ROOM IMPULSE RESPONSE RESHAPING BY JOINT OPTIMIZATION OF MULTIPLE P-NORM BASED CRITERIA

Jan Ole Jungmann[1], *Tiemin Mei*[2], *Stefan Goetze*[3] *and Alfred Mertins*[1]

[1]Institute for Signal Processing, University of Lübeck, D-23562 Lübeck, Germany
email: {jungmann, mertins}@isip.uni-luebeck.de
web: www.isip.uni-luebeck.de

[2]School of Information Science and Engineering, Shenyang Ligong Univ. of Tech., Shenyang 110168, China
email: meitiemin@163.com

[3]Fraunhofer Institute for Digital Media Technology, D-26129 Oldenburg, Germany
email: s.goetze@idmt.fraunhofer.de

ABSTRACT

The purpose of room impulse response reshaping is usually to reduce reverberation and thus to improve the perceived quality of the received signal by prefiltering the source signal before it is played with a loudspeaker in a closed room or by postfiltering the recorded microphone signal. The utilization of an infinity- and/or p-norm based objective function in the time domain has shown to be quite effective compared to least-squares approaches. Although this method intrinsically favors solutions with flat overall frequency responses, in some cases noticeable spectral distortion may occur. In this contribution we propose a method to jointly optimize infinity- and/or p-norm based objective functions in the time and frequency domains to achieve a good reshaping while not affecting the perceived quality due to spectral distortions.

1. INTRODUCTION

Equalization of an acoustic system is usually carried out on the basis of one of the following setups: a filter for listening room compensation (LRC) is placed in the signal path in front of a loudspeaker or is used to process a recorded microphone signal. The goal is to reduce the influence of the room impulse response (RIR) in order to obtain a signal that is hardly distinguishable by a human listener from the original signal [10].

In recent years, there have been different approaches to design such equalizers. Classically, the equalizer is computed in such a way that it minimizes the difference between the global impulse response (that is the impulse response of the concatenated system containing the equalizer and the RIR) and a given target system in a least-squares sense [5, 1]; usually, the target system is chosen to be a bandpass-filtered and/or delayed unit pulse.

A more relaxed requirement than choosing a bandpass as the target system is to try to concentrate most of the energy of the reshaped response within the first 50 ms after the first peak, thus maximizing the D50-measure [6] for intelligibility of speech known from psychoacoustics. However, while it seems appealing to directly maximize the D50-measure, it was shown in [4] that a *shaping* rather than a *shortening* of the global impulse response is preferable in practice, because the temporal masking effect of the human auditory system can be more efficiently exploited.

The idea of reshaping the room impulse response to fit to an average temporal masking curve was further developed in [9]. Instead of using the least-squares measure, it was proposed to minimize an objective function based on the infinity- and p-norm, respectively. 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 global impulse response's time coefficients while favoring overall impulse responses that have one dominant tap and thus typically have a flat frequency response. However, in practical experiments with very long measured impulse responses it was observed that the method may lead to spectral distortions in some cases, as no spectral requirements are captured by the solely time-domain based objective function used in [9].

In the next section we briefly summarize the approach based on the p-norm optimization of the time-domain representation of the global impulse response to design the equalizer. In Section 3 we motivate and derive the formulas for the proposed extended objective function. In Section 4 we present some simulations and results of our proposed algorithm and, finally, in Section 5 we give some short conclusions.

Notation. Vectors and matrices are printed in boldface. The superscripts ^{T} and ^{H} denote transposition and the hermitian transpose, respectively. The asterisk $*$ denotes convolution. The discrete time index is denoted by n. The operator $\mathrm{diag}\{\cdot\}$ turns a vector into a diagonal matrix, $\max\{\cdot\}$ gives out the maximum component of a vector variable and $\mathrm{sign}\{\cdot\}$ produces a sign vector of its input vector variable, whereat the sign of a complex number is defined as its projection on the unit circle of the complex plane. Finally, $\Re\{\cdot\}$ returns the real part of its input variable.

2. ROOM IMPULSE RESPONSE RESHAPING

Let $c(n)$ and $h(n)$ denote the room impulse response and the equalizer of lengths L_c and L_h, respectively. To simplify matters, $c(n)$ combines the actual RIR and the transfer functions of the loudspeaker and the microphone. The global impulse response (GIR) of the concatenation of the equalizer and the room impulse response is given by $g(n) = h(n) * c(n)$. Its length is denoted as L_g.

