SELECTIVE TIME-REVERSAL BLOCK SOLUTION TO THE STEREOPHONIC ACOUSTIC ECHO CANCELLATION PROBLEM

Dinh-Quy Nguyen, *Woon-Seng Gan, and Andy W. H. Khong*

Digital Signal Processing Laboratory, Nanyang Technological University, Singapore

Email: {n060075, ewsgan, andykhong}@ntu.edu.sg

ABSTRACT

Stereophonic acoustic echo cancellation (SAEC) plays an important role in delivering realistic teleconferencing experience. However, the problem of stereophonic acoustic echo cancellation is challenging due to the requirement of uniquely identifying two acoustic paths. In this paper, we present a novel method of selective time-reversal block transformation that significantly reduces the misalignment without noticeably affecting the audio quality. The proposed method employs a magnitude detector so that input blocks of one channel with average magnitude less than a specified threshold are time-reversed in order to decorrelate the other channel. Simulation results show that the proposed method achieves higher convergence rate, better spatial information with less audio distortion compared to the well-known half-wave rectifier method.

Index Terms — *Decorrelation method, magnitude detector, nonlinear transformation, stereophonic acoustic echo cancellation, time-reversal.*

1. INTRODUCTION

Stereophonic acoustic echo cancellation (SAEC) enhances spatial information and provides a more immersive experience in teleconferencing. However, for a realistic SAEC system, where impulse response of the transmission room is longer than that of the adaptive filters, the adaptive filters suffer from poor misalignment. This poor misalignment is due to high interchannel coherence of the transmitted signals [1]. Since the analysis of this problem has been published in [1], several algorithms have been proposed to decorrelate the transmitted input signals so as to achieve good convergence performance for the adaptive filters. It is important to realize, however, that any such pre-processing should not degrade the audio quality and/or adversely affect the stereophonic image. Several decorrelation techniques have since been developed to address this misalignment problem. This includes the use of random noise addition to the loudspeaker signals [2], nonlinear transformation of the loudspeaker signals [1], audio coding [3], new configuration with nonlinear pre-processing [4], first-order time-varying allpass filters [5], spectral-shaped random noise [6], Autoregressive (AR) analysis [7] and adaptive noise addition [8]. Among these methods, the nonlinear transformation provides an effective approach to achieve interchannel decorrelation and results in good convergence performance. The nonlinear transformation was investigated using different types of nonlinearities [9] and it has been shown that the half-wave rectifier (HWR) achieves a good tradeoff in terms of stereo quality as well as convergence rate. It is noted however, that the stereo perception of the signals are somewhat degraded especially for musical signals.

In this paper, we propose the use of selective time-reversal block transformation to achieve a higher convergence rate without noticeably affecting the audio quality. This technique employs the magnitude detector to select input blocks with small magnitude in order to perform time-reversal transformation. The motivation of selecting such blocks is that by doing so, it does not significantly degrade the stereo audio quality and spatial information of the SAEC system.

The paper is organized as follows. Section 2 describes the SAEC problem. The proposed selective time-reversal block solution is presented in Section 3. Following that, simulation results and discussions are presented in Section 4 while Section 5 concludes the paper.

2. THE SAEC PROBLEM

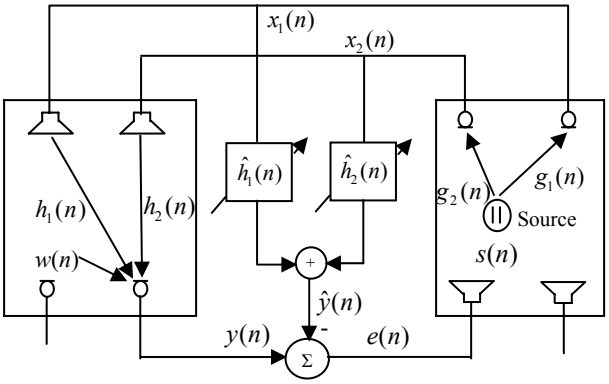

Figure 1 - Schematic diagram of SAEC system.

