

GENERALIZED CROSSTALK CANCELLATION AND EQUALIZATION USING MULTIPLE LOUDSPEAKERS FOR 3D SOUND REPRODUCTION AT THE EARS OF MULTIPLE LISTENERS

Yiteng (Arden) Huang[†], Jacob Benesty[‡], and Jingdong Chen[†]

[†] Bell Laboratories, Alcatel-Lucent
600 Mountain Avenue
Murray Hill, New Jersey 07974, USA
{arden, jingdong}@research.bell-labs.com

[‡] Université du Québec, INRS-EMT
800 de la Gauchetière Ouest, Suite 6900
Montréal, Québec, H5A 1K6, Canada
benesty@emt.inrs.ca

ABSTRACT

A classical crosstalk cancellation and equalization (CTCE) system uses two loudspeakers and assumes only one listener. In this paper, the idea is generalized to the design of using multiple loudspeakers for multiple listeners. We will show that if the number of loudspeakers is equal to the number of ears, only a least-squares (LS) solution can be obtained, while using more loudspeakers than ears we have more options: either an LS solution or an exact solution for perfect CTCE. Via simulations using real impulse responses measured in the varechoic chamber at Bell Labs, we learn that a CTCE system employing more loudspeakers will be more robust to errors in the estimated acoustic impulse responses.

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

1. INTRODUCTION

Spatial audio systems have the potential to be used in many applications such as computer gaming and multi-party teleconferencing over the IP networks, where there is a great need for the participants to be able to differentiate competing sounds or voices. But wearing a tethered headphone to enjoy spatial audio is anyway inconvenient and undesirable, if not cumbersome. Alternatively 3D sound can be delivered to a listener using loudspeakers. But crosstalk arises and the rendered binaural signals are distorted by room reverberation when arriving at the listener's two ears, which leads to the need for a crosstalk cancellation and equalization (CTCE) system.

The concept of CTCE was invented by Atal and Schroeder [1] and Bauer [2] in the early 1960's. As illustrated in Fig. 1, a classical CTCE system employs two loudspeakers and assumes only one listener.

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 of CTCE is then to find the filters g_{mp} ($m, p = 1, 2$) in such a way that crosstalk signals are suppressed and the effect of the channel impulse responses h_{nm} ($n, m = 1, 2$) from the loudspeakers to the ears is reduced. This is equivalent to demanding ideally $y_n(k) = s_n(k - \kappa)$ ($n = 1, 2$) with κ being a constant delay.

The loudspeaker signals are:

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

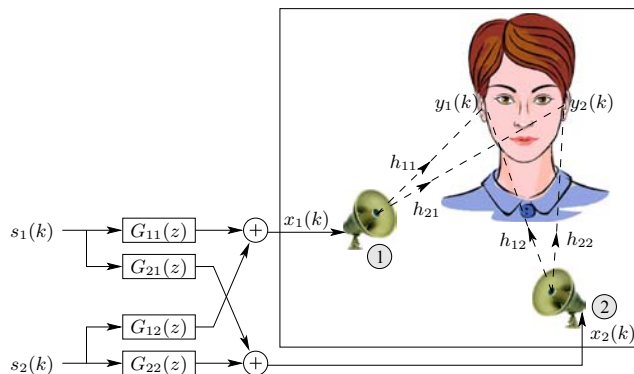


Fig. 1. Schematic diagram of a classical CTCE system using two loudspeakers.

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

$$y_n(k) = x_1(k) * h_{n1} + x_2(k) * h_{n2}, \quad n = 1, 2. \quad (2)$$

Substituting (1) into (2), we get:

$$y_n(k) = \sum_{p=1}^2 (g_{1p} * h_{n1} + g_{2p} * h_{n2}) * s_p(k), \quad n = 1, 2, \quad (3)$$

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

$$y_n(k) = \sum_{p=1}^2 \mathbf{s}_{L,p}^T(k) \mathbf{H}_n \mathbf{g}_p, \quad n = 1, 2, \quad (4)$$

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

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_{\cdot 1} & \mathbf{g}_{\cdot 2} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_{11} & \mathbf{g}_{12} \\ \mathbf{g}_{21} & \mathbf{g}_{22} \end{bmatrix}$$

