On Crosstalk Cancellation and Equalization With Multiple Loudspeakers for 3-D Sound Reproduction

Yiteng (Arden) Huang, *Member, IEEE*, Jacob Benesty, *Senior Member, IEEE*, and Jingdong Chen, *Member, IEEE*

***Abstract*—People prefer to be able to enjoy spatial audio without wearing a headphone. Such a tethered device is anyway inconvenient and undesirable, if not cumbersome. Alternatively, 3-D sound can be delivered to a listener with loudspeakers. However, crosstalk arises, and the rendered binaural signals are distorted by room reverberation when arriving at the listener's two ears, which lead to the need for a crosstalk cancellation and equalization (CTCE) system. Classical CTCE systems employ only two loudspeakers, and their performance is usually unsatisfactory in practice. While the idea of using more loudspeakers has been investigated, it was never shown why using more loudspeakers is theoretically more advantageous for CTCE. In this letter, we will study this problem and demonstrate that with two loudspeakers, only a least-squares (LS) solution can be obtained, while using multiple loudspeakers, we have more options: either an LS solution or an exact solution for perfect CTCE. These findings are justified by simulations using real impulse responses measured in the varechoic chamber at Bell Labs.**

***Index Terms*—Crosstalk cancellation, equalization, inverse filtering, multichannel acoustic signal processing, 3-D sound reproduction.**

I. Introduction

The 3-D audio technology based on head-related transfer functions (HRTFs) (for synthesizing binaural signals) can position sound sources in 3-D space with only a two-loudspeaker presentation [1]. This is not possible with classical stereo systems. Therefore, 3-D systems have the potential to be used in many applications such as computer gaming and multiparty teleconferencing over the IP networks, where there is a great need for the participants to be able to differentiate competing sounds or voices. However, the delivery of these binaural signals to the listener's ears (assuming that loudspeakers are used and not headphones) is not straightforward. Indeed, each ear receives the so-called crosstalk components, and moreover, the direct signals are distorted by the reverberation of the room. Therefore, an inverse filter is required before playing out the binaural signals through the loudspeakers.

The concept of crosstalk cancellation and equalization (CTCE) was first invented by Atal and Schroeder [2] and Bauer [3] in the early 1960s. Many other sophisticated algorithms have been proposed since then, using two or more loudspeakers for rendering the binaural signals. Most of these algorithms are based on least-squares techniques [4]–[7].

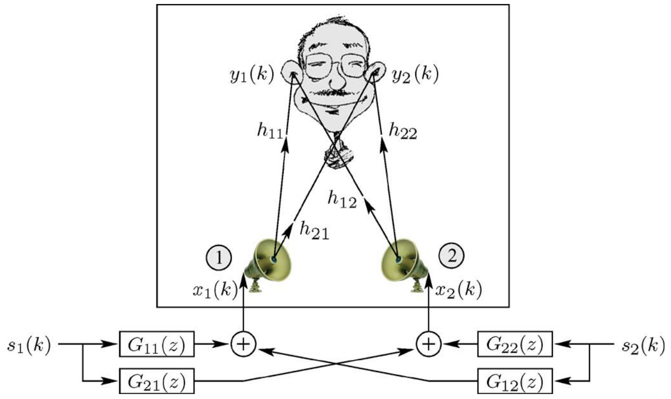

Fig. 1. Schematic diagram of a classical CTCE system using two loudspeakers.

II. Classical Least-Squares Approach With Two Loudspeakers

The least-squares approach is the most used technique for CTCE. In this section, we explain this method in a two-loudspeaker presentation and try to show why it may be limited in practice.

Let $s_1(k)$ and $s_2(k)$ be the binaural signals (k is the time index), $x_1(k)$ and $x_2(k)$ the two loudspeaker signals, and $y_1(k)$ and $y_2(k)$ the signals at the listening points (i.e., the two ears). The objective is to find the filters $g_{mi}, m, i = 1, 2$ in such a way that crosstalk signals are suppressed and the effect of the channel impulse responses ($h_{jm}, j, m = 1, 2$) from the loudspeakers to the ears is reduced. This is equivalent to demanding ideally $y_j(k) = s_j(k - \kappa), j = 1, 2$, with κ being a constant delay. See Fig. 1 for the principle of this scheme.