Figure 1 shows an SAEC system where two microphones in the transmission room pick up speech signals from a source $s(n)$ through two acoustic impulse responses

$\mathbf{g}_{i,M} = \left[g_{i,0}\ g_{i,1}\ \ldots\ g_{i,M-1} \right]^T$ where i=1,2 and M is defined as the length of $\mathbf{g}_i$. The stereo signals $x_i(n)$, are transmitted to the receiving room which are in turn picked up by both microphones via another set of acoustic echo paths $\mathbf{h}_{i,L} = \left[h_{i,0}\ h_{i,1}\ \ldots\ h_{i,L-1} \right]^T$, where L is defined as the length of $\mathbf{h}_i$. Similar to that of [1], we show the problem of SAEC for only one microphone signal.

A pair of finite impulse response (FIR) adaptive filters $\hat{\mathbf{h}}_{i,L}(n) = \left[\hat{h}_{i,0}(n)\ \hat{h}_{i,1}(n)\ \ldots\ \hat{h}_{i,L-1}(n) \right]^T$ are used to identify the unknown acoustic echo paths $\mathbf{h}_{i,L}(n)$ in the receiving room and the output of these adaptive filters are given by

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

where $\mathbf{x}_{i,L}(n) = \left[x_{i,0}(n), x_{i,1}(n) \ldots, x_{i,L-1}(n) \right]^T$ is the tap-input vector of length L. The microphone signal in the receiving room is then given by

$$y(n) = \mathbf{h}_1^T(n)\mathbf{x}_1(n) + \mathbf{h}_2^T(n)\mathbf{x}_2(n) + w(n)\ , \qquad (2)$$

where $w(n)$ is defined as the background noise. Employing (1) and (2), the acoustic echo is then given by

$$e(n) = \left[\mathbf{h}_1^T(n) - \hat{\mathbf{h}}_1^T(n) \right]\mathbf{x}_1(n) + \left[\mathbf{h}_2^T(n) - \hat{\mathbf{h}}_2^T(n) \right]\mathbf{x}_2(n) + w(n)\ . \qquad (3)$$

It has been shown and described comprehensively in [1] that, for a realistic SAEC system where $L<M$, a unique solution exists. However, due to the high interchannel coherence between $\mathbf{x}_1(n)$ and $\mathbf{x}_2(n)$, the convergence rate of the adaptive filters is reduced significantly. Thus, it is important to understand that in order to achieve high rate of convergence, the coherence between $\mathbf{x}_1(n)$ and $\mathbf{x}_2(n)$ must be reduced, besides any processing of $\mathbf{x}_1(n)$ and $\mathbf{x}_2(n)$ should not degrade the audio quality and/or adversely affect the stereophonic image.

3. THE PROPOSED SELECTIVE TIME-REVERSAL BLOCK (STRB) ALGORITHM

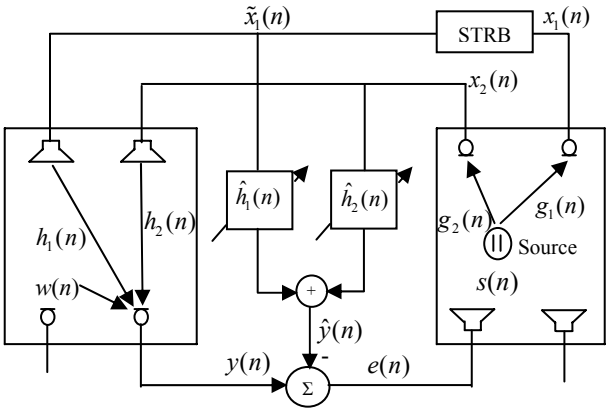

Figure 2 - Schematic diagram of SAEC with STRB transformation.

Time-reversal signal processing is a technique use to reverse a given process or a signal in the time domain. It has been applied widely for sound focusing applications [10]. In this paper, we propose to apply time-reversal for the SAEC problem. The motivation behind this approach is that by time-reversing signal of only one channel, the interchannel coherence can be reduced. Due to the similarity among two channels, we can choose any channel to apply time-reversal technique. In this paper, channel 1 is chosen to perform time-reversal. However, time-reversal technique should be applied with caution since time-reversal process will distort the stereo-image and degrade the speech quality. From this view, we propose a selective time-reversal block method that only selects and time-reverses input blocks with average magnitude less than a specified threshold.