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

$$\mathbf{g}_{mp} = [g_{mp,0} \ g_{mp,1} \ \cdots \ g_{mp,L_g-1}]^T, \quad m, p = 1, 2,$$

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

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

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

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

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

$$\mathbf{h}_{nm} = [h_{nm,0} \ h_{nm,1} \ \cdots \ h_{nm,L_h-1}]^T, \quad m, n = 1, 2,$$

is the acoustic impulse response, of length L_h , from the m th loudspeaker to the n th ear, and

$$\mathbf{s}_{L,p}(k) = [s_p(k) \ s_p(k-1) \ \cdots \ s_p(k-L+1)]^T, \quad p = 1, 2,$$

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

Then the conditions for CTCE are mathematically expressed as follows:

$$\mathbf{H}\mathbf{G} = \begin{bmatrix} \mathbf{u}_\kappa & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_\kappa \end{bmatrix}, \quad (5)$$

where $\mathbf{u}_\kappa = [0 \ \cdots \ 0 \ 1 \ 0 \ \cdots \ 0]^T$ is a vector of length L , whose κ th 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 [3], [4], i.e.,

$$\mathbf{G}^{\text{LS}} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \begin{bmatrix} \mathbf{u}_\kappa & \mathbf{0} \\ \mathbf{0} & \mathbf{u}_\kappa \end{bmatrix}. \quad (6)$$

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's not clear what LS does best, crosstalk cancellation or equalization. In other words, we can not 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 [5]. The idea of using more loudspeakers has been proposed [6], [7]. In an earlier study [9], we have shown that using more loudspeakers is more advantageous for CTCE. But these CTCE algorithms can deliver 3D sound to only one listener and therefore are found inadequate for multi-user spatial audio applications (e.g., teleconferencing with immersive audio). In the following section, we will generalize the idea of CTCE to the design of using multiple loudspeakers for multiple listeners.

2. GENERALIZED CTCE USING MULTIPLE LOUSPEAKERS FOR MULTIPLE LISTENERS

Here we suppose to use M loudspeakers and try to deliver P ($P \geq 2$) channels of signals $s_p(k)$ ($p = 1, 2, \dots, P$) to a number of listeners whose overall N ears are at various positions, as shown in Fig. 2. Apparently $P \leq N$ and presumably $M \geq N$. In general N is even (since everyone has two ears) but this is not necessarily required.

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

$$y_n(k) = \sum_{p=1}^P \mathbf{s}_{L,p}^T(k) \mathbf{H}_{n:\mathbf{g}_p}, \quad n = 1, 2, \dots, N, \quad (7)$$

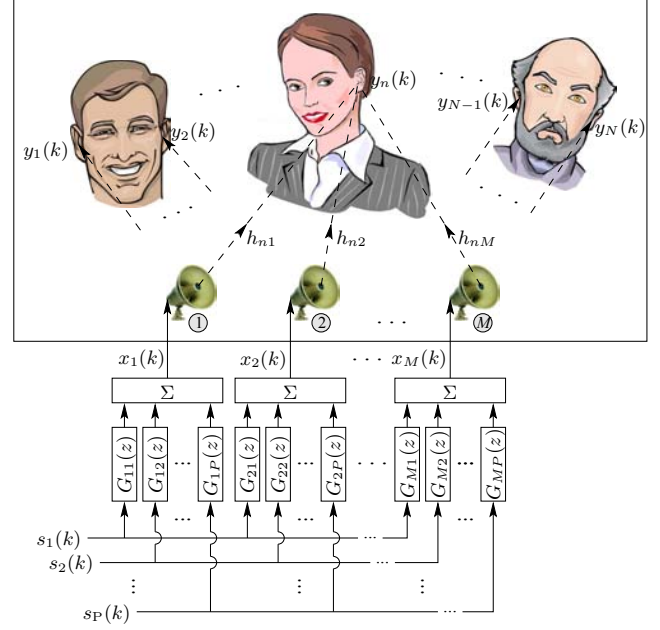


Fig. 2. Illustration of the generalized CTCE system using M loudspeakers for 3D sound reproduction at the overall N ears of a number of listeners.

where this time,

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

is a matrix of size $ML_g \times P$, and

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{1:} \\ \mathbf{H}_{2:} \\ \vdots \\ \mathbf{H}_{N:} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \cdots & \mathbf{H}_{1M} \\ \mathbf{H}_{21} & \mathbf{H}_{22} & \cdots & \mathbf{H}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{N1} & \mathbf{H}_{N2} & \cdots & \mathbf{H}_{NM} \end{bmatrix}$$

is the channel impulse response matrix of size $NL \times ML_g$.

