A NOVEL ADAPTIVE CROSSTALK CANCELLATION USING PSYCHOACOUSTIC MODEL FOR 3D AUDIO

JunSeong Kim, SangGyun Kim, and Chang D.Yoo

Dept. of EECS, Div. of EE, KAIST
373-1 Guseong-dong, Yuseong-gu, Daejon 305-701, Republic of Korea
s_a_p@kaist.ac.kr, zom@eeinfo.kaist.ac.kr, and cdyoo@ee.kaist.ac.kr

ABSTRACT

In rendering a virtual sound over two loudspeakers, adaptive inverse filtering is required for crosstalk cancellation. Although various adaptive algorithms have been proposed for crosstalk cancellation, few have been effective especially in time-varying environments where fast convergence rate is required. Until now, the least-mean square (LMS) algorithm known for its simplicity and robustness has been the predominant algorithm used, but its convergence rate is considered slow for colored inputs. In this paper, human perceptual characteristics which have never been incorporated in an LMS algorithm is introduced. In our experiment, the proposed algorithm achieved higher perceptual accuracy and faster convergence rate than the conventional LMS algorithm.

Index Terms— crosstalk cancellation, 3-D audio

1. INTRODUCTION

Crosstalk cancellation to adjust the sound level reaching each ear is required, when rendering virtual sound over two loudspeakers. Various crosstalk cancellations have been proposed in the past. Atal and Schroeder [1] first put it into practice using analog crosstalk cancellation algorithm. Damaske [2], and Cooper and Bauck [3] refined it using digital crosstalk cancellation algorithms. Ideally, crosstalk cancellation should find the inverse matrix of the acoustic transfer function (ATF) matrix from each of the speakers to each of the listener's eardrums. The ATF matrix, however, is not guaranteed to be minimum-phase and invertible [4], thus direct inversion may not be possible. An adaptive inversion algorithm can produce an approximation of the inverse matrix of the ATF matrix [5], although a certain amount of training time for convergence is required.

When we listen to music, the listening environment often changes due to the movement of the listener and variation in the listening environment. This causes the ATF matrix to change, and the crosstalk cancellation algorithm has to adapt to the changing environment. For real-time implementation, the predominant algorithm used is the least mean-squared (LMS) algorithm [6]. Although it is relatively simple and accurate, the LMS has slow convergence rate for a colored input.

The step size of the proposed adaptive filter changes according to the masking effect of desired output. The proposed algorithm reduces crosstalk in the perceptual sense. In our experiment, we have found that the proposed algorithm converged faster than the conventional LMS.

This paper is organized as follows. Section 2 describes the conventional crosstalk cancellation using the LMS. Section 3 describes the novel adaptive filtering algorithm based on a psychoacoustically motivated step-size. Section 4 presents experimental results, and Section 5 concludes and summaries the paper.

2. TYPICAL ADAPTIVE CROSSTALK CANCELLATION

A typical listening situation based on two fixed speakers with a general crosstalk cancellation structure is illustrated in Fig. 1. The frequency magnitudes of desired binaural signals are denoted by X_1 and X_2 for sounds reaching the left and right ear, respectively. Here, Y_1 and Y_2 are frequency magnitudes reaching each ear of the listener. This system can be described by the following matrix equation in the frequency domain

$$\mathbf{Y} = \mathbf{CHX} \tag{1}$$

where $\mathbf{Y} = [Y_1, Y_2]^T$ and $\mathbf{X} = [X_1, X_2]^T$, that is,

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \tag{2}$$

where C_{lm} is the (l,m)th element of the ATF matrix **C**, and H_{lm} is the (l,m)th element of the crosstalk cancellation matrix **H** ($l, m = 1, 2$). In order to render the desired signals at each ear of the listener, **H** should be the inverse of **C**. However, the direct inversion of C does not always guarantee the existence of crosstalk cancellation, since the elements do not satisfy the minimum-phase condition [4]. So, the problem of reducing crosstalk is approached using the LMS algorithm. For mathematical simplicity, (2) is rearranged as follows