The loudspeaker signals are

$$x_m(k) = s_1(k) * g_{m1} + s_2(k) * g_{m2}, \quad m = 1, 2 \qquad (1)$$

where the operator "$*$" denotes convolution. We can now write the signals at the listener's ears as

$$y_j(k) = x_1(k) * h_{j1} + x_2(k) * h_{j2}, \quad j = 1, 2. \qquad (2)$$

Substituting (1) into (2), we get

$$y_j(k) = \sum_{i=1}^{2} (g_{1i} * h_{j1} + g_{2i} * h_{j2}) * s_i(k), \quad j = 1, 2 \qquad (3)$$

Manuscript received February 13, 2007; revised March 14, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alfred Hanssen.

Y. Huang and J. Chen are with Bell Laboratories, Alcatel-Lucent, Murray Hill, NJ 07974 USA (e-mail: arden@research.bell-labs.com; jingdong@research.bell-labs.com).

J. Benesty is with the Université du Québec, INRS-EMT, Montréal, QC H5A 1K6, Canada (e-mail: benesty@emt.inrs.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2007.898329

1070-9908/$25.00 © 2007 IEEE

which we can put in a more convenient vector/matrix form as follows:

$$y_j(k) = \sum_{i=1}^{2} \mathbf{s}_{L,i}^T(k)\mathbf{H}_{j:}\mathbf{g}_{:i}, \quad j = 1, 2 \tag{4}$$

where $(\cdot)^T$ denotes a vector/matrix transpose

$$\mathbf{G} = [\mathbf{g}_{:1} \quad \mathbf{g}_{:2}] = \begin{bmatrix} \mathbf{g}_{11} & \mathbf{g}_{12} \\ \mathbf{g}_{21} & \mathbf{g}_{22} \end{bmatrix}$$

is a matrix of size $2L_g \times 2$

$$\mathbf{g}_{mi} = \begin{bmatrix} g_{mi,0} & g_{mi,1} & \cdots & g_{mi,L_g-1} \end{bmatrix}^T \quad m, i = 1, 2,$$

is an FIR filter of length L_g, whose input and output are $s_i(k)$ and $x_m(k)$, respectively

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{1:} \\ \mathbf{H}_{2:} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{bmatrix}$$

is the channel impulse response matrix of size $2L \times 2L_g$, with $L = L_g + L_h - 1$

$$\mathbf{H}_{jm} = \begin{bmatrix} h_{jm,0} & \cdots & h_{jm,L_h-1} & 0 & \cdots & 0 \\ 0 & h_{jm,0} & \cdots & h_{jm,L_h-1} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_{jm,0} & \cdots & h_{jm,L_h-1} \end{bmatrix}^T$$

is a Sylvester matrix of size $L \times L_g$

$$\mathbf{h}_{jm} = \begin{bmatrix} h_{jm,0} & h_{jm,1} & \cdots & h_{jm,L_h-1} \end{bmatrix}^T, \quad j, m = 1, 2$$

is the acoustic impulse response, of length L_h, from the mth loudspeaker to the jth ear, and

$$\mathbf{s}_{L,i}(k) = \begin{bmatrix} s_i(k) & s_i(k-1) & \cdots & s_i(k-L+1) \end{bmatrix}^T, \quad i = 1, 2$$

is a vector containing the L most recent samples of the source signal s_i.

The conditions for CTCT are mathematically expressed as follows:

$$\mathbf{HG} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_2 \end{bmatrix} \tag{5}$$

where $\mathbf{u}_j = [0 \cdots 0\, 1\, 0 \cdots 0]^T$ is a vector of length L, whose κth component is equal to 1, and $\mathbf{0}$ is also a vector of length L containing only zeroes. Assuming that $\mathbf{H}$ has full column rank and $L_h > 1$, it is easy to see from (5) that this linear system has more equations than unknowns since $2L > 2L_g$. In this situation, the best (and only) estimator that we can derive from (5) is the least-squares solution, i.e.,

$$\mathbf{G}^{\mathrm{LS}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T \begin{bmatrix} \mathbf{u}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_2 \end{bmatrix}. \tag{6}$$

However, this solution may not be good enough in practice for several reasons. First, we do not know how to determine L_g. Second, $\mathbf{H}$ may not even be of full rank. Third, from this approach, it is not clear what LS does best, crosstalk cancellation or equalization. In other words, we cannot quantify in a clean way the residual crosstalk signals or the equalization error. Fourth, it is very well known that this method is not very robust to head movements [8].

