OPTIMIZED PREPROCESSING FOR SPATIALLY ROBUST ROOM IMPULSE RESPONSE RESHAPING

Jan Ole Jungmann, Radoslaw Mazur, and Alfred Mertins

University of Lübeck, Institute for Signal Processing, 23562 Lübeck
{*jungmann, mazur, mertins*}*@isip.uni-luebeck.de*

ABSTRACT

The purpose of room impulse response reshaping is to reduce reverberation and thus to improve the perceived quality of the received signal by prefiltering the source signal before it is played back with a loudspeaker. The optimization of an infinity- and/or p-norm based objective function has proven to be quite effective compared to least-squares methods. Multi-position approaches have been developed in order to increase the robustness against small movements of the listener. The drawback, however, of the multi-position approach is the great amount of measurements that need to be done prior to equalizer design. A recent method considered the system perturbations in the case of small spatial mismatch and with arbitrary weighting for the reverberation tail. The drawback of this approach is the computation effort required to preprocess the data. In this paper we present a method to significantly speed up this preprocessing step.

Index Terms— room impulse response, reshaping, robustness, optimization, p-norm.

1. INTRODUCTION

The task of listening room compensation (LRC) aims at neutralizing the convolutional distortions that are added to an audio signal by transmission in a closed room. A filter is placed before the loudspeaker to preprocess the audio signal. The goal is to reduce the influence of the room impulse response (RIR) in order to obtain a signal that is hardly distinguishable from the original signal by a human listener [1].

Early approaches computed the equalizer by minimizing the difference between the global impulse response (GIR, that is the convolution of the RIR and the equalizer) and a desired target system in a least-squares sense [2].

A more relaxed requirement is to define arbitrary desired *shapes* for the GIR. It has been shown in [3] that a *shaping* rather than a *shortening* of the GIR is preferable in practice, because by shaping the GIR the temporal masking effect of the human auditory system can be exploited efficiently.

In [4] the least-squares measure has been generalized to a p-norm based optimality criterion. It has been shown that by adequately choosing the involved parameters, the optimization process leads to an equalizer that distributes the perceivable errors evenly across the GIR's time coefficients. The objective function has been extended in [5] to explicitly control the frequency response of the overall system.

Unfortunately, all of these approaches lack spatial robustness. In case of small spatial mismatch (e.g. due to the listener moving his head slightly) the performance of the equalizer degrades greatly [6].

There are, in general, two approaches to improve spatial robustness. In [7] the approach from [4] has been extended to achieve reshaping at multiple positions. If the spatial sampling is dense enough, then the listener is allowed to move in a small volume without perceiving a degraded performance. This method has been extended by a frequency-domain based regularization term to guarantee a flat overall frequency response [8]. The second method is to explicitly consider the system errors in the optimization problem [9]. In [8] a stochastic model for the system perturbations was presented with an arbitrary weighting for the reverberant tail. However, a large computation effort was needed to determine required weighting matrices based on the RIRs.

In this paper we propose a method to significantly reduce the time required to preprocess the RIRs.

This paper is organized as follows. In Section 2 we give a review of the p-norm based design of reshaping filters and of the frequency-domain based regularization term. In Section 3 we review the proposed model to capture the system perturbations and derive the new objective function. Results are given in Section 4. Finally, we give some conclusions in Section 5.

Notation: Lowercase boldface characters denote vectors, while uppercase boldface characters denote matrices. The superscript ^{T} denotes transposition. The asterisk $*$ denotes convolution. The operator $\mathrm{diag}\{\cdot\}$ turns a vector into a diagonal matrix, and $\|\cdot\|_p$ returns the ℓ_p-norm (short p-norm) of a vector. Furthermore, $\mathrm{E}\{\cdot\}$ denotes the expectation operator.

2. ROOM IMPULSE RESPONSE RESHAPING

For the reshaping we use the method from [8], where we proposed a comprehensive optimality criterion that captures both the time- and the frequency-domain representations of the

GIR. The approach was originally formulated for an arbitrary number of microphones (i.e. listening positions) and loudspeakers. However, in this paper we consider only one microphone and, for the sake of simplicity, the equations are formulated accordingly. Considering N loudspeakers, the GIR of length L_g at the reference position is given by $g(n) = \sum_{\ell=1}^{N} h_\ell(n) * c_\ell(n)$, where $c_\ell(n)$ is the RIR of length L_c from loudspeaker ℓ to the listening position and $h_\ell(n)$ is the prefilter of length L_h for the ℓ-th loudspeaker. The reshaping filters are designed by defining two window functions $w_d(n)$ and $w_u(n)$ to determine the *desired* and the *unwanted* parts of the GIR. The desired and the unwanted parts are given by $g_d(n) = g(n)\, w_d(n)$ and $g_u(n) = g(n)\, w_u(n)$, respectively.