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} C_{11}X_1 & C_{12}X_1 & C_{11}X_2 & C_{12}X_2 \\ C_{21}X_1 & C_{22}X_1 & C_{21}X_2 & C_{22}X_2 \end{bmatrix} \begin{bmatrix} H_{11} \\ H_{21} \\ H_{12} \\ H_{22} \end{bmatrix}. \tag{3}$$

From (3), $C_{lm}X_i$ is the input to H_{mk}, and Y_i should ideally be a delayed version of X_i. Fig. 2 shows a crosstalk cancellation block diagram using LMS. In the figure, the lowercase variables represent the time domain signal, and d is the system delay. A binaural signal $x_i[n]$ is passed through the filter C_{lm} to generate output $r_{ilm}[n]$, and this output is passed through the adaptive filter h_{mi}.
The signal received at the ear is given by

$$\begin{aligned} y_l[n] =& r_{1l1}[n] * h_{11}[n] + r_{1l2}[n] * h_{21}[n] + r_{2l1}[n] * h_{12}[n] \\ &+ r_{2l2}[n] * h_{22}[n] \end{aligned} \tag{4}$$

for $l = 1, 2$. The LMS adjusts the adaptive filter coefficients by minimizing the cost function

$$J = E[e[n]^2] = E[(d[n] - y[n])^2] \tag{5}$$

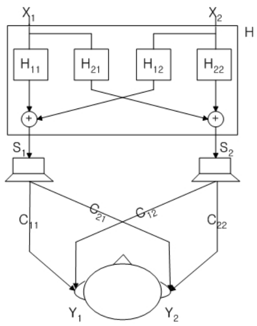


Fig. 1. Speaker-based spatial audio rendering system showing the acoustic transfer functions (**C**) and the structure of crosstalk cancellation (**H**). Here, C_{lm} is the (l,m)th element of the ATF matrix **C**.

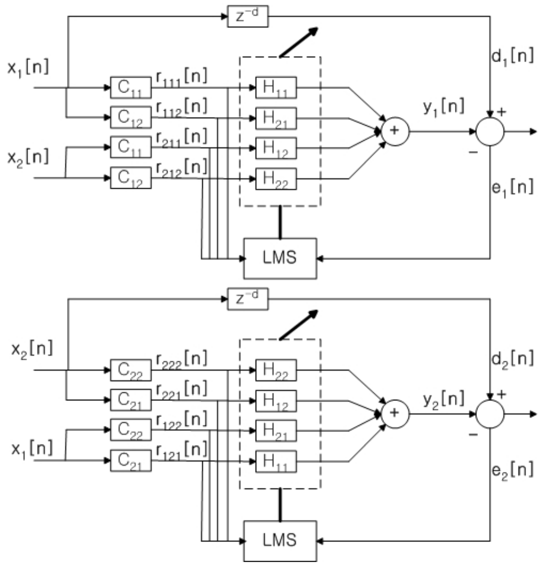


Fig. 2. A crosstalk-cancellation block diagram using LMS.

using the steepest descent method, where $e[n]$, $d[n]$ and $y[n]$ are defined respectively as

$$e[n] = \left[\begin{array}{c} e_1[n] \\ e_2[n] \end{array} \right], d[n] = \left[\begin{array}{c} d_1[n] \\ d_2[n] \end{array} \right], y[n] = \left[\begin{array}{c} y_1[n] \\ y_2[n] \end{array} \right].$$

This adaptive inverse filtering algorithm could be modified to the normalized frequency domain adaptive filter(NFDAF)-LMS algorithm using the overlap-save method [7][8]. The update of filter coefficient of the NFDAF-LMS is established by block-by-block in the frequency domain. Thus the filter coefficient is updated according to each frequency component. The general update form of the NFDAF-LMS for crosstalk cancellation can be expressed as follows

$$H[k+1] = H[k] + 2\mu X^*[k]E[k] \tag{6}$$

where k is block number, $X[k]$ is normalized filter input, and $E[k]$ is error between the desired output and the filtered output.