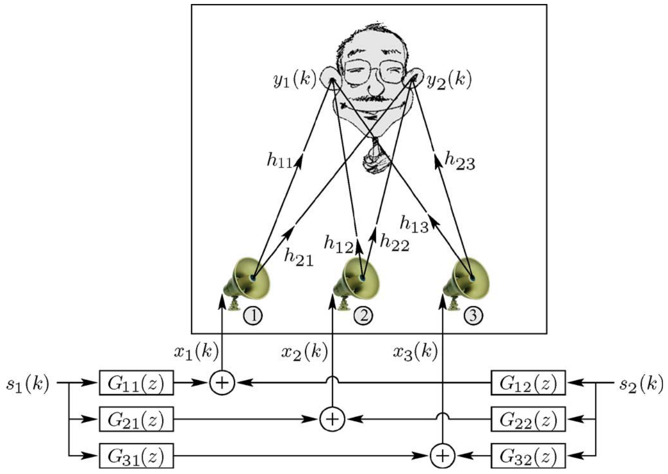

Fig. 2. Illustration of the proposed CTCE system using three loudspeakers for 3-D sound reproduction.

III. Approach to CTCE With More Loudspeakers

In this section, we are going to show that by using channel diversity (i.e., using more than two loudspeakers), more options are available to us to find a robust and reliable solution to our problem.

Here again we have our binaural signals $s_1(k)$ and $s_2(k)$, but this time, we will use $M \geq 3$ loudspeakers to try rendering the sound as exactly as possible at the listener's ears. Fig. 2 depicts the principle of this approach with $M = 3$.

Using the same notation as the previous section, we can easily see that the signals at the listener's ears are

$$y_j(k) = \sum_{i=1}^{2} \mathbf{s}_{L,i}^T(k)\mathbf{H}_{j:}\mathbf{g}_{:i}, \quad j = 1, 2 \tag{7}$$

where this time

$$\mathbf{G} = [\mathbf{g}_{:1} \quad \mathbf{g}_{:2}] = \begin{bmatrix} \mathbf{g}_{11} & \mathbf{g}_{12} \\ \mathbf{g}_{21} & \mathbf{g}_{22} \\ \vdots & \vdots \\ \mathbf{g}_{M1} & \mathbf{g}_{M2} \end{bmatrix}$$

is a matrix of size $(M \cdot L_g) \times 2$, and

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{1:} \\ \mathbf{H}_{2:} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \cdots & \mathbf{H}_{1M} \\ \mathbf{H}_{21} & \mathbf{H}_{22} & \cdots & \mathbf{H}_{2M} \end{bmatrix}$$

is the channel impulse response matrix of size $2L \times (M \cdot L_g)$.

We now deduce the conditions for CTCE as follows:

$$\mathbf{HG} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_2 \end{bmatrix}. \tag{8}$$

The linear system (8) has $(2 \times 2L)$ equations and $(2 \times M \cdot L_g)$ unknowns. Assume that $\mathbf{H}$ has full column rank. Depending on how we choose L_g, we have three very different solutions.

A. Least-Squares Solution

To obtain the least-squares solution [10], we should take L_g in such a way that $2L > M \cdot L_g$, which implies that $L_g < 2(L_h - 1)/(M-2)$. Therefore

$$\mathbf{G}^{\mathrm{LS}} = \left(\mathbf{H}^T\mathbf{H} + \delta\mathbf{I}_{ML_g \times ML_g}\right)^{-1}\mathbf{H}^T \begin{bmatrix} \mathbf{u}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_2 \end{bmatrix} \quad (9)$$

where δ is a nonnegative regularization factor and $\mathbf{I}_{ML_g \times ML_g}$ is the identity matrix of size $ML_g \times ML_g$. Regularization is used to reduce its sensitivity to errors in the measured impulse responses [9].

B. Exact Solution

An exact solution can be derived if we can make $\mathbf{H}$ a square matrix. This is possible if $2L = M \cdot L_g$, which implies that $L_g = 2(L_h - 1)/(M-2)$ if the result of such division is an integer. Hence

$$\mathbf{G}^{\mathrm{E}} = \left(\mathbf{H} + \delta\mathbf{I}_{ML_g \times ML_g}\right)^{-1} \begin{bmatrix} \mathbf{u}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_2 \end{bmatrix}. \quad (10)$$

C. Minimum-Norm Solution

This solution can be obtained if we decide to have more equations than unknowns [10], i.e., $L_g > 2(L_h - 1)/(M-2)$. Hence

$$\mathbf{G}^{\mathrm{MN}} = \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1} \begin{bmatrix} \mathbf{u}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_2 \end{bmatrix}. \quad (11)$$

The first important thing to notice is that, compared to the two-loudspeaker presentation, we have a pretty good idea on how to determine the length L_g of the CTCE filters g_{mi}. While this number of FIR filters is doubled, its required length will probably be much smaller than the length of the filters of the classical least-squares solution (with two loudspeakers), with likely much better performances.