One aims to design an equalizer $h(n)$ that *shapes* the GIR in a way that it has a faster decay than $c(n)$ and/or to meet certain other requirements that may be defined by psychoacoustic findings. To actually *shape* or *shorten* an impulse response one usually formulates two window functions $w_d(n)$

and $w_u(n)$ that determine a *desired* part of the GIR and an *unwanted* part. These windows directly model the demands one makes on the global impulse response. The desired and the unwanted parts can now be expressed as $g_d(n) = g(n)\, w_d(n)$ and $g_u(n) = g(n)\, w_u(n)$, respectively.

2.1 Impulse Response Reshaping with p-Norm Optimization

In [9] the approach from [4] (that was based on [7, 8]) was generalized by explicitly incorporating an average temporal masking curve and replacing the least-squares criterion with the more flexible p-norm measure. The approach was developed by formulating the following optimization problem:

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

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

$$f_u(\mathbf{h}) = \|\mathbf{g}_u\|_{p_u} = \left(\sum_{n=0}^{L_g-1} |g_u(n)|^{p_u}\right)^{\frac{1}{p_u}}. \tag{3}$$

The optimization of (1) leads to a minimization of the p-norm of the unwanted part while simultaneously maximizing the p-norm of the desired part of the GIR. By choosing $p_d = p_u = 2$, the solution of (1) degenerates to a least-squares solution.

The advantage of this procedure is that by selecting appropriately large values for p_d and p_u, the error is distributed evenly across the time coefficients in the unwanted part of the global impulse response while favoring the production of one dominant tap in the desired part, which, overall, yields a *good* shaping.

The decay of the overall system depends on the choice of the window functions $w_d(n)$ and $w_u(n)$. We use the windows from [9] defined by:

$$\mathbf{w}_d = [\underbrace{0,0,\ldots,0}_{N_1},\underbrace{1,1,\ldots,1}_{N_2},\underbrace{0,0,\ldots,0}_{N_3}]^T \tag{4}$$

and

$$\mathbf{w}_u = [\underbrace{0,0,\ldots,0}_{N_1+N_2},\underbrace{\mathbf{w}_0^T}_{N_3}]^T \tag{5}$$

where $N_1 = t_0 \cdot f_s$, $N_2 = 0.004\text{s} \cdot f_s$ and $N_3 = L_g - N_1 - N_2$ with f_s being the sampling frequency and t_0 being the time taken by the direct sound. The window $\mathbf{w}_0$ is defined as

$$w_0(n) = 10^{\frac{3}{\log(N_0/(N_1+N_2))}\log\left(\frac{n}{N_1+N_2}\right)+0.5} \tag{6}$$

with $N_0 = (0.2\text{s} + t_0)\, f_s$ and time index n ranging from $N_1 + N_2 + 1$ to $L_g - 1$. These windows represent the compromise temporal masking limit of the human auditory system [2, 9].

3. EXTENDED OBJECTIVE FUNCTION

The sole optimization of (1) considers only the time-domain representation of the GIR, which, in some cases, can lead to spectral distortions. Examples will be presented in Section 4. In this section, a new approach is introduced, which aims at jointly optimizing a time- and a frequency-domain based criterion.

To cope with frequency-domain distortions, we extend the objective function (1) by adding a frequency-based regularization term. We are considering two methods to express the additional criteria:

1. A first approach constrains the frequency response of the equalizer $h(n)$ in such a way that the equalizer does not produce significant spectral peaks.
2. The second method considers the GIR $g(n)$ and tries to avoid peaks of the frequency response of the GIR, while allowing notches.

Both aims can be expressed in a unified form by formulating the following p-norm based optimality criterion:

$$\min_{\mathbf{h}} : q(\mathbf{h}) = f(\mathbf{h}) + \alpha \cdot y(\mathbf{h}) \tag{7}$$

with

$$y(\mathbf{h}) = \left\|\mathbf{a}_f\right\|_{p_f} \tag{8}$$

where $f(\mathbf{h})$ is the same as in (1), $\alpha > 0$ is the weighting factor for the introduced regularization term $y(\mathbf{h})$, $\mathbf{a}_f$ is a vector made up by the discrete Fourier transform (DFT) of a sequence $a(n)$, which is either the equalizer itself (Case 1: $a(n) = h(n)$) or the global impulse response (Case 2: $a(n) = g(n)$). Its length is denoted as L_a. By setting $\alpha = 0$ the new approach is equivalent to (1).