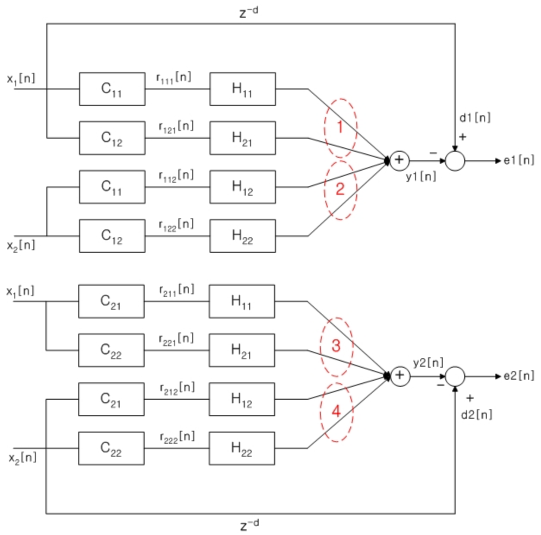


Fig. 3. Diagram of the adaptive LMS filter for crosstalk cancellation where the filtered outputs are divided into four groups as 1, 2, 3, and 4.

3. NOVEL ADAPTIVE FILTERING ALGORITHM BASED ON STEP SIZE VARIATION

The masking threshold is defined as the upper bound below which the human cannot perceive the presence of noise or any other sound. The calculation of the masking threshold is well summarized in the literature [9]. The steps involved in determining the masking threshold are as follows:

1. Critical band analysis: sum up the power spectrum in each critical band (bark), where the power spectrum is obtained by magnitude squaring the FFT coefficient.
2. Spreading: convolve with a spreading function to take into account the effect of adjacent critical bands.
3. Offset: subtract the offset by considering the tone-like or noise-like nature of the speech.
4. Re-normalization: convert the spread spectrum back to bark domain.
5. Absolute threshold: compare with the absolute threshold and choose the maximum between them.

In Fig. 3, the LMS structure is redrawn with the ouput of H_{lm} grouped from 1 to 4 as shown by dashed circles. The outputs combined as group 2 and group 3 using dashed circles are crosstalk components, and the outputs combined as group 1 and group 4 are not crosstalk components. Ideally, the outputs of groups 1 and 4 must converge to the desired outputs $d_1[n]$ and $d_2[n]$, respectively, while the outputs of both groups 2 and 3 must converge to 0. In other words, the adaptive filters H_{11} and H_{21} must work as a pair so that the output of group 1 equals $d_1[n]$ while the output of group 3 equals 0, and the adaptive filters H_{12} and H_{22} must work as a pair so that the output of group 2 equals 0 while the output of group 4 equals $d_2[n]$. These requirements imposed on the pairs are very stringent and can lead to slow convergence rates and large mean-square errors. The motivation behind the proposed algorithm is to alleviate these stringent conditions for faster convergence rate without co3mpromising the perceptual performance.

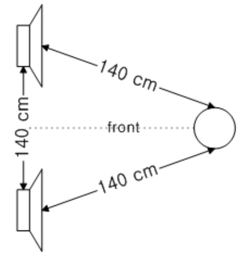


Fig. 4. Geometrical arrangement for crosstalk cancellation experiment.

When the output of group 1 has a large masking threshold in a particular frequency bin, the output of group 2 will be masked by the output of group 1 with high probability. Since the desired output of group 2 is 0, the error of group 2 is itself and is generally small compared to the output of group 1. Therefore, when the masking threshold of the output of group 1 is large in the frequency bin, the error will be masked with high probability and may not affect the listener in the perceptual sense. Under such situation, the proposed algorithm allows the adaptation of H_{12} and H_{22}, which must work as a pair so that the output of group 2 equals 0 while the output of group 4 equals $d_2[n]$, to focus more on making the sum of the output of group 4 close to $d_2[n]$ and less on making the sum of the output of group 2 to be 0 in the frequency bin for allowing unperceived errors in the output of group 2. This is done by placing a larger weight on $e_2[n]$ than on $e_1[n]$. When the masking threshold of the output of group 1 is relatively small for a particular frequency bin, equal weights are placed on both error terms as in the conventional algorithm. Similar argument can be made for the adaptation for the pair H_{11} and H_{21}.