2.1. RIR Reshaping with p-Norm Optimization

The time-domain representation of the GIR is optimized by solving the optimization problem given by

$$\min_{\mathbf{h}} \quad : \quad f(\mathbf{h}) = \log\left(\frac{f_u(\mathbf{h})}{f_d(\mathbf{h})}\right) \tag{1}$$

with

$$f_d(\mathbf{h}) = \|\mathbf{g}_d\|_{p_d} = \left(\sum_{n=0}^{L_g-1} |g_d(n)|^{p_d}\right)^{\frac{1}{p_d}} \tag{2}$$

and $f_u(\mathbf{h}) = \|\mathbf{g}_u\|_{p_u}$. The target vector $\mathbf{h} = \left[\mathbf{h}_1^T, \ldots, \mathbf{h}_N^T\right]^T$ is the concatenation of the prefilters for the N loudspeakers. The optimization is carried out by applying a gradient-descent procedure.

The advantage of (1) in comparison to a least-squares measure is that by choosing appropriately large values for p_d and p_u (usually chosen between 10 and 20), the error is distributed evenly across the time coefficients in the unwanted part of the GIR.

For the weighting we use the window functions from [4] that capture the temporal masking property of the human auditory system.

2.2. Frequency Domain Based Regularization

It has been shown recently that one has to consider both the time- and the frequency-domain representations of the GIRs to achieve a *good* reshaping without degrading the perceived quality due to high spectral peaks [5].

The regularization term proposed in [5] is given by

$$y(\mathbf{h}) = \|\mathbf{a}_f\|_{p_f}\,, \tag{3}$$

where the vector $\mathbf{a}_f$ is made up by the discrete Fourier transform of the GIR. The method has been generalized in [8] to incorporate an arbitrary number of loudspeakers and microphones.

3. ROBUST RESHAPING USING STATISTICAL KNOWLEDGE

The problem of designing an equalizer for a reference position and then moving the microphone away has been studied by Radlović et al. [6]. In [8] we presented a method to incorporate statistical knowledge about the system perturbations in the case of spatial mismatch into the optimization process. As a novelty we allowed for an arbitrary weighting of the reverberation.

3.1. System Perturbations Caused by Spatial Mismatch

Let $\omega = 2\pi f$ denote the radial frequency and let $C(\omega)$, $P(\omega)$ and $H(\omega)$ be the Fourier transforms of the RIR $c(t)$, its perturbation caused by microphone movement $p(t)$, and the equalizer $h(t)$, respectively. The frequency-dependent error due to misplacement is then given as in [6] by

$$F(\omega) = \mathrm{E}\left\{\left|\left[C(\omega) + P(\omega)\right] H(\omega) - 1\right|^2\right\}. \tag{4}$$

Assuming perfect equalization in the reference position (i.e., $H(\omega) = 1/C(\omega)$) and being in the far field in reverberant environments, the distance measure (as in [6]) reads

$$F(\omega) = \frac{\mathrm{E}\left\{\left|P(\omega)\right|^2\right\}}{\left|C(\omega)\right|^2} = 2 - 2\frac{\sin(\omega D/v)}{\omega D/v}, \tag{5}$$

where v is the speed of sound and D is the deviation of the microphone from the reference location in meters. Solving (5) for $\mathrm{E}\left\{\left|P(\omega)\right|^2\right\}$ yields

$$\mathrm{E}\left\{\left|P(\omega)\right|^2\right\} = \left|C(\omega)\right|^2\left(2 - 2\frac{\sin(\omega D/v)}{\omega D/v}\right). \tag{6}$$

Assuming a bandlimited input signal with a maximum radial frequency ω_c and fulfilling the sampling theorem, the continuous-time signals and impulse responses can be replaced by their discrete-time equivalents (namely $c(n)$, $p(n)$ and $h(n)$). With respect to (6), the autocorrelation sequence for $p(n)$ is given by

$$r_{pp}(n) = r_{cc}(n) * f(n)\,, \tag{7}$$

where $r_{cc}(n) = c(n) * c(-n)$. The sequence $f(n)$ is computed by sampling $F(\omega)$ according to (5) at discrete frequencies and applying the inverse discrete Fourier transform.