The proposed objective function (7) is minimized by applying a gradient-descent procedure. The update rule reads:

$$\mathbf{h}^{l+1} = \mathbf{h}^l - \mu^l \left(\nabla_{\mathbf{h}} f\left(\mathbf{h}^l\right) + \alpha \nabla_{\mathbf{h}} y\left(\mathbf{h}^l\right)\right). \tag{9}$$

with μ^l being the adaptive step-size in iteration l. The gradient for $f_d(\mathbf{h})$ is computed as

$$\nabla_{\mathbf{h}} f_d(\mathbf{h}) = \left(\sum_{n=0}^{L_g-1} |g_d(n)|^{p_d}\right)^{\frac{1}{p_d}-1} \mathbf{C}^T \mathbf{b}_d \tag{10}$$

where

$$\mathbf{b}_d = \operatorname{diag}\{\operatorname{sign}\{\mathbf{g}_d\}\} \operatorname{diag}\{\mathbf{w}_d\} \cdot |\mathbf{g}_d|^{(p_d-1)}, \tag{11}$$

and $\mathbf{C}$ denotes the $L_g \times L_h$ convolution matrix made up of $c(n)$; the computation of $\nabla_{\mathbf{h}} f_u(\mathbf{h})$ and $\mathbf{b}_u$ is analogue.

Finally, the gradient for $f(\mathbf{h})$ reads:

$$\nabla_{\mathbf{h}} f(\mathbf{h}) = \frac{1}{f_u(\mathbf{h})} \nabla_{\mathbf{h}} f_u(\mathbf{h}) - \frac{1}{f_d(\mathbf{h})} \nabla_{\mathbf{h}} f_d(\mathbf{h}). \tag{12}$$

A more in-depth derivation of the gradient $\nabla_{\mathbf{h}} f(\mathbf{h})$ is given in [9].

3.1 Regularization based on p-Norm

By choosing p_f to be finite, (8) reads

$$y(\mathbf{h}) = \left(\sum_{k=0}^{L_a-1} |A(k)|^{p_f} \right)^{\frac{1}{p_f}} \tag{13}$$

where $A(k)$ is the DFT of $a(n)$, with discrete frequency index k. As the DFT can be expressed by a matrix multiplication, the derivation of $\nabla_{\mathbf{h}} y(\mathbf{h})$ is quite similar to that of $\nabla_{\mathbf{h}} f_d(\mathbf{h})$ and $\nabla_{\mathbf{h}} f_u(\mathbf{h})$.

Let $\mathbf{F}$ be the matrix for the DFT of compatible size, then $\mathbf{a}_f = \mathbf{FTh}$, where $\mathbf{T}$ is either $\mathbf{C}$, for the case one aims to optimize the frequency response of the global impulse response $g(n)$, or the identity matrix, if one aims to optimize only the frequency response of the equalizer $h(n)$. The gradient of $y(\mathbf{h})$ reads

$$\nabla_{\mathbf{h}} y(\mathbf{h}) = \Re \left\{ \left(\sum_{k=0}^{L_a-1} |A(k)|^{p_f} \right)^{\frac{1}{p_f}-1} (\mathbf{FT})^H \mathbf{b}_f \right\} \tag{14}$$

with

$$\mathbf{b}_f = \operatorname{diag}\left\{ \operatorname{sign}\left\{ \mathbf{a}_f \right\} \right\} \cdot \left| \mathbf{a}_f \right|^{(p_f-1)}. \tag{15}$$

3.2 Regularization based on Infinity-Norm

By choosing $p_f = \infty$, the regularization term becomes

$$y(\mathbf{h}) = \left\| \mathbf{a}_f \right\|_\infty = \max\left\{ \left| \mathbf{a}_f \right| \right\}. \tag{16}$$

With I_m being the index of the maximum entry of $\left|\mathbf{a}_f\right|$ we define

$$A_m(k) = \begin{cases} A(k), & \text{for } k = I_m, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

A vector $\mathbf{a}_m$ is made up by $A_m(k)$.