We now deduce the conditions for the generalized CTCE:

$$\mathbf{H}\mathbf{G} = \mathbf{U} = \begin{bmatrix} \mathbf{u}_{11} & \mathbf{u}_{12} & \cdots & \mathbf{u}_{1P} \\ \mathbf{u}_{21} & \mathbf{u}_{22} & \cdots & \mathbf{u}_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_{N1} & \mathbf{u}_{N2} & \cdots & \mathbf{u}_{NP} \end{bmatrix}, \quad (8)$$

where for $n = 1, 2, \dots, N$, $p = 1, 2, \dots, P$,

$$\mathbf{u}_{np} = \begin{cases} \mathbf{u}_\kappa & \text{if supposedly } y_n(k) = s_p(k - \kappa), \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

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

1. Least-Squares Solution:

To obtain the least-squares solution, we should take L_g in such a way that $NL > ML_g$, which anyway holds if $M = N$

and $L_h > 1$. But for $M > N$, this condition implies that $L_g < N(L_h - 1)/(M - N)$. In such cases,

$$\mathbf{G}^{\text{LS}} = (\mathbf{H}^T \mathbf{H} + \delta \mathbf{I}_{ML_g \times ML_g})^{-1} \mathbf{H}^T \mathbf{U}, \quad (9)$$

where δ is a non-negative 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 [8].

2. Exact Solution:

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

$$\mathbf{G}^{\text{EX}} = (\mathbf{H} + \delta \mathbf{I}_{ML_g \times ML_g})^{-1} \mathbf{U}. \quad (10)$$

3. Minimum-Norm Solution:

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

$$\mathbf{G}^{\text{MN}} = \mathbf{H}^T (\mathbf{H} \mathbf{H}^T)^{-1} \mathbf{U}. \quad (11)$$

Discussion: The first important thing to notice is that when using more loudspeakers (i.e., $M > N$), we have a pretty good idea on how to determine L_g (the length of the CTCE filters). While there are more CTCE filters, their required length will probably be much smaller than the length of the filters in the case of $M = N$, 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 for which $L_g = N(L_h - 1)/(M - N)$ can be seen as a generalization of the multiple-input/output inverse theorem (MINT) [10]. 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 multichannel 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 loudspeaker is low, L_g for the minimum-norm solution can be much longer than the length of the acoustic impulse responses h_{nm} .

3. SIMULATIONS

In this section, we evaluate the performance of the proposed CTCE system by simulations.

3.1. Performance Measures

Similar to [9], two performance measures are employed in these simulations: signal-to-crosstalk ratio (SCTR) and signal-to-distortion ratio (SDR). Let's first denote

$$\mathbf{H} \mathbf{G} = \mathbf{F} = \begin{bmatrix} \mathbf{f}_{11} & \mathbf{f}_{12} & \cdots & \mathbf{f}_{1P} \\ \mathbf{f}_{21} & \mathbf{f}_{22} & \cdots & \mathbf{f}_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{f}_{N1} & \mathbf{f}_{N2} & \cdots & \mathbf{f}_{NP} \end{bmatrix}_{N \times P}. \quad (12)$$

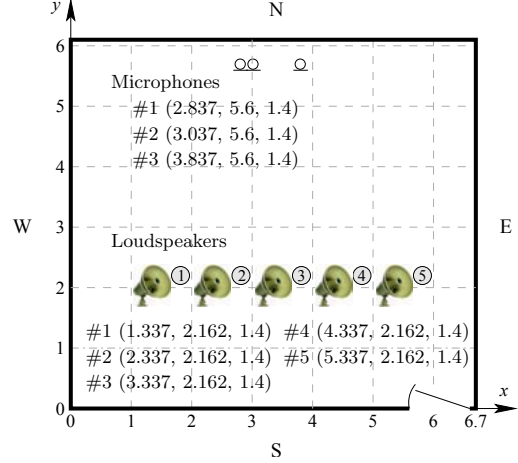


Fig. 3. Floor plan of the varechoic chamber at Bell Labs (coordinate values measured in meters).