The least-squares technique may be the most interesting in practice since it gives an upper bound for L_g. Moreover, $\mathbf{H}$ may not be full column rank. In this case, we can reduce the length L_g until we get an acceptable solution.

The exact solution $[L_g = 2(L_h - 1)/(M-2)]$ can be seen as a generalization of the multiple-input/output inverse theorem (MINT) [11]. Recall that the MINT can exactly equalize any number of points in a room using a monaural signal only. Here, we generalized the idea to binaural signals.

The minimum-norm solution seems useless from a practical system design point of view because there is no good reason why we should choose L_g much longer than necessary. In particular, when the number of used loudspeakers is low, L_g for the minimum-norm solution can be much longer than the length of the acoustic impulse responses h_{jm}.

IV. Simulations

In this section, we evaluate the performance of the proposed CTCE algorithm in a comparison with the classical method by simulations.

A. Performance Measures

In this letter, two performance measures are employed: signal-to-crosstalk ratio (SCTR) and signal-to-distortion ratio (SDR). Let us first denote

$$\mathbf{H}\mathbf{G} = \mathbf{F} = \begin{bmatrix} \mathbf{f}_{11} & \mathbf{f}_{21} \\ \mathbf{f}_{12} & \mathbf{f}_{22} \end{bmatrix}. \quad (12)$$

So what the jth ear hears due to the signal $s_i(k)$ is found as

$$y_{j,s_i}(k) = \mathbf{s}_{L,i}^T(k)\mathbf{f}_{ji}, \quad i,j = 1,2. \quad (13)$$

Therefore, the SCTR at the two ears would be

$$\mathrm{SCTR}_1 = \frac{E\left\{y_{1,s_1}^2(k)\right\}}{E\left\{y_{1,s_2}^2(k)\right\}}, \quad \mathrm{SCTR}_2 = \frac{E\left\{y_{2,s_2}^2(k)\right\}}{E\left\{y_{2,s_1}^2(k)\right\}} \quad (14)$$

where $E\{\cdot\}$ denotes mathematical expectation. Substituting (13) into (14) leads to

$$\mathrm{SCTR}_1 = \frac{\mathbf{f}_{11}^T\mathbf{R}_{s_1s_1}\mathbf{f}_{11}}{\mathbf{f}_{12}^T\mathbf{R}_{s_2s_2}\mathbf{f}_{12}}, \quad \mathrm{SCTR}_2 = \frac{\mathbf{f}_{22}^T\mathbf{R}_{s_2s_2}\mathbf{f}_{22}}{\mathbf{f}_{21}^T\mathbf{R}_{s_1s_1}\mathbf{f}_{21}} \quad (15)$$

where $\mathbf{R}_{s_is_i} = E\{\mathbf{s}_{L,i}(k)\mathbf{s}_{L,i}^T(k)\}$, $i = 1, 2$ is the autocorrelation matrix of $s_i(k)$. In general, the SCTRs depend not only on the CTCE filters g_{mi} but also on the binaural signals. Since our interest is merely in the CTCE system, without any loss of generality, we can assume that the binaural signals are white and have the same strength. Then $\mathbf{R}_{s_is_i} = \sigma_{s_i}^2\mathbf{I}_{L\times L}$. Consequently, the SCTRs are calculated as follows:

$$\mathrm{SCTR}_1 = \frac{\mathbf{f}_{11}^T\mathbf{f}_{11}}{\mathbf{f}_{12}^T\mathbf{f}_{12}}, \quad \mathrm{SCTR}_2 = \frac{\mathbf{f}_{22}^T\mathbf{f}_{22}}{\mathbf{f}_{21}^T\mathbf{f}_{21}} \quad (16)$$

and the average SCTR is given by $\mathrm{SCTR} = (\mathrm{SCTR}_1 + \mathrm{SCTR}_2)/2$.

