On the Robustness of Room Impulse Response Reshaping

Tiemin Mei
School of Information Science and Engineering
Shenyang Ligong Univ. of Tech., Shenyang 110168 China
Email: meitiemin@163.com

Alfred Mertins
Institute for Signal Processing
University of Lübeck, Lübeck 23562 Germany
Email: alfred.mertins@isip.uni-luebeck.de

***Abstract*—In room impulse response (RIR) equalization and reshaping, one of the difficulties is the spatial robustness, because RIRs are very sensitive to the movements of both the signal source and the receiver. For example, the reshaping filter designed for one pair of loudspeaker/receiver or source/microphone positions will be ineffective for another pair. In this paper, we concentrate on loudspreaker/receiver pairs and propose a novel approach in which we use multiple prefilters to reshape simultaneously the RIR samples in a given area of interest (listening area). According to the RIR sampling principle, we prove statistically that the listening area will be reshaped if only the RIR samples in this area are reshaped. In simulations, we show that the proposed approach is valid.**

I. INTRODUCTION

Room impulse response equalization and reshaping in a loudspeaker/receiver setting aims to provide a listening-room compensation (LRC) by pre-processing loudspeaker signals in a suitable way. It helps to improve the speech intelligibility in reverberant rooms and enables new applications in audio-visual communications and virtual acoustics. Similarly, for improving the quality of far-field microphone recordings, a post-filtering stage may be introduced for the received signals. Both problems are mathematically equivalent, and for the sake of conciseness, we describe our method for the LRC problem.

LRC compensates the channel so that the received signals are perceived without reverberation. For a scenario where the loudspeaker and the receiver are fixed to their positions, several LRC approaches have been proposed [1], [2], [3], [4], [5], [6]. The classical approaches in [1], [2], [3] try to equalize the acoustic channel as best as possible. More recent methods [4], [5], [6], [7] try to shorten or reshape the channel response, which leave more degrees of freedom in the filter design and promises better echo suppression. However, whatever method is used for the filter design, if the loudspeaker or the receiver move around in the neighborhood of their supposed positions, or if there are moving objects in the room, the RIR will be changing with time, which will lead to a degradation of the LRC performance.

For a given LRC system, if the performance is not affected by changes of the environment or by movements of the loudspeaker or receiver, the LRC system is spatially robust, otherwise, it is not spatially robust. If only one loudspeaker is used for source-signal playback, it is impossible to keep the reshaping system spatially robust within a given area in a room, because the RIR is very sensitive to the position of the loudspeaker and/or receiver. But in contrary, if multiple loudspeakers are used and sufficiently many RIR samples in the listening area are equalized or reshaped, according to the spatial sampling principle of RIR [8], it is theoretically possible to control the RIRs in a given limited listening area. The basic idea of the multiple loudspeaker reshaping approach proposed in this paper is that if samples of RIRs taken in the listening area are reshaped, so as to cancel the reverberation effects, and if these samples are homogeneously distributed and are dense enough in the listening area, then the RIR of an unsampled point in the listening area will be reshaped as well, so the whole listening area is reshaped. When receivers move in this listening area, there will be no reverberation heard.

II. SPATIAL RIR INTERPOLATION

Let us denote the RIR from one point $P_L(x_L, y_L, z_L)$ to another point $P_M(x_M, y_M, z_M)$ in a listening room by $c_{P_L,P_M}(t)$, where $P_L(x_L, y_L, z_L)$ is the position of the loudspeaker and $P_M(x_M, y_M, z_M)$ is the position of the microphone or receiver. The RIR is not only a function of time but also a function of the coordinates of loudspeaker/microphone positions. Clearly, the time-domain sampling rate $f_{t.s.}$ is equal to or higher than two times the highest frequency f_0 in the signal. From the point of view of wave equation, for a given position of the loudspeaker, the impulse response $c_{P_L,P_M}(t)$ is a bandlimited spatial function for any time t. If $c_{P_L,P_M}(t)$ is bandlimited in the time domain to f_0, then the spatial frequency is limited to f_0/v, where v is the speed of sound. So the space-domain sampling rate $f_{s.s.}$ of the RIR is [8]:

$$f_{s.s.} \geq 2f_0/v. \tag{1}$$

In general, we use

$$f_{s.s.} \geq f_{t.s.}/v. \tag{2}$$

We can do the space-domain sampling by either moving a microphone from place to place in the listening area while keeping the loudspeaker fixed outside the listening area or vice versa. In a room, the listening area is quite limited compared with the entire room space, so for the purpose of RIR reshaping, we just focus on the listening area.