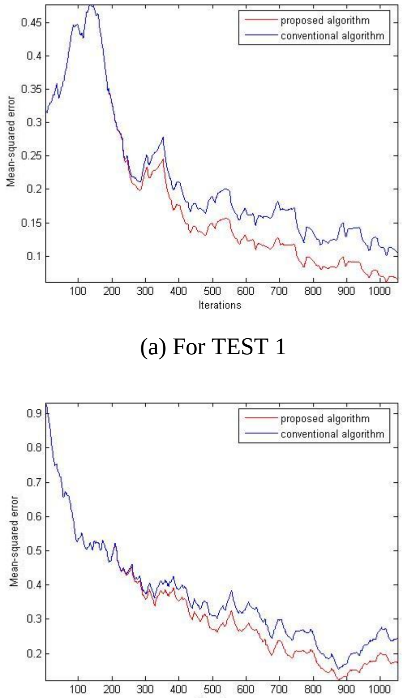
When soft stereophonic sounds are used as input, the following problem occurs. Due to the low input level, the masking threshold of the non-crosstalk components are also low. Because the step-size variation depends on the level of the masking threshold, the proposed algorithm will not perform better than the conventional algorithm. Thus, the proposed algorithm follows the conventional algorithm when the input is small.

4. EXPERIMENTAL RESULTS

Experiment is conducted with the geometrical arrangement of the loudspeakers and the dummy-head as shown in Fig. 4. First, stereophonic sound was played over the two speakers and recorded using the microphone at each ear of the dummy head. Second, using both the stereophonic input and the recorded output, adaptive filtering experiments were conducted. Measured head-related transfer functions (HRTFs) were used.

As test inputs, two songs from each of the 7 genres (classical, country, dance, jazz, newage, R&B, and rock) were used. Two tests were conducted. In Test 1, $x_1[n]$ and $x_2[n]$ are two different songs of different genre. In Test 2, they are stereophonic sound of the same song.

The convergence performance measured as the mean-square error (MSE) as a function of number of iterations are obtained for non-crosstalk components. Fig. 5 shows the MSE between the filtered output of one non-crosstalk component and the desired output for TEST 1 and TEST 2. The results show that the proposed algorithm is more accurate and has faster convergence rate than the conventional algorithm. For all genre, the proposed algorithm produced similar results. Crosstalk component, which occurs more due to the relatively small weight, had to be below masking threshold of non-crosstalk component. Fig. 6 shows the rate that means crosstalk component is greater than masking threshold of non-crosstalk component for several frequencies during iterations for TEST 1 and TEST 2. Although the proposed algorithm exhibits faster and more accurate convergence for non-crosstalk components than the conventional algorithm, it allows more errors in the crosstalk components. But these increased errors could be masked by the filtered outputs of non-crosstalk components. Fig. 7 shows the total mean-squared errors for TEST 1 and TEST 2. To simulate a changing environment, listener moves to the right about 0.2 m at 7.54 sec which is equal to 650 iteration. Fig. 8 shows the convergence rate of time-varying listening situation. The proposed algorithm is more accurate and has faster convergence rate in non-crosstalk components than the conventional one. Crosstalk components were almost below masking threshold of non-crosstalk components.



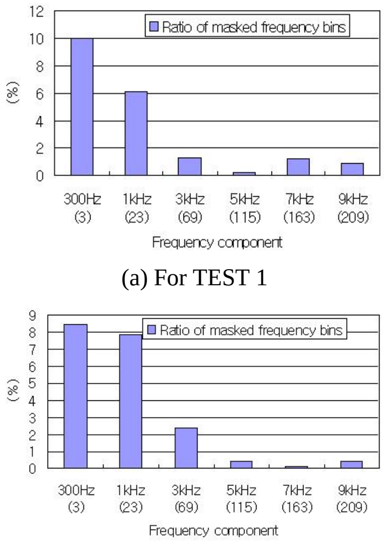
(a) For TEST 1

(b) For TEST 2

Fig. 5. Convergence performance of non-crosstalk component.