It is also useful to note that time-reversal technique is easy and simple to implement in real-time processing. The time-reversal operation can be easily achieved using circular buffer that is commonly provided on DSP processors [11].

3.1 Interchannel coherence reduction through time-reversal

Assume that the system (transmission room) is linear and time invariant; therefore, the linear relation among stereo signals in SAEC system [1] is

$$\mathbf{x}_{1,M}^T(n)\mathbf{g}_{2,M}(n) = \mathbf{x}_{2,M}^T(n)\mathbf{g}_{1,M}(n)\ , \qquad (4)$$

where $\mathbf{x}_{i,M} = \left[x_i(n)\ x_i(n\text{-}1)\ \ldots\ x_i(n-M+1) \right]^T$, i=1,2.

Thus, if we apply time-reversal transformation in signal of channel 1 in SAEC system, we have:

$$\bar{\mathbf{x}}_{1,M}^T(n)\mathbf{g}_{2,M}(n) = \mathbf{x}_{2,M}^T(n)\mathbf{g}_{1,M}(n)\ . \qquad (5)$$

where $\bar{\mathbf{x}}_{1,M} = \left[x_1(n-M+1)\ \ldots\ x_1(n\text{-}1)\ x_1(n) \right]^T$.

There is a linear relation between $\bar{\mathbf{x}}_{1,M}$ and $\mathbf{x}_{2,M}$ in (5) if and only if:

$$\bar{\mathbf{x}}_{1,M}(n) = \mathbf{x}_{1,M}(n)\ . \qquad (6)$$

This can happen if the signal chosen to time-reverse is symmetric in time domain. However, in practice this case never occurs because all realistic signals (speech, music or noise) are random signals. Thus, time-reversal transformation in one channel can reduce much interchannel coherence.

Continuously, to check the interchannel coherence in SAEC system, we apply the proposed STRB method in two channels. In this example, we compute the interchannel coherence of $\mathbf{x}_1(n)$ and $\mathbf{x}_2(n)$ with block length L=512. We generate the impulse responses in the transmission room by using the method of images [12] with the source at {2.2, 1.5, 1.6} m while the microphones are placed at {1.5, 2.5, 1.6} m and {2.5, 2.5, 1.6} m. We can see from this example that the interchannel coherence magnitude of the proposed selective time-reversal block (STRB) algorithm is smaller than that of the original signal. More importantly, the proposed STRB method achieves smaller coherence magnitude than half-wave rectifier (HWR) method with α=0.5 [1] across most frequency bins and especially for lower frequencies. The average reduction in interchannel

coherence across all frequencies for the STRB method over the HWR method is 0.28 dB, as shown in Fig. 3.

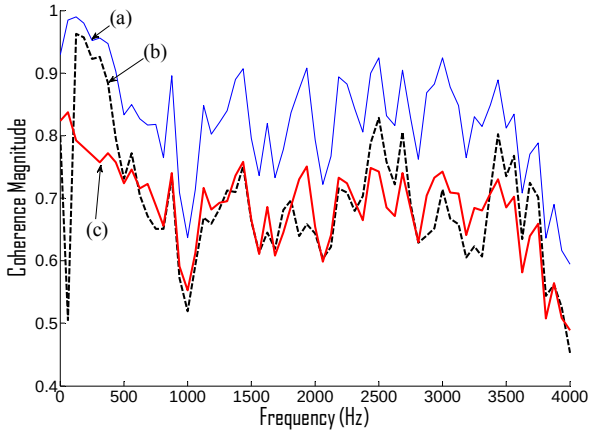

Figure 3 - The interchannel coherence plot for (a) no decorrelation, (b) HWR transformation, and (c) STRB transformation.

3.2 Reducing audio distortion using magnitude threshold

It is expected that the proposed method of time-reversed transformation will greatly degrade the audio quality of the transmitted signals $\mathbf{x}_1(n)$ and $\mathbf{x}_2(n)$ if this process is applied to most input signal blocks. In order to address this problem, we propose to select and process input blocks having magnitudes less than a specified threshold ε for performing time-reversal. By processing such blocks with small magnitude computed using a magnitude detector; we reduce the distortion introduced by our proposed method since these blocks are relatively inaudible.