At the jth ear, the signal distortion is defined as

$$d_j(k) = y_{j,s_j}(k) - \mathbf{s}_{L,j}^T(k)\mathbf{u}_j, \quad j = 1,2. \quad (17)$$

Substituting (13) into (17) produces

$$d_j(k) = \mathbf{s}_{L,j}^T(k)(\mathbf{f}_{jj} - \mathbf{u}_j), \quad j = 1,2. \quad (18)$$

Then the SDR at the jth ear is determined by

$$\begin{aligned}\mathrm{SDR}_j &= \frac{E\left\{\left[\mathbf{s}_{L,j}^T(k)\mathbf{u}_j\right]^2\right\}}{E\left\{d_j^2(k)\right\}} \\ &= \frac{\mathbf{u}_j^T\mathbf{R}_{s_js_j}\mathbf{u}_j}{(\mathbf{f}_{jj} - \mathbf{u}_j)^T\mathbf{R}_{s_js_j}(\mathbf{f}_{jj} - \mathbf{u}_j)}, \quad j = 1,2. \quad (19)\end{aligned}$$

Using the assumption of white binaural signals, we deduce that

$$\mathrm{SDR}_j = \frac{1}{(\mathbf{f}_{jj} - \mathbf{u}_j)^T(\mathbf{f}_{jj} - \mathbf{u}_j)}, \quad j = 1,2 \quad (20)$$

and the average $\mathrm{SDR} = (\mathrm{SDR}_1 + \mathrm{SDR}_2)/2$.

B. Simulation Setup

The simulations were carried out using the impulse responses measured in a real, reverberant environment: the

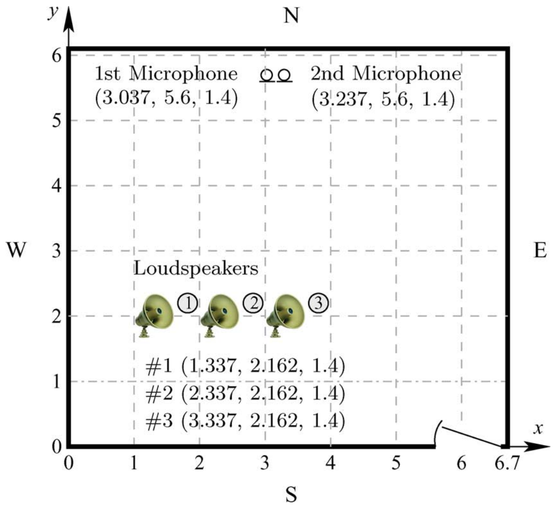

Fig. 3. Floor plan of the varechoic chamber at Bell Labs (coordinate values measured in meters).

varechoic chamber at Bell Labs [12]. The chamber is a rectangular room $(6.7\ \mathrm{m} \times 6.1\ \mathrm{m} \times 2.9\ \mathrm{m})$ with 368 electronically controlled panels that vary the acoustic absorption of the walls, floor, and ceiling [13]. Therefore, the level of room reverberation is well controlled by the percentage of open panels. Three panel configurations were investigated: 75%, 30%, and 0% open panels. Their average T_{60} reverberation times are approximately 310 ms, 380 ms (moderately reverberant), and 580 ms (highly reverberant), respectively. The original impulse responses were measured at 8 kHz and had 4096 samples. For the three panel configurations, the impulse responses were truncated to $L_h = 310$, 380, and 580 samples, respectively. Gaussian random noise is added in the impulse responses with 30 or 15 dB signal-to-noise ratio (SNR). The regularization factor δ is specified as 0.01 and 0.5, respectively, at 30 and 15 dB SNR. We investigated using two and three loudspeakers for CTCE. The positions of the loudspeakers and the two microphones (to simulate a listener's two ears) are shown in Fig. 3.

C. Simulation Results

The simulation results are summarized in Table I. It is clearly demonstrated that the performance of a CTCE system is significantly improved by using multiple loudspeakers in either lightly or heavily reverberant environments. If L_g is chosen such that the number of parameters (denoted as N_p, which determines the computational complexity of the CTCE algorithm) is the same, a CTCE system using three loudspeakers at 15 dB SNR can achieve almost the same performance as a CTCE system using only two loudspeakers at 30 dB SNR. In the case of using three loudspeakers, the exact solution is ideal in the absence of noise in the impulse responses. However, in practice, where the impulse responses cannot be precisely measured, the study suggests to use an LS algorithm and choose a L_g that is only slightly smaller than $2(L_h - 1)/(M - 2)$.

