A CLIPPING-BASED ADAPTIVE FILTERING APPROACH FOR STEREOPHONIC ACOUSTIC ECHO CANCELLATION

Mehdi Bekrani 1, Andy W. H. Khong 2, Mojtaba Lotfizad 1

1 Tarbiat Modares University, Iran,
2 Nanyang Technological University, Singapore

ABSTRACT

The use of partial-updating algorithm for reducing interchannel coherence in stereophonic acoustic echo cancellation has been proposed recently. In this work, we show that this algorithm suffers from the lack of robustness against source positions in the transmission room. To address this, we present an insight into this problem and propose a center clipping algorithm that improves the joint optimization between reducing interchannel coherence and increasing energies of selected taps for different source positions. Simulation results using both colored and speech inputs verify the robustness of the proposed algorithm against source positions.

Index Terms— stereophonic acoustic echo cancellation, cross-correlation, center-clipping, tap-selection, partial update

1. INTRODUCTION

Stereophonic acoustic echo cancellation (SAEC) is an important part of stereo audio/visual communication systems that suppresses acoustic echoes arising from the acoustic coupling between the microphones and loudspeakers in the receiving room. The advantage of these systems over monophonic systems is the increase in spatial realism which provides a better understanding of the sound.

In a practical scenario, the performance of stereophonic acoustic echo cancellers is often limited by the mismatch between the adaptive filter coefficients and the acoustic impulse responses. This performance degradation is due to the misalignment problem [1]. Several algorithms have been proposed to address this problem by reducing the interchannel coherence between the two transmitted signals so as to improve the convergence rate of the adaptive filters. These include the addition of controlled quantities of independent noises [2] and the use of comb filtering [3]. These algorithms often require a trade-off between audio quality as well as achieving good convergence performance for the adaptive filters. It is therefore apparent that adaptive algorithms for SAEC needs to decorrelate the transmitted signals without significantly degrading the audio quality or destroying the stereo image of the transmission room.

One of the most popular methods of achieving efficient decorrelation is the use of a non-linear (NL) pre-processor which adds a nonlinear function of the transmitted signal in each channel to the signal itself [1] [4]. More recent advances involve decorrelation of the tap-input vectors as oppose to the transmitted signals [5] [6]. Among these recent methods, the exclusive-maximum (XM) tap-selection algorithm [5] has been shown to be an attractive solution. This algorithm achieves signal decorrelation by ensuring that only an exclusive set of filter coefficients from each channel are selected for adaptation. In order to reduce any degradation in convergence performance due to this subselection, the algorithm further ensures that the energies of these exclusive tap-inputs are maximized. As shown in [5], the XM tap-selection has been incorporated with the NL preprocessor and the normalized least-mean-square (NLMS) algorithm. The resulting XMNL-NLMS has shown to achieve higher rate of convergence compared to that of NL-NLMS.

In this paper, we propose to improve the robustness and convergence performance of XMNL-NLMS. Our motivation is derived from the poor initial convergence and steady-state performance of XMNL-NLMS when the interchannel coherence between the two tap-input vectors is relatively low such as occur when the source position is away from the microphone pair centroid in the transmission room. To address this problem, we propose to employ a center-clipping algorithm in order to ensure that the low interchannel coherence and maximization of tap-input energy criterion are jointly optimized further. The proposed algorithm ensures that the convergence and the steady-state performance of the adaptive filters will be robust to the source position in the transmission room.

2. STEREOPHONIC ACOUSTIC ECHO CANCELLATION

Figure 1 illustrates an SAEC system where a source \( s(n) \) generates stereophonic signals \( x_1(n) \) and \( x_2(n) \) that are received by the microphones via acoustic paths \( g_{1}(n) \) and \( g_{2}(n) \) in the transmission room. These stereophonic signals are then transmitted to the loudspeakers in the receiving room which produces an echo

\[
y(n) = h_{1}^{T}(n)x_1(n) + h_{2}^{T}(n)x_2(n) + w(n) ,
\]

where \( h_{i}(n) = [h_{i,0}(n), h_{i,1}(n), \ldots, h_{i,L-1}(n)]^T \) is the unknown impulse response, \( x_{i}(n) = [x_{i}(n), \ldots, x_{i}(n - L + 1)]^T \) and \( i = 1, 2 \) is the channel index while \( w(n) \) is the background noise and \( L \) is the length of \( h_{i}(n) \). For simplicity, we consider only one microphone in the receiving room, since similar analysis can be applied to the other channel [1]. For effective SAEC, two adaptive