The gradient of $y(\mathbf{h})$ then reads

$$\nabla_{\mathbf{h}} y(\mathbf{h}) = \Re \left\{ (\mathbf{FT})^H \operatorname{sign}\{\mathbf{a}_m\} \right\}. \tag{18}$$

Due to the special structure of the matrix $\mathbf{T}$ (which is either the convolution matrix $\mathbf{C}$ or the identity matrix) and the DFT-matrix $\mathbf{F}$, many of the computations can be performed efficiently in the frequency-domain by utilizing the fast Fourier transform and its inverse.

4. SIMULATIONS

To show the effect of the proposed method, we applied it to a simulated RIR of length $L_c = 2000$ taps and a measured RIR which was estimated with a length of $L_c = 24000$ taps, both sampled at a rate of $f_s = 16$ kHz.

For the simulation, the reverberation time was set to $\tau_{60} = 200$ ms. Figure 1(a) shows the simulated RIR, and Figure 1(b) depicts its reshaped version and the corresponding frequency response. The filter design was carried out according to the method from [9], with $p_d = p_u = 10$ and $\mathbf{w}_d$ and $\mathbf{w}_u$ chosen as in (4) and (5). As one can see from the plots, both the obtained decay and the frequency response appear acceptable.

Figure 2(a) shows the measured RIR, which is much longer than the first one. The result of the reshaping based

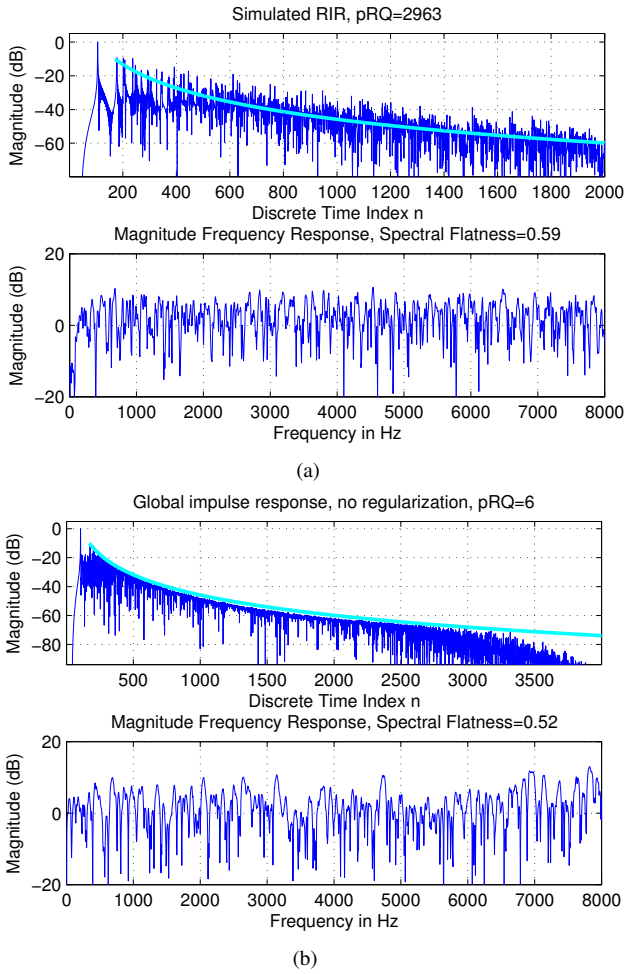


Figure 1: Example of reshaping a simulated RIR. (a) The simulated RIR and its frequency response. (b) The reshaped RIR in the time and frequency domain using the method from [9]. The exponentially descending curve is the average temporal masking limit.

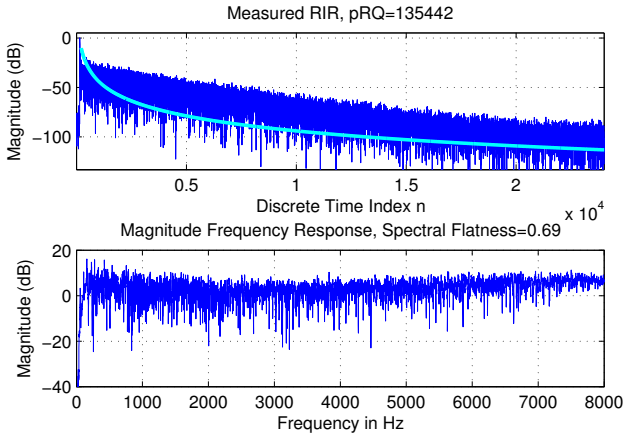
on the method from [9] is depicted in Figure 2(b). In this case, the time-domain reshaping works as desired, but the obtained frequency response shows a strong amplification of higher frequencies. The postfilter method from [4] could be used to equalize the frequency response to an extent that the spectral distortion becomes hardly audible while the RIR decay remains almost as desired. However, in this paper we aim at giving a direct method to obtain the desired time- and frequency-domain characteristics.