Denoting the discretized RIRs by $c_{Q_L,Q_M}(t)$, where $Q_L(i, j, k)$ represents the position of a loudspeaker and $Q_M(l, m, n)$ is the position of a microphone in the listening area, for each loudspeaker position, we sample the RIR in

the listening area. If the space-sampling condition is fully satisfied, then for any point $P_M(x,y,z)$ in the listening area, the RIR caused by the loudspeaker at position $Q_L(i,j,k)$ can be reconstructed through interpolation:

$$c_{Q_L(i,j,k),P_M(x,y,z)}(t) = \sum_{l,m,n} c_{Q_L(i,j,k),Q_M(l,m,n)}(t)\phi(x,y,z,l,m,n), \quad (3)$$

where the interpolating function is $\phi(x,y,z,l,m,n) \equiv \phi_0(x-l\Delta x, y-m\Delta y, z-n\Delta z)$ for convenience; Δx, Δy and Δz are the spatial sampling periods in the three spatial directions; $\phi_0(x,y,z)$ is the interpolation function, which is independent of the loudspeaker positions. For example, it is well known that possible interpolating functions are the so-called sampling functions given by $\phi_0(x) = \frac{sin(\pi x/\Delta x)}{\pi x/\Delta x}$ and $\phi_0(x,y) = \frac{sin(\pi x/\Delta x)}{\pi x/\Delta x}\frac{sin(\pi y/\Delta y)}{\pi y/\Delta y}$ for one- and two-dimensional spaces, respectively, and

$$\phi_0(x,y,z) = \frac{sin(\pi x/\Delta x)}{\pi x/\Delta x}\frac{sin(\pi y/\Delta y)}{\pi y/\Delta y}\frac{sin(\pi z/\Delta z)}{\pi z/\Delta z}$$

for the three-dimensional case. The frequency supporting domain is accordingly a line $([-f_{s.s.}/2, f_{s.s.}/2])$, a square $([-f_{s.s.}/2, f_{s.s.}/2])^2$, and a cubic $([-f_{s.s.}/2, f_{s.s.}/2])^3$, respectively.

For two- and three-dimensional space, the orthogonal sampling meshes are not optimal. Regular triangle for 2D space and the so-called "body-centered cubic" for 3D space are optimal [9].

Let the prefilter of loudspeaker $Q_L(i,j,k)$ be $h_{ijk}(t)$, then the reshaped impulse response at $P_M(x,y,z)$ in the listening area is as follows:

$$\begin{aligned} g_{P_M(x,y,z)}(t) &= \sum_{i,j,k} h_{ijk}(t) * c_{Q_L(i,j,k),P_M(x,y,z)}(t) \\ &= \sum_{l,m,n} \phi(x,y,z,l,m,n) g_{Q_M(l,m,n)}(t), \end{aligned} \quad (4)$$

where $g_{Q_M(l,m,n)}(t) = \sum_{i,j,k} h_{ijk}(t) * c_{Q_L(i,j,k)Q_M(l,m,n)}(t)$ is the reshaped RIR of the sampling point $Q_M(l,m,n)$ in the listening area; '$*$' is the convolution operator.

For a good RIR reshaping, as described in [10][6], the attenuation of the reshaped RIR should be limited under an upper bound which is defined by the average masking window that is derived from the auditory masking effect, i.e., this upper-bound decays -10 dB at 4 ms after the direct impulse and then it decays exponentially to -70 dB at 200 ms [11]. We define this upper bound as $g_0(t)$. In addition, for the investigation of spatial robustness of RIR reshaping, the spatial characteristics of RIRs must be taken into account, so the reshaped RIR should satisfy the following hypotheses: For a given time instant t, $g_{Q_M(l,m,n)}(t)$ is a spatially stationary field and $|g_{Q_M(l,m,n)}(t)|$ is limited in the following ways:

- $\mathrm{E}[g_{Q_M(l,m,n)}(t)] = \mathrm{const}(t)$,
- $\mathrm{E}[g_{Q_M(l,m,n)}(t) g_{Q_M(i,j,k)}(t)] = r_g(t, l-i, m-j, n-k)$,
- $|g_{Q_M(l,m,n)}(t)| \le |g_0(t)|$ for $t > 4$ ms,

where $\mathrm{E}[.]$ is the spatial ensemble average operator. The third hypothesis implies that $r_g(t,0,0,0) = \mathrm{E}[g^2_{Q_M(l,m,n)}(t)] \le \mathrm{E}[g_0^2(t)]$ for $t > 4$ ms.