![Fig. 1. SAEC in teleconferencing. Thick line in the transmission room shows the source locus used for experiments.](image-url)
filters $\hat{h}_1(n)$ and $\hat{h}_2(n)$ are employed to estimate $h_1(n)$ and $h_2(n)$. The error signal between $y(n)$ and its estimate is thus given by

$$e(n) = y(n) - [\hat{h}_1(n)x_1(n) + \hat{h}_2(n)x_2(n)],$$

(2)

where $\hat{h}_i(n) = [\hat{h}_{i,0}(n), \hat{h}_{i,1}(n), \ldots, \hat{h}_{i,L-1}(n)]^T$ is the vector of adaptive filter coefficients of the $i^{th}$ channel.

3. THE XM TAP-SELECTION FOR SAEC

Due to the high interchannel coherence between $x_1(n)$ and $x_2(n)$, the SAEC problem is ill-conditioned and as a result, NLMS suffers from poor convergence [1]. Several algorithms have since been proposed to decorrelate the audio signals without severely degrading the audio quality and stereophonic image. One of the most recent is the XMNL-NLMS algorithm [5]. This algorithm incorporates the nonlinear preprocessor [1] and a tap-selection scheme that reduces the interchannel coherence by selecting an exclusive set of filter coefficients across both channels for adaptation. The degradation in convergence due to this partial updating is then minimized by jointly maximizing the $L_2$ norm of the selected tap-inputs across both channels. The benefit of XMNL-NLMS is that it operates on the tap-input vectors instead of the transmitted signals which as a consequence reduces any distortion as oppose to NL pre-processing.

As explained in [5], the XMNL-NLMS algorithm requires the NL preprocessor such that the processed signals are given by

$$x'_1(n) = x_1(n) + 0.5\alpha[x_1(n) + x_1(n)],$$

(3)

$$x'_2(n) = x_2(n) + 0.5\alpha[x_2(n) + x_2(n)],$$

(4)

where $\alpha$ controls the amount of non-linearity and a value of $\alpha = 0.5$ brings about a good compromise between speech quality and convergence rate of the adaptive algorithm [1]. The weight update for the XMNL-NLMS algorithm is then performed by

$$\hat{h}_i(n+1) = \hat{h}_i(n) + \frac{\mu}{\|x'(n)\|^2_2} e(n)q_i(n)x'_i(n),$$

(5)

where $x'(n) = [x'_1(n) \ x'_2(n)]^T$ and $\mu$ is the step-size while $\epsilon$ is the regularization parameter to prevent division by zero. The $L \times L$ matrix $Q_i(n) = \text{diag}(q_i(n))$ is the tap-selection matrix for the $i^{th}$ channel given that elements in the $L \times 1$ vector $q_i(n) = [q_{i,1}(n), \ldots, q_{i,L}(n)]^T$ are given by

$$q_{1,u}(n) = \begin{cases} 1 & p_u \in \{0.5L \text{ maxima of } p(n)\} \\ 0 & \text{otherwise} \end{cases},$$

(6)

$$q_{2,u}(n) = \begin{cases} 1 & p_u \in \{0.5L \text{ minima of } p(n)\} \\ 0 & \text{otherwise} \end{cases},$$

(7)

$$p(n) = [x_1(n) - x_2(n)],$$

(8)

where $u, v = 1, 2, \ldots, L$ represent element $u$ of $q_{1}(n)$ and element $v$ of $q_{2}(n)$ such that $[x'_1(n)] = [x_1(n)], \ldots, [x'_1(n-L+1)]$. We note that $e(n)$ in (5) is computed using (2).

3.1. Robustness of the XMNL-NLMS algorithm

One of the problems that has not been investigated for XMNL-NLMS is its robustness to the source positions in the transmission room. To illustrate this, we vary, by way of simulation using the method of images [7], the source position starting from the front of the microphone array centroid at (2.85, 1.85, 1.6) m to the front of the right microphone at (3.1, 1.85, 1.6) m as shown in Fig. 1. In this experiment, the microphones are positioned at (2.7, 2.1, 5) m and (3.2, 1.5) m in a room of dimension $7 \times 7 \times 4$ m.