3.2. Weighting of the Reverberation

By using N loudspeakers for playback, the global impulse response at the reference position is given by

$$g(n) = \sum_{\ell=1}^{N} c_\ell(n) * h_\ell(n)\,. \tag{8}$$

Outside the reference position the GIR is modeled by

$$g(n) = \sum_{\ell=1}^{N} \left[c_\ell(n) + p_\ell(n)\right] * h_\ell(n)\,, \tag{9}$$

where $p_\ell(n)$ denotes the perturbations of the ℓ-th channel caused by displacement.

Assuming perfect equalization in the reference point, the weighted error due to microphone movement is then given by

$$e(n) = w(n) \sum_{\ell=1}^{N} p_\ell(n) * h_\ell(n)\,, \tag{10}$$

where $w(n)$ is a sequence of positive weights, usually chosen as $w(n) = w_u(n)$.

With $\mathbf{W} = \operatorname{diag}\{\mathbf{w}\}$, $\mathbf{P}_\ell$ being the Toeplitz-structured convolution matrix of size $L_g \times L_h$ made up by $p_\ell(n)$, and $\mathbf{h}_\ell$ being the vector made up by the sequence $h_\ell(n)$, the mean squared error due to spatial movement is given by

$$O = \mathrm{E}\left\{ \sum_{\ell=1}^{N} \|\mathbf{W}\mathbf{P}_\ell \mathbf{h}_\ell\|_2^2 \right\}. \tag{11}$$

Assuming the perturbations to be uncorrelated and $\mathbf{H}_\ell$ being the convolution matrix made up by $\mathbf{h}_\ell$, (11) can be simplified to $O = \sum_{\ell=1}^{N} O_\ell$ with

$$\begin{aligned} O_\ell &= \mathrm{E}\left\{ \|\mathbf{W}\mathbf{P}_\ell \mathbf{h}_\ell\|_2^2 \right\} = \mathrm{E}\left\{ \mathbf{h}_\ell^T \mathbf{P}_\ell^T \mathbf{W}^T \mathbf{W} \mathbf{P}_\ell \mathbf{h}_\ell \right\} \\ &= \mathrm{E}\left\{ \mathbf{p}_\ell^T \mathbf{H}_\ell^T \mathbf{W}^T \mathbf{W} \mathbf{H}_\ell \mathbf{p}_\ell \right\}. \end{aligned} \tag{12}$$

By exploiting the properties of the trace of a matrix, namely $\operatorname{tr}\{\mathbf{AB}\} = \operatorname{tr}\{\mathbf{BA}\}$, O_ℓ can be rewritten as

$$\begin{aligned} O_\ell &= \mathrm{E}\left\{ \operatorname{tr}\left\{ \mathbf{H}_\ell^T \mathbf{W}^T \mathbf{W} \mathbf{H}_\ell \mathbf{p}_\ell \mathbf{p}_\ell^T \right\} \right\} \\ &= \operatorname{tr}\left\{ \mathbf{H}_\ell^T \mathbf{W}^T \mathbf{W} \mathbf{H}_\ell \mathbf{R}_{pp}^{(\ell)} \right\}, \end{aligned} \tag{13}$$

where $\mathbf{R}_{pp}^{(\ell)} = \mathrm{E}\left\{ \mathbf{p}_\ell \mathbf{p}_\ell^T \right\}$ is the autocorrelation matrix for the perturbation. For an impulse response $c_\ell(n)$ and an assumed average displacement D, $\mathbf{R}_{pp}^{(\ell)}$ can be set up as a Toeplitz matrix from the sequence $r_{pp}^{(\ell)}$, which can be computed as stated in (7).

For a tractable computation of the gradient we presented an algorithm in [8] to construct N matrices $\mathbf{M}_\ell$ so that

$$O_\ell = \mathbf{h}_\ell^T \mathbf{M}_\ell \mathbf{h}_\ell. \tag{14}$$

The algorithm from [8] is based on the Cholesky decomposition of autocorrelation matrices of the dimension $L_c \times L_c$ and requires the element-wise sum over L_c matrices of size $L_h \times L_h$. Usually, the lengths of the prefilters and the RIRs are in the range of several thousand taps. Due to the size of the matrices, the calculations are very time consuming and computationally demanding. In the following we present an optimized method for computation of the weighting matrices.