Let us look at the statistical properties of the reshaped RIR of any given point in the listening area. If $x = l\Delta x$, $y = m\Delta y$, $z = n\Delta z$, then $g_{P_M(x,y,z)}(t) = g_{Q_M(l,m,n)}(t)$. If $x \neq l\Delta x$, $y \neq m\Delta y$, $z \neq n\Delta z$, then $g_{P_M(x,y,z)}(t)$ is the linear combination of $g_{Q_M(l,m,n)}(t)$ shown in (4). The spatial ensemble average of $g_{P_M(x,y,z)}(t)$ is,

$$\begin{aligned} \mathrm{E}[g_{P_M(x,y,z)}(t)] &= \sum_{l,m,n} \phi(x,y,z,l,m,n) \mathrm{E}[g_{Q_M(l,m,n)}(t)] \\ &= \mathrm{const}(t) \sum_{l,m,n} \phi(x,y,z,l,m,n) \end{aligned} \quad (5)$$

for $t > 4$ ms.

Let us consider the average energy of a RIR tap for any given point in the listening area, we get,

$$\mathrm{E}[g^2_{Q_M(x,y,z)}(t)] = \sum_{k_x,k_y,k_z=-\infty}^{\infty} r_g(t,k_x,k_y,k_z)\Psi(x,y,z,k_x,k_y,k_z), \quad (6)$$

where $\Psi(x,y,z,k_x,k_y,k_z) = \sum_{l,m,n} \phi(x,y,z,l,m,n) \times \phi(x,y,z,l+k_x,m+k_y,n+k_z)$.

Using the properties of sampling function, it is easy to prove that

$$\sum_{l,m,n} \phi(x,y,z,l,m,n) = 1 \quad (7)$$

and

$$\Psi(x,y,z,k_x,k_y,k_z) = \begin{cases} 1 & k_x,k_y,k_z = 0, \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

so we obtain

$$\mathrm{E}[g_{P_M(x,y,z)}(t)] = \mathrm{const}(t) \quad (9)$$

and

$$\mathrm{E}[g^2_{P_M(x,y,z)}(t)] = r_g(t,0,0,0) \le \mathrm{E}[g_0^2(t)]. \quad (10)$$

This means that for any given point $P_M(x,y,z)$ in the listening area, the impulse response $g_{P_M(x,y,z)}(t)$ will decay statistically as fast as those sampling points' reshaped impulse responses on which the reshaping filters are designed, so we can say that the listening area is reshaped.

III. Algorithm development

For the RIR reshaping, N_s loudspeakers and N_m microphones are used. The microphones are used for the sampling of RIR in this listening area and the loudspeakers are used for playback of source signals. Let $c_{ij}(n)$ denote the impulse response from the jth loudspeaker to the ith microphone, and let L_c be the length of $c_{ij}(n)$. Moreover, let $h_k(n)$ denote the impulse response of the kth prefilter with length L_h, then the ith global impulse response of this prefilter-loudspeaker-room system is as follows, where we have subsumed the loudspeaker responses as a part of the room impulse responses:

$$g_i(n) = \sum_{k=1}^{N_s} h_k(n) * c_{ik}(n) = \sum_{k=1}^{N_s} \mathbf{C}_{ik}\mathbf{h}_k \quad (11)$$

with $\mathbf{C}_{ik}$ being an L_g-by-L_h convolution matrix made up of sequence $c_{ik}(n)$. With 'global' we mean that it is the response from the original source over the N_s loudspeakers to the ith microphone position. The length of $g_i(n)$ is $L_g = L_c + L_h - 1$. Our aim is to design prefilters which make the global impulse responses $g_i(n)$ not only attenuate faster than the impulse response of the room but also allow them to satisfy certain psychoacoustic conditions so that there will be no audible echoes. For this, we define an unwanted part of the reshaped RIRs as

$$g_{ui}(n) = w_{ui}(n) g_i(n), \tag{12}$$

where $i = 1, 2, \ldots, N_m$, and $w_{ui}(n)$ is a suitable window function (details will be given in Section IV). Other than in [10], the wanted part of the impulse response is not explicitly specified.