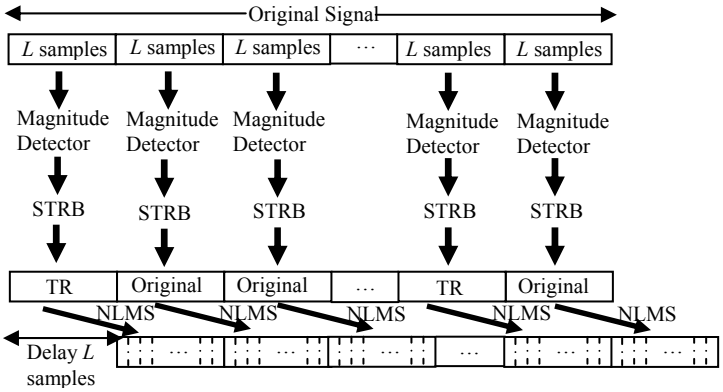

Figure 4 - Illustration of STRB-NLMS algorithm.

We first define an input block of the first channel by

$$\mathbf{x}_1(m) = \left[x_1(mL)\; x_1(mL-1)\; \ldots\; x_1(mL-L+1)\right]^T, \qquad (7)$$

where m is defined as the block index. The time reversed input signal of the m^{th} block in channel 1 is then defined as

$$\overline{\mathbf{x}}_1(m) = \left[x_1(mL-L+1)\; \ldots\; x_1(mL-1)\; x_1(mL)\right]^T. \qquad (8)$$

Besides, the mean absolute magnitude of each block in channel 1 is also defined by

$$k(m) = \frac{1}{L}\sum_{i=0}^{L-1}\left|x_1(mL-i)\right|. \qquad (9)$$

In the second step of the STRB process, the mean absolute magnitude $k(m)$ is then compared with a specified magnitude threshold ε to perform time-reversal given by

$$\tilde{\mathbf{x}}_1(m) = \begin{cases} \overline{\mathbf{x}}_1(m), & \text{if } k(m) < \varepsilon \\ \mathbf{x}_1(m), & \text{if } k(m) \geq \varepsilon \end{cases}. \qquad (10)$$

It is important to realize that this specified magnitude threshold serves as a tradeoff between audio quality and convergence performance (brought about by the reduction in interchannel coherence).

A final step of the STRB algorithm is to apply a delayed version of the normalized least-mean-square (NLMS) algorithm to the stereo channels. This delay corresponds to L samples, as shown in Fig. 4. We introduce this delay so as to process the input signals in blocks to perform magnitude detector and time-reversal. Defining δ as the regularization parameter, the proposed STRB-NLMS algorithm is listed in Table I.

TABLE I
STRB-NLMS ALGORITHM.

$$k(m) = \frac{1}{L}\sum_{i=0}^{L-1}\left|x_1(mL-i)\right|$$

$$\tilde{\mathbf{x}}_1(m) = \begin{cases} \overline{\mathbf{x}}_1(m), & \text{if } k(m) < \varepsilon \\ \mathbf{x}_1(m), & \text{if } k(m) \geq \varepsilon \end{cases}$$

$$\tilde{\mathbf{x}}(m) = \left[\tilde{\mathbf{x}}_1^T(m)\; \mathbf{x}_2^T(m)\right]^T$$

$$\hat{\mathbf{h}}(m) = \left[\hat{\mathbf{h}}_1^T(m)\; \hat{\mathbf{h}}_2^T(m)\right]^T$$

$$y(n-L) = \hat{\mathbf{h}}^T(n)\tilde{\mathbf{x}}(n-L) \quad \text{where } n>L$$

$$e(n) = d(n-L) - y(n-L)$$

$$\mathbf{h}(n+1) = \mathbf{h}(n) + \mu\frac{\tilde{\mathbf{x}}(n-L)e(n)}{\left\|\tilde{\mathbf{x}}(n-L)\right\|^2 + \delta}$$

4. SIMULATIONS AND DISCUSSIONS

In this simulation, a male speech with duration of about 10 s is used to verify the effectiveness of the proposed STRB technique compared to that of the HWR method for an SAEC system. The two microphone signals in the transmission room are obtained by convolving the speech with two impulse responses each of 512 points in length while the receiving room impulse responses are also 512 points in length. All impulse responses are generated using the method of images [12] with the source at {2.2, 1.5, 1.6} m while the microphones are placed at {1.5, 2.5, 1.6} m and {2.5, 2.5, 1.6} m in the transmission room. A sampling frequency of 8 kHz and the two-channel NLMS algorithm with a fixed step size μ=0.5 are used throughout the simulation. The reverberation time (T60) is 0.064 s.

