Combined Acoustic MIMO Channel Crosstalk Cancellation and Room Impulse Response Reshaping

Jan Ole Jungmann, *Student Member, IEEE*, Radoslaw Mazur, *Member, IEEE*, Markus Kallinger, *Member, IEEE*, Tiemin Mei, and Alfred Mertins, *Senior Member, IEEE*

***Abstract*—Virtual 3-D sound can be easily delivered to a listener by binaural audio signals that are reproduced via headphones, which guarantees that only the correct signals reach the corresponding ears. Reproducing the binaural audio signal by two or more loudspeakers introduces the problems of crosstalk on the one hand, and, of reverberation on the other hand. In crosstalk cancellation, the audio signals are fed through a network of prefilters prior to loudspeaker reproduction to ensure that only the designated signal reaches the corresponding ear of the listener. Since room impulse responses are very sensitive to spatial mismatch, and since listeners might slightly move while listening, robust designs are needed. In this paper, we present a method that jointly handles the three problems of crosstalk, reverberation reduction, and spatial robustness with respect to varying listening positions for one or more binaural source signals and multiple listeners. The proposed method is based on a multichannel room impulse response reshaping approach by optimizing a p-norm based criterion. Replacing the well-known least-squares technique by a p-norm based method employing a large value for p allows us to explicitly control the amount of crosstalk and to *shape* the remaining reverberation effects according to a desired decay.**

***Index Terms*—Crosstalk cancellation, optimization, room impulse response (RIR) reshaping, spatial robustness.**

I. Introduction

THREE-DIMENSIONAL audio reproduction with loudspeakers in a room can be achieved by using a prefilter network that processes the binaural source signals prior loudspeaker reproduction in such a way that the individual signals arrive only at the designated ears of the listener, or even at the designated ears of multiple listeners. Thus, all acoustic crosstalk need to be cancelled out. To keep up the perceived quality of the audio signal, no spectral distortion or reverberation should be introduced along the signal paths. Early approaches assumed symmetric propagation paths and aimed at the equalization of head related transfer functions (HRTFs) and the cancellation of crosstalk [1]. Later designs considered the individual transmission paths from the loudspeakers to the ears and tried to tackle the above mentioned equalization problem in more detail, as described follows.

Signal propagation from N loudspeakers to M listener ears can be expressed by a network of $M \times N$ system functions $C_{m\ell}(z)$ that describe the transmission from loudspeaker ℓ to ear m. Given Q sources, a preprocessing network can be defined by another set of $N \times Q$ system functions $H_{\ell q}(z)$ which determine the transmission from source q to loudspeaker ℓ. The concatenation of the prefilter network and the acoustic multichannel system yields a global (overall) system with Q inputs and M outputs. The system functions of the global system will be denoted by $G_{mq}(z)$ with $q = 1, 2, \ldots, Q$ and $m = 1, 2, \ldots, M$ in the following. An ideal prefilter network would lead to system functions $G_{mq}(z)$ that are equal to one (or to a delay term z^{-n_0} with some delay of n_0 samples) for desired signal paths and zero for undesired ones. It is relatively straightforward to achieve the goal of perfect crosstalk cancellation (i.e., making all undesired paths equal to zero), as this is algebraically related to forming the adjugate of a matrix of system functions. However, it is very demanding to obtain perfect equalization for the desired paths (even with some delay), because this requires the inversion of systems that typically have many zeros on or close to the unit circle of the z-plane [2]–[4].

Nelson *et al.* [5] proposed a least-squares design that aimed to achieve both, equalization and crosstalk cancellation in one step. This method has been extended by Ward [6], who simultaneously considered multiple head positions in order to increase spatial robustness. Kallinger and Mertins [7] proposed a spatially robust least-squares method by considering perturbations of the measured systems based on statistical knowledge [8] of the acoustic transfer functions inside a closed room.