We exploit the p-norm for defining the objective function. The corresponding optimization problem is given by

$$\begin{aligned} \mathrm{MIN}_{\mathbf{h}}: \ f(\mathbf{h}) &= \log\left(\|\mathbf{g}_u\|_p\right) \\ &= \log\left(\left(\sum_{i=1}^{N_m}\sum_{k=0}^{L_g-1} |g_{ui}(k)|^p\right)^{\frac{1}{p}}\right) \end{aligned} \tag{13}$$

where $\mathbf{h} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{N_s}]$ and $\mathbf{g}_u = [\mathbf{g}_{u1}^{\mathrm{T}}, \mathbf{g}_{u2}^{\mathrm{T}}, \ldots, \mathbf{g}_{uN_m}^{\mathrm{T}}]^{\mathrm{T}}$. For simplicity, p is usually set to be an even integer, so the problem is simplified as follows,

$$\mathrm{MIN}_{\mathbf{h}}: \ f(\mathbf{h}) = \frac{1}{p}\log\left(\phi_{f_u}(\mathbf{h})\right), \tag{14}$$

where $\phi_{f_u}(\mathbf{h}) = \sum_{i=1}^{N_m}\sum_{k=0}^{L_g-1} g_{ui}^p(k)$. All of the $g_{ui}(n)$'s are dealt with as a whole, so all of the sampled RIRs will be reshaped simultaneously.

From (14), the steepest descent learning rule reads

$$\mathbf{h}^{l+1} = \mathbf{h}^l - \mu p^{-1} \phi_{f_u}^{-1}(\mathbf{h}^l) \nabla_{\mathbf{h}} \phi_{f_u}(\mathbf{h}^l), \tag{15}$$

where

$$\nabla_{\mathbf{h}}\phi_{f_u}(\mathbf{h}) = \left[\sum_{k=1}^{N_m}\mathbf{C}_{k1}^{\mathrm{T}}\mathbf{b}_k, \sum_{k=1}^{N_m}\mathbf{C}_{k2}^{\mathrm{T}}\mathbf{b}_k, \ldots, \sum_{k=1}^{N_m}\mathbf{C}_{kN_s}^{\mathrm{T}}\mathbf{b}_k\right] \tag{16}$$

and

$$b_k(n) = w_{uk}(n) g_{uk}^{p-1}(n). \tag{17}$$

The fast Fourier transform (FFT) can be exploited for the calculation of $\mathbf{C}_{ki}^{\mathrm{T}}\mathbf{b}_k$, where $k = 1, 2, \ldots, N_m$ and $i = 1, 2, \ldots, N_s$, so the algorithm (15) is very computationally efficient [10].

IV. SIMULATIONS

Simulations were performed for a room of dimensions $5\ \mathrm{m} \times 4\ \mathrm{m} \times 2.5\ \mathrm{m}$, using the method in [12]. The simulated room impulse responses were of $L_c = 2000$ taps at a sampling frequency of $f_{t.s.} = 16\mathrm{kHz}$. 13 loudspeakers were used to playback source signals, and 57 microphones were considered to sample the RIRs in a limited listening area of dimensions $2.0\ \mathrm{cm} \times 8.0\ \mathrm{cm} \times 8.0\ \mathrm{cm}$. For the loudspeakers array, six of the 13 loudspeakers were uniformly distributed on a circle of a 40 cm diameter, one was located at the center of this circle, the last 6 loudspeakers were located at the center of the equilateral triangles which are defined by the 6 loudspeakers on the circle and the centered loudspeaker.

We define the unwanted part's window function of the global RIR as described in [10]. For the kth microphone,

$$\mathbf{w}_{uk} = [\underbrace{0, 0, \ldots, 0}_{N_{1k}+N_{2k}}, \underbrace{\mathbf{w}_{0k}^{\mathrm{T}}}_{N_{3k}}]^{\mathrm{T}}, \tag{18}$$

where $N_{1k} = t_{0k} f_s$, $N_{2k} = 0.004 f_s$, and $N_{3k} = L_g - N_{1k} - N_{2k}$ with f_s being the sampling frequency and t_{0k} being the minimal time taken by the direct sound from the N_s loudspeakers to the kth microphone. The window $\mathbf{w}_{0k}$ is defined as