So what the n th ear hears due to the signal $s_p(k)$ is found as

$$y_{n,s_p}(k) = \mathbf{f}_{np}^T \mathbf{s}_{L,p}(k), \quad n = 1, 2, \dots, N, \quad p = 1, 2, \dots, P. \quad (13)$$

Suppose that we intend to deliver at the n th ear the signal $s_{p_n}(k)$. Then the SCTR at the n th ear would be

$$\text{SCTR}_n = \frac{E \{ y_{n,s_{p_n}}^2(k) \}}{\sum_{p=1, p \neq p_n}^P E \{ y_{n,s_p}^2(k) \}}, \quad (14)$$

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

$$\text{SCTR}_n = \frac{\mathbf{f}_{np_n}^T \mathbf{R}_{s_{p_n} s_{p_n}} \mathbf{f}_{np_n}}{\sum_{p=1, p \neq p_n}^P \mathbf{f}_{np}^T \mathbf{R}_{s_p s_p} \mathbf{f}_{np}}, \quad (15)$$

where $\mathbf{R}_{s_p s_p} = E \{ \mathbf{s}_{L,p}(k) \mathbf{s}_{L,p}^T(k) \}$ ($p = 1, 2, \dots, P$) is the auto-correlation matrix of $s_p(k)$. In general, the SCTR depends not only on the CTCE filters g_{mp} , but also on the spatial sound signals $s_p(k)$. Since our interest is merely in the CTCE system, without any loss of generality, we can assume that the spatial sound signals are white and have the same strength. Then $\mathbf{R}_{s_p s_p} = \sigma_{s_p}^2 \mathbf{I}_{L \times L}$. Consequently, the SCTR_n are calculated as follows

$$\text{SCTR}_n = \frac{\mathbf{f}_{np_n}^T \mathbf{f}_{np_n}}{\sum_{p=1, p \neq p_n}^P \mathbf{f}_{np}^T \mathbf{f}_{np}}, \quad (16)$$

and the average SCTR is given by $\text{SCTR} = \sum_{n=1}^N \text{SCTR}_n / N$.

At the n th ear, the signal distortion is defined as

$$d_n(k) = y_{n,s_{p_n}}(k) - s_{p_n}(k - \kappa) = y_{n,s_{p_n}}(k) - \mathbf{u}_\kappa^T \mathbf{s}_{L,p_n}(k). \quad (17)$$

Substituting (13) into (17) produces

$$d_n(k) = (\mathbf{f}_{np_n} - \mathbf{u}_\kappa)^T \mathbf{s}_{L,p_n}(k). \quad (18)$$

Then the SDR at the n th ear is determined by

$$\begin{aligned} \text{SDR}_n &= E \left\{ \left[\mathbf{u}_\kappa^T \mathbf{s}_{L,p_n}(k) \right]^2 \right\} / E \{ d_n^2(k) \} \\ &= \frac{\mathbf{u}_\kappa^T \mathbf{R}_{s_{p_n} s_{p_n}} \mathbf{u}_\kappa}{(\mathbf{f}_{np_n} - \mathbf{u}_\kappa)^T \mathbf{R}_{s_{p_n} s_{p_n}} (\mathbf{f}_{np_n} - \mathbf{u}_\kappa)}. \end{aligned} \quad (19)$$

Using the assumption of white spatial sound signals, we deduce that

$$\text{SDR}_n = 1/[(\mathbf{f}_{npn} - \mathbf{u}_\kappa)^T (\mathbf{f}_{npn} - \mathbf{u}_\kappa)], \quad (20)$$

and the average $\text{SDR} = \sum_{n=1}^N \text{SDR}_n / N$.

3.2. Simulation Setup

The simulations were carried out using the impulse responses measured in a real, reverberant environment: the varechoic chamber at Bell Labs [11]. A detailed description of the chamber can be found in [12]. In our simulations, 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 these panel configurations, the impulse responses were truncated to $L_h = 310, 380$, and 580 samples, respectively. These numbers were carefully chosen to avoid a too complicated CTCE system with a memory requirement that our laptops cannot sustain. 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 $M = 3$ and 5 loudspeakers for CTCE with $P = 2$ and $N = 3$. The ears with odd index are supposed to hear $s_1(k)$ and the ears with even index hear $s_2(k)$. The positions of the loudspeakers and the three microphones (to simulate the ears of the listeners) are shown in Fig. 3.