To quantify the spectral distortion, we utilize the *spectral flatness measure* (SF), that is defined as

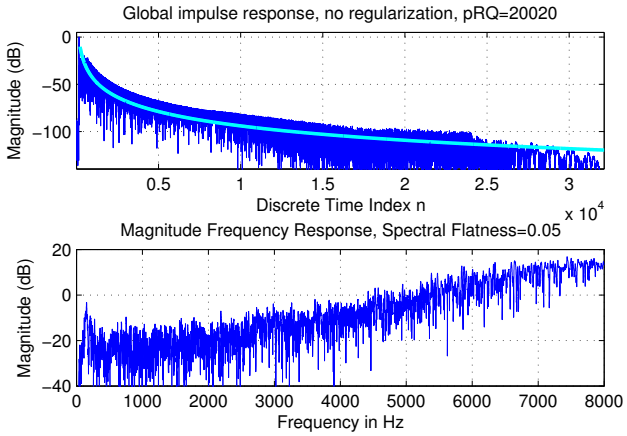
$$\mathrm{SF} = \frac{\sqrt[L_g]{\prod_{k=0}^{L_g-1} |G(k)|^2}}{\frac{1}{L_g} \sum_{k=0}^{L_g-1} |G(k)|^2} \tag{19}$$

with $G(k)$ being the frequency response of the GIR [3].

The reason we define the masking window $w_u(n)$ is that we take $\frac{1}{w_u(n)}$ as the average temporal masking limit. If the RIR or the reshaped RIR exceeds this limit, it implies that audible reverberation exists. For an accurate quantitative description, we propose the *perceivable reverberation quanti-*

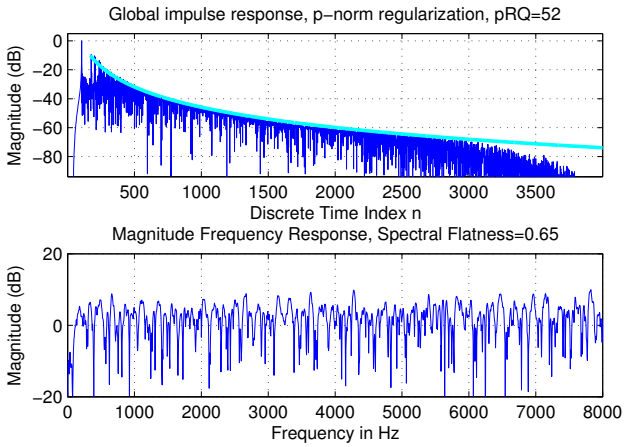


(a)

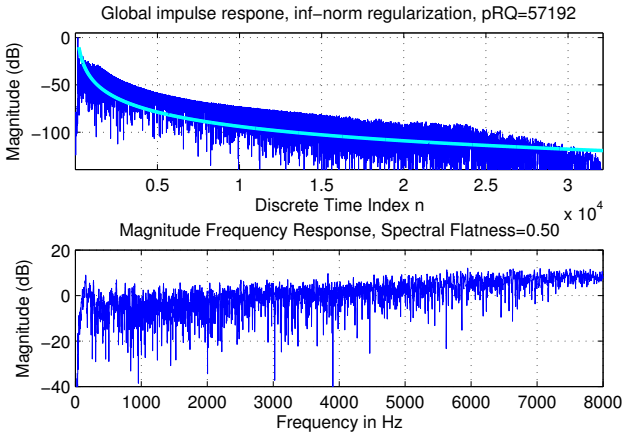


(b)

Figure 2: Example of reshaping a measured RIR. (a) The RIR and its frequency response. (b) The reshaped RIR in the time and frequency domain using the method from [9]. The exponentially descending curve is the average temporal masking limit.