$$w_{0k}(n) = 10^{\frac{3}{\log(N_{0k}/(N_{1k}+N_{2k}))}\log(\frac{n}{N_{1k}+N_{2k}})+0.5} \tag{19}$$

with $N_{0k} = (0.2 + t_{0k}) f_s$ and time index n ranging from $N_{1k} + N_{2k} + 1$ to $L_g - 1$. The reason for defining the window $w_u(n)$ in this form is that we can take $\frac{1}{w_u(n)}$ as the average masking limit. If the RIR or reshaped RIR exceed this limit, it implies that audible reverberation exists. For an accurate quantitative description, we define the following parameter:

$$\mathrm{RQ} = \frac{\sum_{n=N_0}^{L_g-1} g_{\mathrm{EM}}^2(n)}{\sum_{n=N_0}^{L_g-1} g^2(n)} \tag{20}$$

as the reverberation quantization (RQ) parameter, where

$$g_{\mathrm{EM}}(n) = \begin{cases} |g(n)| - \frac{1}{w_u(n)} & \text{for } |g(n)| > \frac{1}{w_u(n)}, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

N_0 in (20) corresponds to the sample which is 4 ms later then the direct impulse of $g(n)$. RQ is the power ratio between the exceeding part of $g(n)$ over the masking limit and $g(n)$ itself. If the RIR is completely reshaped, then $\mathrm{RQ} = 0$.

For a given listening area, it is more convenient to use the logarithmic average of RQs of the N randomly distributed points in this area, we call this logRQ:

$$\mathrm{logRQ} = -10\log\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{RQ}_i\right). \tag{22}$$

The spatially sampled listening area with a sampling period of 2.0 cm is presented in Fig. 1, it is the so-called 'body-centered cubic' mesh. The listening area is of dimension $8\mathrm{cm} \times 8\mathrm{cm} \times 2\mathrm{cm}$. An example of an unreshaped RIR is given in Fig. 2(a); the corresponding averaged global RIR of 30 random points in the listening area is shown in Fig. 2(b).

For the example mentioned above, the reshaping filters have a length of $L_h = 3000$. The logRQ's before and after reshaping are listed in Table I. Before reshaping, the logRQ of the listening area is about 13.86 dB, but after the reshaping, the logRQ is 30.32. The exceeding energy over the masking limit is effectively reduced. Informal listening tests showed that the reverberation is effectively suppressed.

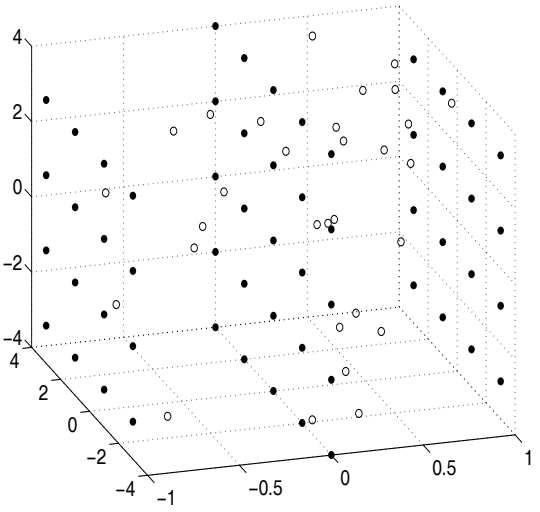

Fig. 1. Spatial sampling: the 57 microphone array (black dots) for spatial sampling of the listening area and the 30 random positions of microphone (circles) for performance evaluation. The spatial sampling period is 2 cm.

TABLE I
THE LOGRQ BEFORE/AFTER RESHAPING.

Spatial sampling period	2.0 cm
logRQ, before reshaping	13.86 dB
logRQ, after reshaping	30.32 dB

V. CONCLUSIONS

The robustness of room impulse response reshaping is practically important for a good LRC system. In this paper we propose to sample the sound field of a room in the listening area to get the spatio-temporal RIRs for the design of reshaping filters. In addition, the p-norm optimization approach is exploited. Simulations show that this approach is encouraging. Real measurements will be done in near future.

ACKNOWLEDGMENT

This work has been supported by the German Research Foundation under Grant No. ME1170/3-1.