3.3. Simulation Results

The simulation results are summarized in Table 1. It is clearly demonstrated that the performance of a CTCE system is significantly improved by using more loudspeakers in either lightly or heavily reverberant environments. From these simulations, we learned that the exact solution is ideal in the absence of noise in the impulse responses. But in practice where the impulse responses cannot be precisely measured, the study suggests to use an LS algorithm and choose an L_g that is only slightly smaller than $L_g^{\text{EX}} \triangleq N(L_h - 1)/(M - N)$.

4. CONCLUSIONS

In this paper, the idea of crosstalk cancellation and equalization (CTCE) was generalized to the design of using multiple loudspeakers for multiple listeners. We showed mathematically that using the same number of loudspeakers as the number of overall ears of the listeners, only a least-squares (LS) solution can be obtained. 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 generalized 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.

5. REFERENCES

- [1] B. S. Atal and M. R. Schroeder, "Apparent sound source translator," U.S. Patent 3,236,949, Feb. 1966 (filed 1962).
- [2] B. B. Bauer, "Stereophonic earphones and binaural loudspeakers," *J. Audio Eng. Soc.*, vol. 9, pp. 148–151, 1961.

Table 1. Performance of the generalized CTCE system using various numbers of loudspeakers and in different acoustic environments.

T_{60} (ms)	L_h	SNR (dB)	δ	M	N	L_g^{EX}	L_g	N_p	SCTR (dB)	SDR (dB)
310	310	30	0.01	3	3	—	738	4428	14.89	13.22
				5	3	463.5	443	4430	19.42	16.54
							463	4630	19.54	16.66
				3	3	463.5	443	4430	11.22	7.15
				5	3	463.5	443	4430	14.59	8.38
							463	4630	14.75	8.44
380	380	30	0.01	3	3	—	913	5478	13.97	11.13
				5	3	568.5	548	5480	18.24	14.92
							568	5680	18.35	15.03
				3	3	—	913	5478	10.95	6.65
				5	3	568.5	548	5480	13.69	7.72
							568	5680	13.79	7.77
580	580	30	0.01	3	3	—	1413	8478	12.49	10.81
				5	3	868.5	848	8480	17.11	15.70
							868	8680	17.24	15.87
				3	3	—	1413	8478	10.02	6.56
				5	3	868.5	848	8480	13.80	7.97
							868	8680	13.88	8.01

Note: $N_p = P \cdot M \cdot L_g$ denotes the number of parameters, which determines the computational complexity of the generalized CTCE system. $L_g^{\text{EX}} \triangleq N(L_h - 1)/(M - N)$.

- [3] 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.
- [4] S. M. Kuo and G. H. Canfield, "Dual-channel audio equalization and cross-talk cancellation for 3-D sound reproduction," *IEEE Trans. Consumer Electronics*, vol. 43, pp. 1189–1196, Nov. 1997.
- [5] D. B. Ward and G. W. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," *IEEE Signal Process. Lett.*, vol. 6, pp. 106–108, May 1999.
- [6] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis, "Inverse filter design for immersive audio rendering over loudspeakers," *IEEE Trans. Multimedia*, vol. 2, pp. 77–87, June 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] Y. Tatekura, Y. Nagata, H. Saruwatari, and K. Shikano, "Adaptive algorithm of iterative inverse filter relaxation to acoustic fluctuation in sound reproduction system," in *Proc. Int. Congress on Acoustics*, 2004, vol. IV, pp. 3163–3166.
- [9] Y. Huang, J. Benesty, and J. Chen, "On crosstalk cancellation and equalization with multiple loudspeakers for 3D sound reproduction," *IEEE Signal Process. Lett.*, vol. 14, pp. 649–652, Oct. 2007.
- [10] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-36, pp. 145–152, Feb. 1988.
- [11] A. Härmä, "Acoustic measurement data from the varechoic chamber," Technical Memorandum, Agere Systems, Nov. 2001.
- [12] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougald, "The new Varechoic chamber at AT&T Bell Labs," in *Proc. Wallace Clement Sabine Centennial Symposium*, 1994, pp. 343–346.