(a)



(b)

Figure 3: Examples of reshaping RIRs while constraining the frequency response of the equalizer. (a) Reshaping result for the simulated RIR ($a(n) = h(n)$, $p_f = 8$, $\alpha = 0.05$). (b) Result of reshaping the measured RIR with the proposed approach ($a(n) = h(n)$, $p_f = \infty$, $\alpha = 0.14$).

zation measure (pRQ), that we define as

$$\mathrm{pRQ} = \sum_{n=N_0}^{L_g-1} g_E(n), \tag{20}$$

with $g_E(n)$ being the amount of energy that exceeds the temporal masking limit on the logarithmic scale and that is above -60 dB compared to the direct sound:

$$g_{\mathrm{E}}(n) = \begin{cases} 20\log_{10}(|g(n)|\,w_u(n)), & \text{for } |g(n)| > \frac{1}{w_u(n)}, \\ & |g(n)| > -60\mathrm{dB}, \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

This dimensionless measure combines the psychoacoustic measures of the reverberation time τ_{60} of a room and the compromise temporal masking limit, to quantify the amount of perceivable reverberation. If the RIR is completely reshaped, then either no time coefficient exceeds the temporal masking limit or the energy of each of the exceeding coefficients is below -60 dB; in both cases $\mathrm{pRQ} = 0$.

By applying the postfiltering method from [4] we could achieve $\mathrm{pRQ} = 649$, $\mathrm{SF} = 0.71$ for the simulated RIR and $\mathrm{pRQ} = 73232$, $\mathrm{SF} = 0.48$ for the measured RIR.

For the first experiments, we chose $a(n) = h(n)$, so we jointly optimized the frequency response of the equalizer and the time-domain representation of the GIR. In Figure 3(a) the result of applying the proposed approach to the simulated RIR is depicted. In some preliminary simulations it turned out, that $p_f = 8$ is a sufficiently large value to reduce spectral peaks of $a(n)$ in this case.

By setting $p_f = 8$ and $\alpha = 0.05$ the pRQ enhances from 2963 to 52, while the SF could be slightly enhanced from 0.59 to 0.65. Figure 3(b) shows the results of applying the proposed approach to the measured RIR. The parameters were chosen as $p_f = \infty$ and $\alpha = 0.14$; the pRQ could be enhanced from 135442 to 57192, while the SF drops from 0.69 to 0.5.

For the second set of experiments, we chose $a(n) = g(n)$, so we jointly optimized the frequency- and time-domain representations of the global impulse response. In Figure 4(a) the result of applying the proposed approach to the simulated RIR is depicted. By setting $p_f = 8$ and $\alpha = 0.4$ the pRQ enhances from 2963 to 18, while the SF could be enhanced from 0.59 to 0.71. Figure 4(b) shows the results of applying the proposed approach to the measured RIR. The parameters were chosen as $p_f = \infty$ and $\alpha = 0.8$; the pRQ could be en-