The proposed STRB algorithm uses a block length of 512 samples (or 64 ms), and a magnitude threshold of ε=0.03. The STRB is benchmarked against the HWR with α=0.5 in the SAEC simulation. The performances of algorithms are evaluated by objective distortion measures and convergence rate of the normalized misalignment.

4.1 Objective distortion measures

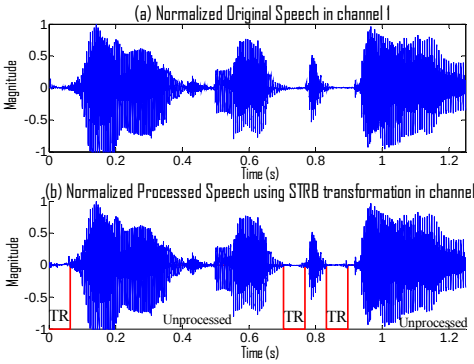

Figure 5 - Original speech and STRB speech in channel 1.

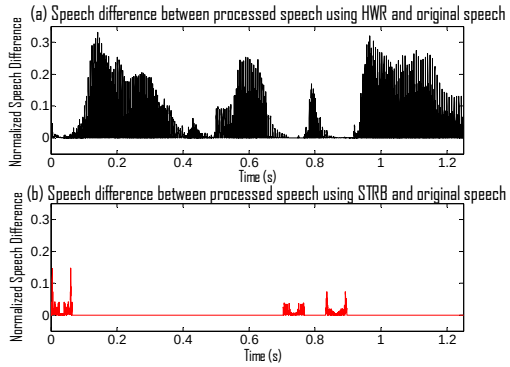

Figure 6 - Comparison of speech difference between HWR and STRB.

We first compare the speech difference between the processed and the original speeches. For clarity of presentation, Fig. 5(a) shows a segment of the speech sequence received from the first microphone. We can see that the processed speech using our proposed STRB, shown in Fig. 5(b), is different with the original speech in only a few segments with small magnitudes indicated by the rectangular blocks. On the other hand, the HWR transformation, when applied to channel 1, adds a value proportional to the magnitude of the received signal for decorrelation. This value is controlled by the variable α which is normally set to $0<\alpha\leq 0.5$ as discussed in [1]. Figure 6(a) shows, for channel 1, the difference between the original and the HWR processed speech signal while Fig. 6(b) shows such differences for the proposed STRB processed speech signals. By comparing Fig. 6(a) and Fig. 6(b), it can be seen that the STRB processor adds less distortion to the speech signal than the HWR.

In order to quantify the effect of the speech difference caused by the proposed STRB decorrelation technique, we define the peak signal-to-difference ratio (PSDR) as

$$\mathrm{PSDR} = 20\log_{10}(1/\sigma), \qquad (11)$$

where σ is a mean speech difference between the processed speech and original speech. Using the STRB method, the mean speech difference σ is always smaller than the chosen magnitude threshold ε, thus we can relate the PSDR of the STRB to the magnitude threshold as:

$$\mathrm{PSDR} = 20\log_{10}(1/\sigma) \geq 20\log_{10}(1/\varepsilon), \qquad (12)$$

With ε=0.03, the minimum PSDR of the proposed STRB method is greater than 30 dB and hence it can sustain good human audio perception due to masking principles [13].

In order to further verify that the proposed STRB method introduces less distortion compared to the HWR method, we employ the Bark Spectral Distortion (BSD) measure whereby a smaller value corresponds to a smaller distortion [14]. The BSD takes into account auditory frequency warping, critical band integration, amplitude sensitivity variations with frequency, and subjective loudness. Thus, the BSD measurement offers a more consistent assessment of the effect of incremental changes in the parameter of a speech coder than informal listening test. The mean BSD of HWR and STRB transformations are found to be 0.0355 and 0.0043 respectively using α=0.5 for the HWR [1] and ε=0.03 for the STRB. Thus, the STRB method shows an improvement of 87.89% in terms of BSD over the HWR method.