3.3. Optimized Preprocessing

Exploiting the special structure of the matrices $\mathbf{W}$, $\mathbf{R}_{pp}^{(\ell)}$ and $\mathbf{H}_\ell$ the optimality criterion (13) can be rewritten as

$$O_\ell = \sum_{n=0}^{L_g-1} \sum_{i=0,j=0}^{L_c-1} h_\ell(n-j)\, h_\ell(n-i)\, r_{pp}^{(\ell)}(i-j)\, w^2(n)\,. \tag{15}$$

By setting (14) equal to (15), a rule can be found to calculate the individual entries of the matrix $\mathbf{M}_\ell$. As $\mathbf{M}_\ell$ is symmetric, only the upper right triangular part needs to be computed, which is then copied to the lower left of the matrix. The individual components $[\mathbf{M}_\ell]_{p,q}$, $1 \le p, q \le L_h$ of the matrix $\mathbf{M}_\ell$ are given by

$$[\mathbf{M}_\ell]_{p,q} = \begin{cases} \sum\limits_{n=q-1}^{p+L_c-2} w^2(n)\, r_{pp}^{(\ell)}(q-p)\,, & q \ge p \\ [\mathbf{M}_\ell]_{q,p}\,, & \text{else.} \end{cases} \tag{16}$$

3.4. Extended Objective Function

With the different optimality criteria given in Sections 2.1, 2.2 and 3, the proposed optimization problem finally reads

$$\min_{\mathbf{h}} : q(\mathbf{h}) = \widetilde{f}(\mathbf{h}) + \alpha y(\mathbf{h}) \quad \text{s.t.} \quad \mathbf{h}^T\mathbf{h} = 1, \tag{17}$$

where

$$\widetilde{f}(\mathbf{h}) = \log\left(\frac{\widetilde{f}_u(\mathbf{h})}{f_d(\mathbf{h})} \right). \tag{18}$$

With the equations derived in Section 3, $\widetilde{f}_u(\mathbf{h})$ is given by

$$\widetilde{f}_u(\mathbf{h}) = f_u(\mathbf{h}) + \beta \underbrace{\left(\sum_{\ell=1}^{N} \mathbf{h}_\ell^T \mathbf{M}_\ell \mathbf{h}_\ell \right)^{\frac{1}{2}}}_{f^{(P)}(\mathbf{h})}, \tag{19}$$

where $\mathbf{M}_\ell$ is given in (16).

The learning rule reads

$$\mathbf{h}^{l+1} = \mathbf{h}^l - \mu^l \left(\nabla_{\mathbf{h}} \widetilde{f}(\mathbf{h}^l) + \alpha \nabla_{\mathbf{h}} y(\mathbf{h}^l) \right), \tag{20}$$

where μ^l is the adaptive positive step size in iteration l. The side condition is fulfilled by renormalizing the target vector $\mathbf{h}^{l+1}$ after every iteration l.

Finally, the gradient for $\widetilde{f}(\mathbf{h})$ is given by

$$\nabla_{\mathbf{h}} \widetilde{f}(\mathbf{h}) = \frac{1}{\widetilde{f}_u(\mathbf{h})} \nabla_{\mathbf{h}} \widetilde{f}_u(\mathbf{h}) - \frac{1}{f_d(\mathbf{h})} \nabla_{\mathbf{h}} f_d(\mathbf{h})\,, \tag{21}$$

where

$$\nabla_{\mathbf{h}} \widetilde{f}_u(\mathbf{h}) = \nabla_{\mathbf{h}} f_u(\mathbf{h}) + \beta \nabla_{\mathbf{h}} f^{(P)}(\mathbf{h})\,. \tag{22}$$

For the derivation of the individual gradients $\nabla_{\mathbf{h}} f_u(\mathbf{h})$, $\nabla_{\mathbf{h}} f^{(P)}(\mathbf{h})$, $\nabla_{\mathbf{h}} f_d(\mathbf{h})$, and $\nabla_{\mathbf{h}} y(\mathbf{h})$ we refer to [8].

4. RESULTS

For the experiments we used four loudspeakers for playback in a typical office room. We measured four impulse responses $c_\ell(n)$ of length $L_c = 4000$ taps with a sampling frequency of $f_s = 16$ kHz. The reshaping filters were designed with

Table 1. Average values for the nPRQ and SF measures before and after reshaping using the different algorithms. For Alg. A the assumed spatial displacement was $D = 2$ cm.