The above-mentioned problem of system inversion for the desired paths is similar to the equalization of room impulse responses (RIR) in the single-channel case, which is usually applied to compensate for the undesired acoustic properties of a closed room, namely reverberation. Early approaches for the inverse filtering of room acoustics [9] decomposed mixed-phase systems into allpass and minimum-phase components and used IIR filters for the inversion of the minimum-phase part. Other methods minimize the mean squared error (MSE) between the output of a desired target system and the concatenation of RIR and equalizer [3], [10]. Although aiming at perfect equalization is quite intuitive and straightforward, practical problems arise when the channel has zeros very close to, or even on the unit circle of the z-plane. In data transmission, the method of

Manuscript received November 22, 2011; revised February 15, 2012; accepted February 16, 2012. Date of publication March 14, 2012; date of current version April 11, 2012. M. Kallinger contributed to this work while he was with the University of Oldenburg. This work was supported by the German Research Foundation under Grant ME1170/3-1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lauri Savioja.

J. O. Jungmann, R. Mazur, and A. Mertins are with the Institute for Signal Processing, University of Lübeck, Lübeck 23562 Germany (e-mail: jungmann@isip.uni-luebeck.de; mazur@isip.uni-luebeck.de; mertins@isip.uni-luebeck.de).

M. Kallinger is with the Fraunhofer Institute for Integrated Circuits IIS, 91058 Erlangen, Germany.

T. Mei is with School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110168 China (e-mail:meitiemin@163.com).

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/TASL.2012.2190929

channel shortening instead of equalization has been introduced for such critical channels. It has originally been proposed by Falconer and Magee to reduce the implementation cost of maximum-likelihood detection via the Viterbi algorithm [11] and is now widely used in orthogonal frequency division multiplex (OFDM) and discrete multitone (DMT) systems to reduce the effective channel order to the length of the guard interval [12], [13]. For listening room compensation, this concept has first been proposed in [14] and has now also been used for post-filtering of microphone signals [15], [16]. In acoustic channel shortening, one does not try to recover the exact source signal but instead to concentrate the energy of the overall impulse response within a certain time frame after the direct sound and thus to maximize, for example, the D50 measure [17]. The D50 measure is a psychoacoustically motivated measure that is defined as the ratio of the energy within the first 50 ms beginning with the direct sound pulse to the energy of the whole impulse response. In [18], the D50-based least-squares method for shortening has been replaced by an infinity-norm criterion that yields much better control of late reverberation. A variation of the channel shortening concept, called channel shaping or reshaping, has first been introduced in [19] with the aim to shape the reverberation tail in a predefined manner. In [20], the least-squares optimality criterion from [19] was generalized by formulating a p-norm based one which allows for a better control of the decay behavior of the obtained global impulse response (GIR) over its full length. Specifically, the decay was shaped according to the temporal masking property of the human auditory system. The masking property means that reverberation is not audible if it remains below a certain limit that is induced by the direct sound [21]. While the exact masking limit is signal dependent and difficult to obtain, a compromise masking limit, found as an average over several stimuli and conditions, has been proposed in [22] and was used for the filter design in [20]. This made it possible to shape impulse responses in such a way that the reverberation tail strictly stays under the average temporal masking limit and, for many signals, no reverberation is audible.

In this paper, we extend the impulse-response shaping method from [20], [23]–[25] to the design of robust crosstalk cancellers that keep control of the amount of crosstalk that occurs due to small head movements and the audibility of spectral distortions and reverberation. Robustness is achieved by two different approaches. In the first one, we consider the design of a set of prefilters that jointly reshape the global impulse responses for multiple positions in a finite area. This method is an extension of the work in [24], which only considered the common setup with two loudspeakers and two ears. In the second method, we incorporate statistical channel knowledge as in [7] into the design of acoustic MIMO systems. While the prefilter design for multiple positions requires the knowledge of multiple realizations of the channel impulse responses and is computationally expensive, the incorporation of statistical knowledge is an effective extension of the equalization for the reference positions that requires the knowledge of only a single set of RIRs.