4.2 Convergence rate of misalignment

We evaluate the performance of the proposed STRB algorithm in term of normalized misalignment defined by

$$\eta(n) = \frac{\left\|\mathbf{h}-\hat{\mathbf{h}}(n)\right\|^2}{\left\|\mathbf{h}\right\|^2}. \qquad (13)$$

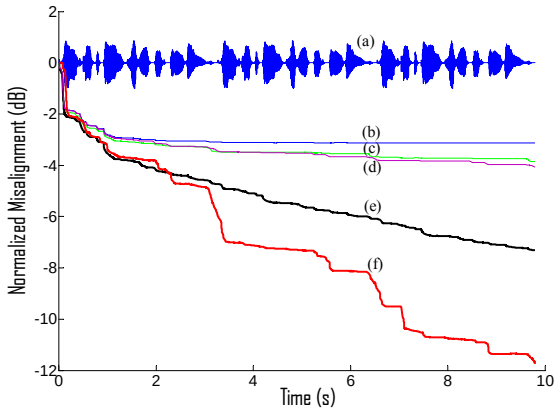

Figure 7 - (a) Speech and misalignment plot for (b) no decorrelation, (c) Random noise addition (RNA) on selected blocks, (d) HWR on selected blocks, (e) HWR and (f) STRB.

Based on the two-channel NLMS algorithm and a male speech, we compare the misalignment performances of STRB transformation with magnitude threshold of ε=0.03, as shown in Fig. 7(e) against that using the HWR transformation as shown in Fig. 7(d). As can be seen from this result, the proposed STRB-NLMS algorithm can achieve about 4.2 dB misalignment reduction during initial convergence compared with the HWR method [1].

We also compare the performance of the proposed STRB algorithm with one that employs the HWR transformation on selected input blocks. In both algorithms, we selected magnitudes with ε=0.03 as the threshold. Normalized misalignment performance for this case of HWR algorithm is plotted as shown in Fig. 7(d). As seen in Fig. 7, the STRB method achieves 7.7 dB improvement in

terms of normalized misalignment compared to that using the HWR method on selected blocks.

Besides, the STRB method is compared with the random noise addition (RNA) method in SAEC system [2], [8]. The normalized misalignment in the STRB method is higher than 7.9 dB compared to that in the RNA method on selected blocks, as shown in Fig. 7(f) and Fig. 7(c). In PSDR measurement, the STRB method also achieves higher 28.2 dB than the RNA method on selected blocks.

Moreover, we also evaluate the performance of the proposed method in different chosen thresholds (ε=0.01, 0.02 and 0.03 corresponding to PSDR=40 dB, 34 dB and 30 dB). As shown in Fig. 8, the smaller threshold is used, the slower misalignment convergence is achieved. However, the performance of STRB method in smaller threshold (ε=0.01) still results in better convergence and less audible distortion compared to the HWR method. Thus, we can adjust the threshold in STRB method to achieve desired misalignment convergence and/or stereo audio quality.

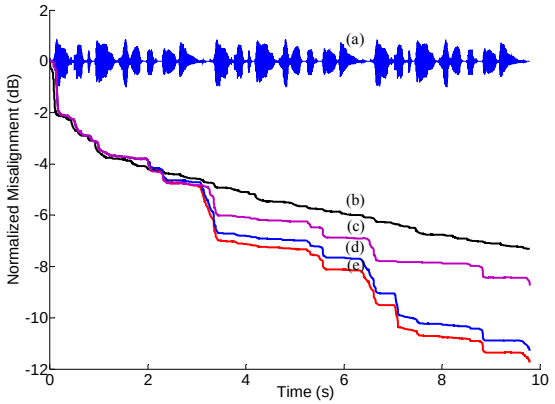

Figure 8 - (a) Speech and misalignment plot for (b) HWR with α=0.5, (c) STRB with ε=0.01, (d) STRB with ε=0.02 and (e) STRB with ε=0.03.