Figure 2 shows the convergence performance, averaged over 10 independent trials, of XMNL-NLMS for different horizontal source positions. The source $s(x)$ is a colored signal obtained by filtering a white Gaussian noise (WGN) through a lowpass finite impulse response (FIR) filter with coefficients $[0.357, 0.9, 0.357]$ at a sampling rate of $f_s = 11.025$ kHz. The convergence of the algorithm is quantified by the normalized misalignment

$$\eta(n) = \frac{\|\hat{h}(n) - h\|^2}{\|h\|^2},$$

(9)

where $\hat{h}(n) = [\hat{h}_1^T(n) \ \hat{h}_2^T(n)]^T$ and $h = [h_1^T \ h_2^T]^T$ in which length of both $h_i$ and $\hat{h}_i(n)$ are equal to 256. The convergence of NL-NLMS when the source is at $(2.85, 1.85, 1.6)$ m has been included for comparison. Additional tests conducted have shown that the convergence of NL-NLMS for various source positions is comparable to that shown in Fig. 2. A step-size of $\mu = 0.6$ is used for XMNL-NLMS while $\mu = 0.8$ is used for NL-NLMS so that its steady-state normalized misalignment reaches that of XMNL-NLMS when the source is in-front-of microphone pair centroid.

As shown in Fig. 2, the XMNL-NLMS algorithm outperforms that of the full-update NL-NLMS when the source is in-front-of the microphone pair centroid at $(2.85, 1.85, 1.6)$ m. The convergence rate of XMNL-NLMS is however reduced significantly when the source is located away from the microphone pair centroid. In order to gain further insight into this degradation in convergence performance for XMNL-NLMS, we consider both the interchannel coherence as well as the ratio of selected tap-input energy to the total tap-input energy.

3.2. Effect of XM tap-selection on interchannel coherence

We first investigate the effect of XM tap-selection on the interchannel coherence for various source locations. Defining $z_i(n) = Q_i(n)x'_i(n)$ as the XM subselected tap-input vector, the interchannel coherence between $z_1(n)$ and $z_2(n)$ is then defined by

$$C_{z_1 z_2}(f) = \frac{|P_{z_1 z_2}(f)|^2}{|P_{z_1 z_1}(f)P_{z_2 z_2}(f)|}$$

(10)

where $P_{z_i z_j}(f), i, j = 1, 2$ is the cross power spectrum between $z_i(n)$ and $z_j(n)$ while $f$ is the normalized frequency.

Figure 3 (a) shows the mean interchannel coherence across different frequencies for various source positions. As can be seen, when the source is in-front-of the microphone pair centroid at $(2.85, 1.85, 1.6)$ m, the XM tap-selection criterion is utilized efficiently to decorrelate $z_1(n)$ and $z_2(n)$ giving a low interchannel coherence of 0.43. Due to this low interchannel coherence, good convergence performance for XMNL-NLMS is therefore expected.

This interchannel coherence between the XM tap-selected inputs, however, increases with x-position beyond which after a certain
x-offset, the interchannel coherence reduces to approximately 0.44. Since high interchannel coherence will degrade convergence performance, it is therefore expected that the convergence rate of XMLE-NLMS will reduce with increasing x-offset until a point where its performance improves again. On the contrary, however, XMLE-NLMS continues to degrade in convergence performance with increasing x-position as can be seen in Fig. 3. We attempt to gain better insights into this contradictory behaviour by studying the effect of XM tap-selection on the energies of the active tap-inputs.

3.3 Effect of XM tap-selection on tap-input energies

To illustrate how XM tap-selection affects the tap-input energies, we employ the $M$-ratio criterion [5] which computes the energies of the selected tap-inputs with respect to the tap-input vector, i.e.,

$$M = \|Q(n)x'(n)\|_2^2/\|x'(n)\|_2^2,$$  

where $Q(n) = \text{diag}\{q_i^2(n) q_j^2(n)\}$ with elements defined by (6)-(8). Figure 3 (b) illustrates how $M$ varies with the position of a WGN source for both NL-NLMS and XMLE-NLMS. As can be seen, the NL-NLMS algorithm has $M = 1$ for all source positions since tap-inputs are not sub-selected. On the other hand, for XMLE-NLMS, $M < 1$ and increases with x-position of the source.

We can now see why XMLE-NLMS degrades in convergence performance when the source is in-front of a microphone. This is because although the interchannel coherence is relatively low, $M$ is not sufficiently high to reduce any degradation in convergence due to tap-selection. As a result of this suboptimal constraint between the need to reduce coherence and maximization of tap-input energies, a reduction in convergence rate of XMLE-NLMS is expected as shown in Fig. 2. However, when the source is in-front of the microphone array centroid at $(2.85, 1.85, 1.6)$ m, the modest degradation in convergence performance due to a reduction in $M$ is offset by a significant improvement in convergence performance due to the reduction in interchannel coherence.