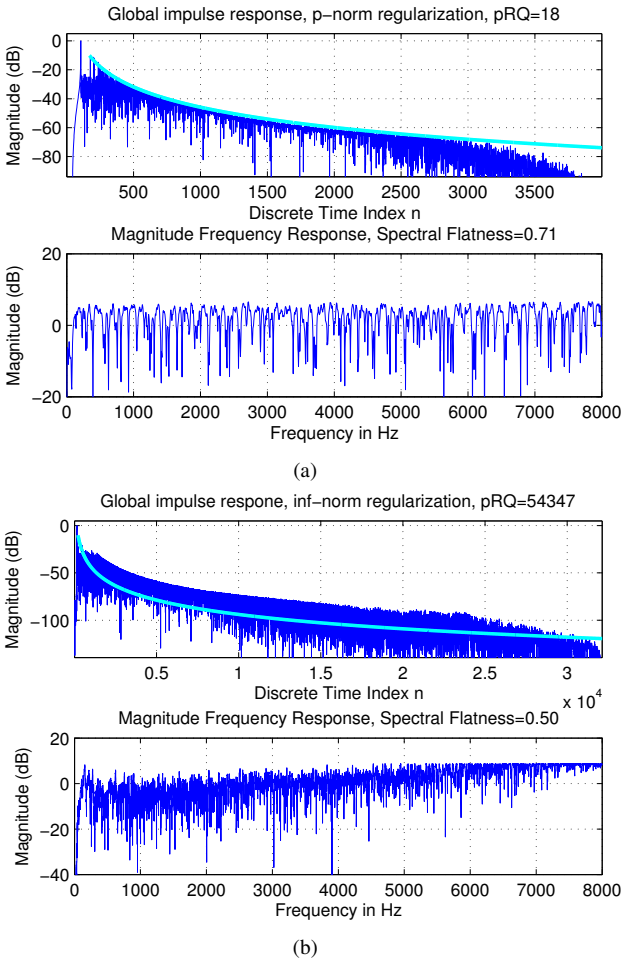


Figure 4: Examples of reshaping RIRs while constraining the frequency response of the global system. (a) Reshaping result for the simulated RIR ($a(n) = g(n)$, $p_f = 8$, $\alpha = 0.4$). (b) Result of reshaping the measured RIR with the proposed approach ($a(n) = g(n)$, $p_f = \infty$, $\alpha = 0.8$).

hanced from 135442 to 54347, while the SF drops from 0.69 to 0.5.

Comparing our proposed approach with the method from [4], it could be shown that our approach outperforms the postfiltering method: for the simulated RIR we could get a SF of 0.71 but the pRQ could be reduced to 18 with the new approach instead of 649; we achieved analogue results for the measured RIR.

Concerning the regularization method, we presented two approaches; circumventing high peaks in the frequency response of the equalizer on the one hand and of the GIR on the other hand. While expressing different aims, simulations showed that both approaches work well and yield comparable results, whereat constraining the GIR performs slightly better (according to the pRQ and SF measures).

The value for p_f has been chosen with respect to some preliminary simulations. The choice of α determines the tradeoff between the requirements on the frequency response and a *good* reshaping; its values have been found empirically. An in-depth investigation of the linking between the RIR, the choice of p_f and the regularization factor α will be done in future work.

5. CONCLUSIONS

In this paper, we proposed to jointly optimize time-domain and frequency-domain based criteria. The time-domain based objective function from [9] was extended by a p-norm based regularization term in the frequency-domain. Simulations showed that our approach can attenuate spectral distortions in the global impulse response at the expense of a slightly degraded dereverberation performance. The amount of dereverberation was measured by the newly introduced pRQ measure. Informal listening tests have confirmed the characteristics of the global impulse response that could be expected from the pRQ and SF measures, namely dereverberation without spectral distortions.

We found that our new approach outperforms the alternative way of first reshaping and then compensating for spectral distortions in an additional postprocessing step, as it has been proposed in [4]. Future work will be directed toward investigating and improving spatial robustness in real-world settings. Besides that, we will investigate the influence of the choice of p_f and also aim for an automatic determination of the regularization factor α. Additional listening tests will be done to further support the use of the newly introduced pRQ measure.

REFERENCES

[1] S. J. Elliott and P. A. Nelson. Multiple-point equalization in a room using adaptive digital filters. *Journal of the Audio Engineering Society*, 37(11):899–907, Nov. 1989.

[2] L. D. Fielder. Practical limits for room equalization. In *AES 111th Conv.*, pages 1–19, 2001.

[3] J. D. Johnston. Transform coding of audio signals using perceptual noise criteria. 6(2):314–323, 1988.

[4] M. Kallinger and A. Mertins. Room impulse response shortening by channel shortening concepts. In *Proc. Conf Signals, Systems and Computers Record of the Thirty-Ninth Asilomar Conf*, pages 898–902, 2005.

[5] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante. Fast deconvolution of multichannel systems using regularization. 6(2):189–194, Mar. 1998.

[6] H. Kuttruff. *Room Acoustics*. Spoon Press, London, 2000.

[7] R. K. Martin, D. Ming, B. L. Evans, and C. R. Johnson Jr. Efficient channel shortening equalizer design. *Journal on Applied Signal Processing*, 13:1279–1290, Dec. 2003.

[8] P. J. W. Melsa, R. C. Younce, and C. E. Rohrs. Impulse response shortening for discrete multitone transceivers. 44(12):1662–1672, 1996.

[9] 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*, 18(2):249–259, 2010.

[10] J. N. Mourjopoulos. Digital equalization of room acoustics. *Journal of the Audio Engineering Society*, 42(11):884–900, Nov. 1994.