4.3 Spatial information

Besides lesser audible distortion and better misalignment convergence, the STRB method also maintains magnitude similarity in one channel (channel 2) and minimizes signal magnitude variation in the other channel (channel 1). Thus, the STRB method retains better spatial information compared to the conventional HWR method in the SAEC system. Hence, the proposed STRB method is an efficient solution to achieve good misalignment performance without distorting the original stereo image in the SAEC system. In summary, the comparison of all decorrelation techniques mentioned in this paper is illustrated in Table II.

TABLE II
COMPARISONS AMONG DECORELATION TECHNIQUES IN SAEC SYSTEM

Types	PSDR	BSD	Misalignment Reduction*
RNA on selected block	17.6 dB	3.91×10^{-4}	0.6 dB
HWR on selected block	26.9 dB	4.17×10^{-4}	0.8 dB
HWR with α=0.5	33.6 dB	355×10^{-4}	4.3 dB
STRB with ε=0.03	45.8 dB	43×10^{-4}	8.5 dB

*The misalignment reductions are evaluated by comparing misalignment differences between decorrelation techniques and original.

5. CONCLUSIONS

In this paper, the STRB transformation is proposed to mitigate the misalignment problem in SAEC. The STRB method employs a magnitude detector so that input blocks of one channel with average magnitude less than the specified threshold are time-reversed in order to decorrelate with other channel. The motivation of this method is that the stereo signals are random signals with many small magnitude portions that are normally inaudible. Thus, the STRB method can be applied in these portions to reduce interchannel coherence and significantly improve in convergence rate of misalignment compared with conventional HWR method in SAEC system.

Besides, as shown in the objective distortion measurements (BSD and PSDR), the proposed method also introduces less audible distortion and better spatial information than the conventional HWR and random noise addition approaches. Hence, the proposed STRB method can overcome the present technical challenges of SAEC and provide a suitable and cost effective solution for multi-channel teleconferencing applications.

6. REFERENCES

[1] 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. on Speech and Audio Processing*, vol. 6, no. 2, pp. 156-165, Mar. 1998.

[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] T. Gansler and P. Eneroth, "Influence of audio coding on stereophonic acoustic echo cancellation," *Proc. ICASSP 1998*, Seattle, USA, May 1998, pp. 3649-3652.

[4] S. Shimauchi, Y. Haneda, S. Makino, and Y. Keneda, "New configuration for a stereo echo canceller with nonlinear pre-processing," *Proc. ICASSP 1998*, Seattle, USA, May 1998, pp. 3685-3688.

[5] M. Ali, "Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation", *Proc. ICASSP 1998*, Seattle, USA, May 1998, pp. 3689-3692.

[6] A. Gilloire and V. Turbin, "Using auditory properties to improve the behaviour of stereophonic acoustic echo cancellers", *Proc. ICASSP 1998*, Seattle, USA, May 1998, pp. 3681-3684.

[7] Y. W. Jung, J. H. Lee, Y. C. Park and D. H. Youn, "A new adaptive algorithm for stereophonic acoustic echo canceller", *Proc. ICASSP 2000*, Istanbul, Turkey, June 2000, pp. 801-804.

[8] P. Surin, N. Tangsangiumvisai, and S. Aramvith, "An adaptive noise decorrelation technique for stereophonic acoustic echo cancellation", *Proc. TENCO*, Nov. 2004, pp. 112-115.

[9] D. R. Morgan, J. L. Hall, and J. Benesty, "Investigation of several types of Nonlinearity for Use in Stereo Acoustic Echo Cancellation", *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 6, Sep. 2001.

[10] S. Yon, M. Tanter, and M. Fink, "Sound focusing in rooms: The time reversal approach," *J. Acoust. Soc. Amer.*, vol. 113, pp. 1533–1543, 2003.

[11] H. D. Hendrix, "Implementing circular buffers with bit-reversed addressing", Digital Signal Processing Solution, Texas Instruments, 1997.

[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol.65, pp.943–950, Apr.1979.

[13] B.C. J. Moore, *An Introduction to the Psychology of Hearing*. New York: Academic, 1989, ch. 3.

[14] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE. Journ. on Selected Areas in Comms.*, vol. 10, no. 5, 1992.