This paper is organized as follows. The problem of crosstalk cancellation (CTC) itself is described in Section II, and the theory of spatial sampling of room impulse responses and introducing system perturbations is described in Section III. The proposed CTC design methods are described in Section IV. In Section V we present the results of the experiments with simulated and measured data. Finally, we close this paper with some conclusions in Section VI.

1) Notation: Lowercase boldface characters denote vectors, while uppercase boldface characters denote matrices. The superscripts ^{T} and * denote transposition and complex conjugation, respectively. 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, $\max\{\cdot\}$ returns the maximum component of its input vector and $\mathrm{sign}\{\cdot\}$ produces a sign vector of its input variable, whereat the sign of a complex number is defined as its projection on the unit circle of the complex plane. Finally, $\Re\{\cdot\}$ captures the real part of the input variable, and $\mathrm{E}\{\cdot\}$ denotes the expected value. The lengths of FIR filters are denoted as L_c and L_h for filters $c(n)$ and $h(n)$, respectively.

II. Crosstalk-Canceller Design

In the following, the crosstalk canceller will be described for a number of Q source channels, N loudspeakers, and M microphones, as depicted in Fig. 1. The presented approach is valid for arbitrary setups, however in practice only configurations with $Q \leq M \leq N$ are relevant. The common setup for crosstalk cancellation consists of just two loudspeakers and two microphones and is, of course, covered by the more general setup considered here. The corresponding formulations can be established by choosing $Q = N = M = 2$. To keep the description close to the existing literature, the problem is formulated in terms of linear systems of equations in this section. Other formulations will be given in later sections when required. For each of the M defined microphone positions, perturbations due to spatial movement are considered. These are alternatively introduced in a statistical manner or by introducing R channel realizations sampled in the vicinity of the reference position.

Assuming FIR filters and system functions $C_{m\ell}(z)$ and $H_{\ell q}(z)$, respectively, the global system functions are given by

$$G_{mq}(z) = \sum_{\ell=1}^{N} C_{m\ell}(z) H_{\ell q}(z),\ q = 1, \ldots, Q,\ m = 1, \ldots, M \tag{1}$$

with $G_{mq}(z)$ denoting the acoustic channel from source q to microphone m, Q being the number of source channels and M being the number of microphones.

The global impulse responses are grouped into *wanted* and *unwanted* (i.e., *crosstalk*) signal paths. For every microphone position m we can define whether we want a specific source q to reach the destination or if it should be handled as undesired crosstalk.

As the prefilters can be designed independently for each of the Q source signals, the following derivations are made with the assumption that just one source signal is active. For Q sources, one would end up designing Q individual sets of prefilters. For *perfect* crosstalk cancellation, the prefilters are designed in such a way that the transmission through the crosstalk channels is suppressed (i.e., $G_{mq}(z) = 0$) and that no audible distortions

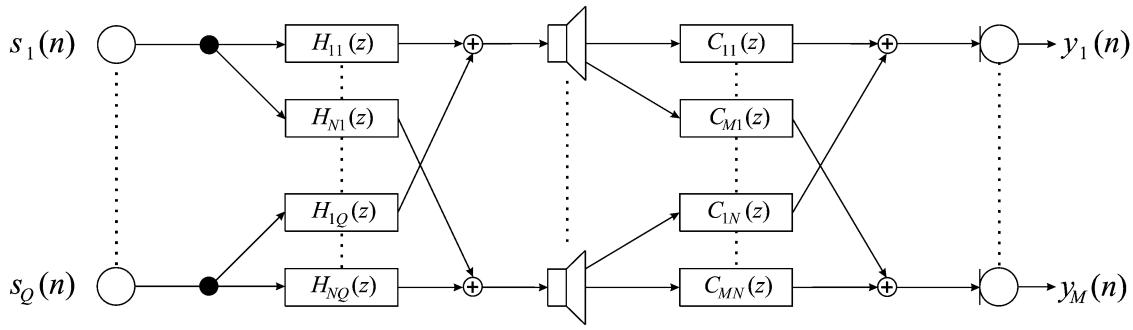

Fig. 1. Setup with Q sources, N loudspeakers, and M microphones. For every microphone, we have perturbations, possibly sampled at R locations, including the reference position.