5. CONCLUSION

In this paper, a novel adaptive algorithm for the crosstalk cancellation using psychoacoustic model was proposed. In our experiment, the proposed algorithm leads to a smaller MSE and faster convergence rate than the conventional LMS algorithm. This is achieved by masking the crosstalk components with the non-crosstalk components when its masking threshold is high. Thus, this algorithm leads to a perceptually more accurate cancellation and faster convergence rate than the conventional algorithm. This algorithm would be suitable for general time-varying listening situation.



(a) For TEST 1

(b) For TEST 2

Fig. 6. Comparison crosstalk component to masking threshold of non-crosstalk component.

6. ACKNOWLEDGMENTS

This work was supported by grant No. R01-2003-000-10829-0 from the Basic Research Program of the Korea Science and Engineering Foundation and by University IT Research Center Project.

7. REFERENCES

[1] B.S. Atal and M.R. Schroeder, “Apparent sound source translator,” U.S. Patent 3 236 949, 1962.

[2] P. Damaske, “Head-related two-channel stereophony with loudspeaker reproduction,” *J. Acoust. Soc. Amer.*, vol. 50, 1971.

[3] D.H. Cooper and J.L. Bauck, “Prospects for transaural recording,” *J. Audio Eng. Soc.*, vol. 37, pp. 3-19, 1989.

[4] S.T. Neely and J.B. Allen, “Invertibility of a room impulse response,” *J. Acoust. Soc. Amer.*, vol. 66, pp. 165-169, 1979.

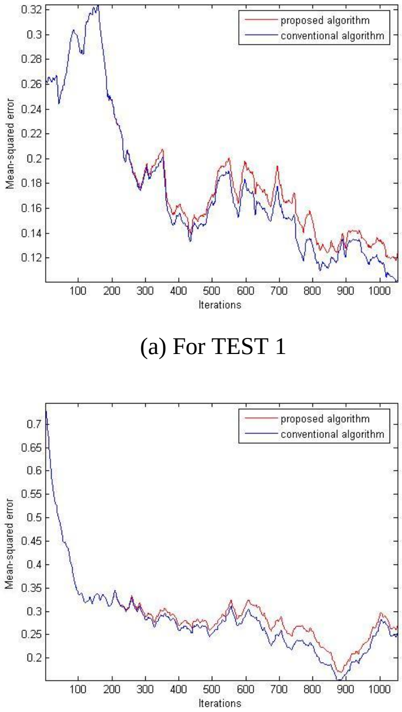
[5] S. Haykin, *Adaptive filter theory, 3rd ed.*, Englewood Cliffs, NJ: Prentice-Hall, 1996.

[6] P.A. Nelson, H. Hamada, and S. J. Elliott, “Adaptive inverse filters for stereophonic sound reproduction,” *IEEE Trans. Signal Process.*, vol. 40, pp. 1621-1632, July, 1992.

[7] E.R. Ferrara, Jr., “Fast implementation of LMS adaptive filters,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-28, no. 4, pp. 474-475, Aug. 1980.

[8] J.S. Lim and C. Kyriakakis, “Multirate Adaptive Filtering for Immersive Audio,”, *Proceedings of the IEEE Acoust. Speech Signal Process. International Conference (ICASSP ’01)*, vol. 5, pp. 3357-3360, May, 2001.

[9] T.Painter and A.Spanias, “Perceptual coding of digital audio,” *Proc. of IEEE*, vol. 88, no. 4, pp. 451-513, Apr. 2000.



(a) For TEST 1

(b) For TEST 2

Fig. 7. Total mean-squared error.

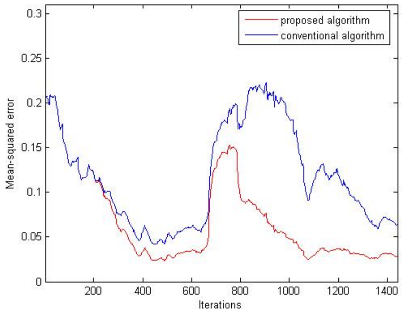


Fig. 8. Convergence rate of time-varying listening situation.