average of 25 different realizations obtained under noise-free conditions with independent speech as input. We see that APA with four OCFs converges at almost the same rate as NLMS-OCF that uses just one additional OCF but with D = 512. However, due to significantly lower computational requirements of fast APA ($2N + 21M$) compared to fast NLMS-OCF ($5NM + 2N + 6M + 4$), APA with $M = 4$ requires less computational effort than NLMS-OCF with $M = 1$. Hence, NLMS-OCF does not provide any computational advantage over APA.

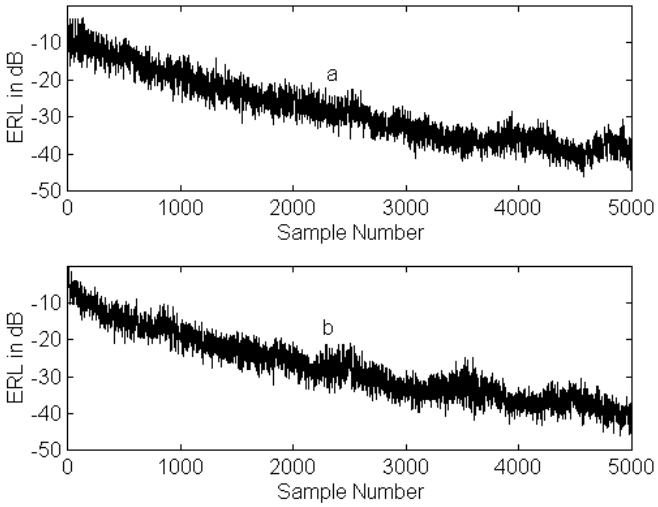


Figure 7.10 Learning Curves of (a) APA with M = 4 and (b) NLMS-OCF with M = 1 and D = 512.

7.5 Conclusion

We presented a stereo echo canceler adapted using the NLMS-OCF algorithm. The NLMS-OCF algorithm is shown to produce faster convergence than the widely used stereo projection algorithm, in terms of faster improvement in echo-return loss as well as faster reduction in impulse response misalignment. Furthermore, NLMS-OCF provides better echo rejection. This is achieved at the cost of increased computational complexity.