Setup	nPRQ [dB]	SF
unreshaped	9.93	0.63
Alg. A, $\alpha = 40, \beta = 0$	10.50	0.64
Alg. A, $\alpha = 1, \beta = 5 \cdot 10^{-4}$	3.88	0.60
Alg. B, $\alpha = 10$	1.23	0.70

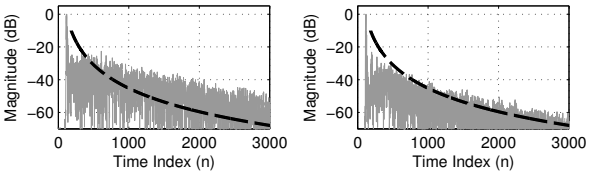

Fig. 1. Global impulse response in the case of small spatial mismatch for the non-robust ($\beta = 0$, left plot) and the robust design method (right plot). The dashed line is the average temporal masking limit.

a length of $L_h = 5000$ taps. For all experiments we chose $p_d = 20$, $p_u = 10$, and $p_f = 8$. To quantify the amount of dereverberation and spectral distortion, we utilize the nPRQ [8] and the *spectral flatness* (SF) measures [10]. The nPRQ measure captures the average overshoot of the time coefficients of an impulse response exceeding the average temporal masking limit and being above -60 dB [8]. The SF measure equals one in the case of a *flat* frequency response and degrades to zero with increasing distortions [10]. To investigate the spatial robustness, we designed the reshaping filters for the reference position and calculated the nPRQ and SF measures for 40 more positions in the vicinity of the reference position. The results are given in a condensed form in Table 1 (denoted as Alg. A) with an assumed displacement of $D = 2$ cm. To compare the proposed method to the multiposition approach (denoted as Alg. B) from [8], we measured 26 more RIRs in the vicinity of the reference position according to the spatial sampling theorem for RIRs. The additional design positions were disjoint to the 40 testing positions. A direct comparison of a GIR in case of small spatial mismatch for the non-robust ($\beta = 0$) and the robust method is given in Fig. 1.

To compare the performance of the new algorithm to the method presented in [8], some measures of the speedup for different lengths of the prefilters and the RIRs were performed. Both algorithms were implemented using MATLAB and benchmarked on a single-core machine running at 3 GHz. The absolute computation times to determine the weighting matrix based on a single RIR and the speedup factors are listed in Table 2, where A.1 denotes the algorithm from [8] and the proposed algorithm is denoted by A.2. We expect a further speedup by using a parallel implementation on multi core systems or dedicated graphics hardware.

Table 2. Computation times in seconds on a single-core machine running at 3 GHz for the former algorithm from [8] (A.1) and the proposed algorithm (A.2).

L_c	L_h	A.1	A.2	Speedup
1000	1000	926 s	4.3 s	**215.3**
2000	2000	14139 s	24.4 s	**579.5**
4000	5000	367112 s	226.3 s	**1622.2**

5. CONCLUSIONS

In this contribution we presented a method to significantly reduce the amount of time to compute spatially robust reshaping filters for listening-room compensation. In comparison to the former implementation, the amount of time needed to set up the required matrices could be lowered by a factor of more than 1600.

6. REFERENCES

[1] J. N. Mourjopoulos, "Digital equalization of room acoustics," *Journal of the Audio Engineering Society*, vol. 42, no. 11, pp. 884–900, Nov. 1994.

[2] S. J. Elliott and P. A. Nelson, "Multiple-point equalization in a room using adaptive digital filters," *Journal of the Audio Engineering Society*, vol. 37, no. 11, pp. 899–907, Nov. 1989.

[3] 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.

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

[5] J. O. Jungmann, T. Mei, S. Goetze, and A. Mertins, "Room impulse response reshaping by joint optimization of multiple p-norm based criteria," in *Proc. European Signal Processing Conference (EUSIPCO 2011)*, Barcelona, Spain, Aug. 2011, pp. 1658–1662.

[6] B. D. Radlović, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: Robustness results," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 311–319, May 2000.

[7] T. Mei and A. Mertins, "On the robustness of room impulse response reshaping," in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC 2010)*, Tel Aviv, Israel, Aug. 2010.

[8] J. O. Jungmann, R. Mazur, M. Kallinger, T. Mei, and A. Mertins, "Combined acoustic mimo channel crosstalk cancellation and room impulse response reshaping," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1829–1842, Aug. 2012.

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

[10] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.