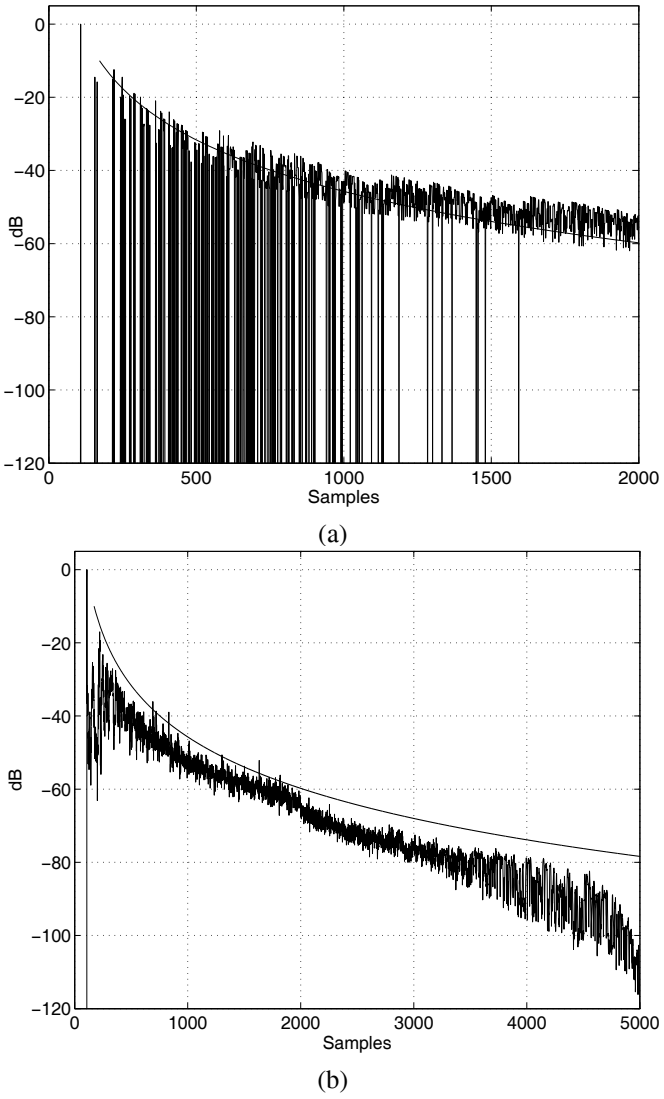

Fig. 2. Spatial sampling: (a) one of the unreshaped RIR; (b) the averaged global RIR of 30 random points in the listening area: $10\log(\sum_{i=1}^{30} g_i^2(n)/30)$. The spatial sampling period of the cubic is 2 cm. The exponentially descending curve is the masking limit.

REFERENCES

[1] S. T. Neely and J. B. Allen, "Invertibility of a Room Impulse Response," *Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, Jul. 1979.

[2] J. N. Mourjopoulos, "Digital Equalization of Room Acoustics," *Journal of the Audio Engineering Society*, vol. 42, no. 11, pp. 884–900, Nov. 1994.

[3] L. D. Fielder, "Practical Limits for Room Equalization," in *AES Convention*, vol. 111, New York, NY, USA, Sep. 2001, preprint No. 5481.

[4] M. Kallinger and A. Mertins, "Room impulse response shortening by channel shortening concepts," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA , USA, Oct. 30 - Nov. 2 2005, pp. 898–902.

[5] ——, "Impulse response shortening for acoustic listening room compensation," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands, Sept. 2005, pp. 197–200.

[6] T. Mei, A. Mertins, and M. Kallinger, "Room impulse response shortening with infinity-norm optimization," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Taipei, Taiwan, April 2009, pp. 3745–3748.

[7] W. Zhang, A. Khong, and P. Naylor, "Adaptive inverse filtering of room acoustics," in *Proc. 42nd Asilomar Conference on Signals, Systems and Computers*, Oct. 2008, pp. 788 –792.

[8] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *IEEE Trans. Signal Processing*, vol. 54, no. 10, pp. 3790 –3804, Oct. 2006.

[9] T. Theussl, T. Moller, and M. E. Groller, "Optimal regular volume sampling," in *Proc. Visualization 2001*, Oct. 2001, pp. 91–98.

[10] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 249–259, Feb. 2010.

[11] L. D. Fielder, "Analysis of traditional and reverberation-reducing methods of room equalization," *J. Audio Eng. Soc.*, vol. 51, no. 1/2, pp. 3–26, Feb. 2003.

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