As an additional note, the source position affects not only the convergence rate of XMLE-NLMS, but also its steady-state normalized misalignment. As can be seen from Fig. 2, the steady-state normalized misalignment is higher than that of NL-NLMS for increasing x-position since the weight update is performed using a subset of tap-inputs. This causes an error in the weight update resulting in an increase in the steady-state normalized misalignment.

4. A CENTER-CLIPPING APPROACH TO SAEC

We propose to improve the convergence and steady-state performance of XMLE-NLMS by employing a center clipping algorithm. The clipping procedure aims to increase the energies of the “inactive” taps when the source is away from the microphone pair centroid so as to compensate for the energy loss due to XM tap-selection. The weight update equation of the proposed clipped XMLE-NLMS (cXMLE-NLMS) algorithm is given by

$$\hat{h}(n+1) = \hat{h}(n) + \frac{\mu}{\|\hat{x}'(n)\|^2 + \epsilon} e(n)\hat{x}'(n),$$  

where $\hat{x}'(n) = [\hat{x}_1'^T(n) \hat{x}_2'^T(n)]^T$ such that

$$\hat{x}_i'(n) = Q_i(n)x_i'(n) + \tilde{Q}_i(n)\tilde{x}_i(n).$$  

In this equation, $Q_i(n)$ is the XM tap-selection matrix defined by (6)-(8) and $\tilde{Q}_i(n) = I_{L \times L} - Q_i(n)$ is the $L \times L$ matrix that is used for identifying the tap-input elements not selected by the XM tap-selection procedure.

The $L \times 1$ vector $\hat{x}'(n) = [\hat{x}_1'(n), \ldots, \hat{x}_L'(n - L + 1)]^T$ has elements that are clipped versions of elements in $x_i'(n)$ such that

$$\hat{x}_i'(n - k) = \begin{cases} x_i'(n - k) - \gamma_i(n) & x_i'(n - k) > \gamma_i(n) \\ x_i'(n - k) + \gamma_i(n) & x_i'(n - k) < -\gamma_i(n) \\ 0 & \text{otherwise} \end{cases},$$  

where $0 \le \gamma_i(n) \le \gamma_{i,\text{max}}(n)$ is a clipping threshold which controls the amount of clipping for $x_i(n)$ given that

$$\gamma_{i,\text{max}}(n) = \max_{x_i(n)}|\hat{x}_i'(n)|.$$  

As can be seen from (14), when $\gamma_i(n) = 0$, we have $\hat{x}_i'(n - k) = x_i'(n - k)$ which results in $\tilde{x}_i'(n) = x_i'(n)$ in (13) and hence, the performance of cXMLE-NLMS is equivalent to that of NL-NLMS. On the other hand, when $\gamma_i(n) = \gamma_{i,\text{max}}(n)$, we have $\tilde{x}_i'(n) = 0$. As a consequence, $\tilde{x}_i'(n) = Q_i(n)x_i'(n)$ and the proposed cXMLE-NLMS algorithm is equivalent to XMLE-NLMS. Therefore, a high value of $\gamma_i(n)$ is desirable when the source is in-front of the microphone pair centroid while a low value of $\gamma_i(n)$ is desirable when the source is in-front of a microphone.

In order for the algorithm to be robust to the source position, we compute $\gamma_i(n)$ adaptively for each channel. We propose to estimate the source position based on the similarity between the energies of $x_1'(n)$ and $x_2'(n)$. This measure should reflect how close the source is to one of the microphones and we propose to employ

$$\delta(n) = \frac{\|\text{mean}\{x_1'(n)\} - \text{mean}\{x_2'(n)\}\|}{\|\text{mean}\{x_1'(n)\}\| + \|\text{mean}\{x_2'(n)\}\|},$$  

to compute this similarity where $\text{mean}\{x_1'(n)\} = \frac{1}{L-1} \sum_{j=0}^{L-1} x_1'(n - j)$. Hence, when the source is in-front of the array centroid, the mean energies of $x_1'(n)$ and $x_2'(n)$ are comparable giving a low $\delta(n)$. To avoid the effects of instantaneous changes of $\delta(n)$ on $\gamma_i(n)$, we further employ a smoothed version of $\delta(n)$ given by

$$\tilde{\delta}(n) = \rho \tilde{\delta}(n - 1) + (1 - \rho)\delta(n),$$  

where $0 \le \rho \le 1$ and $\tilde{\delta}(0) = \delta(0)$. Since we expect that the value of $\delta(n)$ is small when the source is close to microphone pair centroid whereas $\tilde{\delta}(n)$ is large when the source is near to one of the microphones, the clipping threshold $\gamma_i(n)$ should be made a reducing function of $\delta(n)$ in a piecewise linear manner for simplicity as shown in Fig. 4 where, for speech signals, $\beta_L = 0.1$ and $\beta_H = 0.4$ are constants that are determined empirically.