TABLE I
PERFORMANCE OF A CTCE SYSTEM USING VARIOUS NUMBERS OF LOUDSPEAKERS AND IN DIFFERENT ACOUSTIC ENVIRONMENTS

T_{60}	L_h	SNR	δ	M	Solution	L_g	N_p	SCTR	SDR
(ms)		(dB)			Type			(dB)	(dB)
310	310	30	0.01	2	LS	897	3588	14.0	11.1
				3	LS	598	3588	19.6	15.8
				3	Exact	618	3708	5.1	-9.7
		15	0.5	2	LS	897	3588	10.0	6.3
				3	LS	598	3588	14.2	7.7
				3	Exact	618	3708	1.2	-14.7
380	380	30	0.01	2	LS	1107	4428	12.3	10.9
				3	LS	738	4428	19.5	19.8
		15	0.5	2	LS	1107	4428	9.2	6.1
				3	LS	738	4428	12.1	8.6
580	580	30	0.01	2	LS	1707	6828	13.0	10.9
				3	LS	1138	6828	20.3	20.2
		15	0.5	2	LS	1707	6828	10.0	6.4
				3	LS	1138	6828	12.0	10.0

V. CONCLUSIONS

CTCE is a challenging problem, but it is critical for hands-free 3-D sound reproduction. It has been reported that the performance of a CTCE system using only two loudspeakers is usually unsatisfactory in practice. In this letter, we analyzed the problem with use of multiple (more than two) loudspeakers. We showed mathematically that using two loudspeakers only an LS solution can be obtained. Using multiple loudspeakers was recommended. We further showed that by using more loudspeakers, we can take advantage of acoustic channel diversity and get either an LS or an exact solution by choosing a proper length for the CTCE filters. As a result, we can design a more robust CTCE system that is less sensitive to errors in the measured impulse responses. Finally, these analyses were justified by simulations using real impulse responses measured in the varechoic chamber at Bell Labs.

REFERENCES

[1] C. Avendano, "Virtual spatial sound," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Boston, MA: Kluwer, 2004, ch. 13, pp. 345–370.

[2] B. S. Atal and M. R. Schroeder, "Apparent Sound Source Translator," U.S. Patent 3,236,949, Feb. 1966 (filed 1962).

[3] B. B. Bauer, "Stereophonic earphones and binaural loudspeakers," *J. Audio Eng. Soc.*, vol. 9, pp. 148–151, 1961.

[4] P. A. Nelson, H. Hamada, and S. J. Elliott, "Adaptive inverse filters for stereophonic sound reproduction," *IEEE Trans. Signal Process.*, vol. 40, no. 7, pp. 1621–1632, Jul. 1992.

[5] S. M. Kuo and G. H. Canfield, "Dual-channel audio equalization and cross-talk cancellation for 3-D sound reproduction," *IEEE Trans. Consum. Electron.*, vol. 43, no. 4, pp. 1189–1196, Nov. 1997.

[6] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis, "Inverse filter design for immersive audio rendering over loudspeakers," *IEEE Trans. Multimedia*, vol. 2, no. 2, pp. 77–87, Jun. 2000.

[7] S. Miyabe, M. Shimada, T. Takatani, H. Saruwatari, and K. Shikano, "Multi-channel inverse filtering with selection and enhancement of a loudspeaker for robust sound field reproduction," in *Proc. IWAENC*, 2006, pp. 1–4.

[8] D. B. Ward and G. W. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," *IEEE Signal Process. Lett.*, vol. 6, no. 5, pp. 106–108, May 1999.

[9] Y. Tatekura, Y. Nagata, H. Saruwatari, and K. Shikano, "Adaptive algorithm of iterative inverse filter relaxation to acoustic fluctuation in sound reproduction system," in *Prof. Int. Congr. Acoustics*, 2004, vol. IV, pp. 3163–3166.

[10] T. K. Moon and W. C. Stirling*, Mathematical Methods and Algorithms*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[11] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-36, pp. 145–152, Feb. 1988.

[12] A. Härmä, Acoustic Measurement Data From the Varechoic Chamber, Technical Memorandum, Agere Systems, Nov. 2001.

[13] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougald, "The new Varechoic chamber at AT&T Bell Labs," in *Proc. Wallance Clement Sabine Centennial Symp.*, 1994, pp. 343–346.