Finally, it is important to note that when $\gamma_i(n) > 0$, we need to reduce the additional steady-state normalized misalignment that results from the error in weight update introduced by tap-selection.
process. Hence we propose to reduce the clipping threshold \( \gamma_i(n) \) to zero after convergence of the mean-square error (MSE). In order to estimate the convergence of the algorithm, we employ the following recursive relation for the estimation of MSE error \([8]\)

\[
\varepsilon(n) = \theta \varepsilon(n-1) + (1-\theta) e^2(n),
\]

where \( \theta \) is a time constant of the averaging process. Hence when the \( \varepsilon(n) \) reaches below a lower limit, the clipping threshold \( \gamma_i(n) \) will be reduced to zero. This procedure improves the steady-state misalignment performance of cXMNL-NLMS. In view of this and according to Fig. 4 we propose, by defining \( \nu \) as a lower limit,

\[
\gamma_i(n) = \begin{cases} 
\gamma_{i,\text{max}}(n), & \delta_i(n) < \beta_L, \varepsilon(n) > \nu \\
\lambda_i(n) - \beta_H, & \beta_L \leq \delta_i(n) < \beta_H, \varepsilon(n) > \nu \\
0, & \text{otherwise}
\end{cases}
\]

In order to illustrate the effectiveness of the proposed algorithm, we compute, similar to (11), \( M\varepsilon = \|x(n)\|^2 / \|x'(n)\|^2 \) for the case of cXMNL-NLMS where the subscript \( c \) in \( M\varepsilon \) denotes for the center-clipped signals. As can be seen in Fig. 3 (b), when the source moves away from the centroid of microphone pair, the loss of energy is now compensated by (19) and as a consequence convergence rate is improved over XMNL-NLMS.

5. SIMULATION RESULTS

For evaluation purpose, we make use of the experimental setup of the transmission room described in Section 3. The positions of the microphone and loudspeakers in the receiving room are \{3,2,1,5\} m for microphone and \{2.85,1.8,1.6\} m and \{2.4,1.1,1.7\} m for loudspeakers. The receiving room impulse responses are generated synthetically using the method of images \([7]\) and they are of length \( L = 800 \) and \( L_s = 800 \). A WGN signal \( w(n) \) is added to the microphone signal of the receiving room \( y(n) \) to achieve an SNR=30 dB.

Since the steady-state normalized misalignment for XMNL-NLMS varies with the position of the source, we chose the step-size of each algorithm so that their steady-state performance will be equal when the position of the source is in front-of the centroid of the microphone array. This corresponds to \( \mu = 0.8 \) for both NL-NLMS and cXMNL-NLMS and \( \mu = 0.6 \) for XMNL-NLMS. The value of \( \theta \) in (18) was set equal to \( 2L \) and \( \nu \) in (19) was set equal to -60 dB. The input signals \( x_i(n) \), \( i = 1,2 \) are achieved using a colored source signal \( s(n) \) with \( f_s = 16 \) kHz.

The normalized misalignment curves are plotted when the source is at \{2.9,1.85,1.6\} m and \{3,1.85,1.6\} m as shown respectively in the left and right panels of Fig. 5. In addition, Fig. 6 shows an example of speech signal when the position of the source in the transmission room is at \{2.9,1.85,1.6\} m. As can be seen from these results, cXMNL-NLMS achieves the highest rate of convergence of typically 4 dB and steady-state normalized misalignment over existing algorithms for source positions considered in the transmission room. Therefore the overall joint result is a fast converging cXMNL-NLMS that is robust to the position of the source in the transmission room.

6. CONCLUSION

We presented a new center-clipping algorithm for improving the convergence behavior of adaptive filters in SAEC. This approach improves the robustness of XMNL-NLMS against the location of the source relative to microphones in the transmission room. This is achieved by employing a clipping algorithm to increase the energies of the unselected tap-inputs. Simulations performed showed that the proposed algorithm achieves typically 4 dB in convergence performance over XMNL-NLMS.

7. REFERENCES