are introduced by the desired signal paths. Ideal transmission means that the paths for the desired components yield

$$G_{mq}(z) \approx D_{mq}(z) \tag{2}$$

where $D_{mq}(z)$ is a desired target system, that is to be approximated by the corresponding global impulse response. Usually, the target system is chosen as a bandpass system that accounts for the bandpass characteristic of a typical loudspeaker, has an appropriate delay (for example, it has to take the delays by the system $C_{m\ell}(z)$ into account), and does not introduce any audible distortions. When using just a delayed discrete pulse instead of a bandpass, the prefilters will particularly amplify those frequencies that are outside the loudspeakers' range of operation.

By assuming that all involved systems are FIR systems, representing the impulse responses $h_{\ell q}(n)$ and $d_{mq}(n)$ by vectors $\mathbf{h}_{\ell q}$ and $\mathbf{d}_{mq}$, respectively, and denoting the $(L_c + L_h - 1) \times L_h$ dimensional convolution matrices constructed from the individual room impulse responses $c_{m\ell}(n)$ as $\mathbf{C}_{m\ell}$, we can express the general problem as

$$\mathbf{C}\mathbf{h}_q = \mathbf{d}_q \tag{3}$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1N} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{M1} & \cdots & \mathbf{C}_{MN} \end{bmatrix}, \ \mathbf{h}_q = \begin{bmatrix} \mathbf{h}_{1q} \\ \vdots \\ \mathbf{h}_{Nq} \end{bmatrix}, \ \mathbf{d}_q = \begin{bmatrix} \mathbf{d}_{1q} \\ \vdots \\ \mathbf{d}_{Mq} \end{bmatrix} \tag{4}$$

with $\mathbf{d}_{mq}$ either being the zero vector (in cases where $G_{mq}(z)$ is a crosstalk path) or $\mathbf{d}_{mq}$ being a desired target system response when m is the desired listening position for source q. Since predefining $\mathbf{d}_q$ can be problematic, other avenues such as filter shortening and filter shaping have been explored, as described in the introduction. However, given a useful set of desired impulse responses $\mathbf{d}_q$, a classical way would be to solve (3) for $\mathbf{h}_q$ in the sense of least squares by utilizing the Moore–Penrose pseudoinverse of the channel matrix $\mathbf{C}$. According to the multichannel inversion theorem (MINT) [26], even exact solutions are possible when the number of loudspeakers is sufficiently large. However, when spatial robustness is desired and system perturbations are considered during the design, perfect multichannel inversion cannot be achieved.

In order to increase the robustness of the equalizers against small spatial movements, statistical knowledge about acoustic transfer functions in a closed room can be integrated as a perturbation system into the CTC setup [7], [8]. Mathematically, the perturbation, which results from moving the microphone away from its reference position by some distance D, can be expressed as an additive error term $p_{m\ell}(n)$ on the RIR $c_{m\ell}(n)$ from loudspeaker ℓ to microphone m. With $\mathbf{P}_{m\ell}$ being the convolution matrix made up by a sequence $p_{m\ell}(n)$, (3) can be reformulated to yield prefilters which take into account the stochastic perturbations as follows:

$$(\mathbf{C} + \mathbf{P})\mathbf{h}_q = \mathbf{d}_q, \quad \text{where} \quad \mathbf{P} = \begin{bmatrix} \mathbf{P}_{11} & \cdots & \mathbf{P}_{1N} \\ \vdots & \ddots & \vdots \\ \mathbf{P}_{M1} & \cdots & \mathbf{P}_{MN} \end{bmatrix}. \tag{5}$$

An alternative approach to improve the spatial robustness is to design the equalizers to consider R different locations of each microphone jointly. Similar to [6], the linear system of (3) can be reformulated as

$$\begin{bmatrix} \mathbf{C}^{(1)} \\ \vdots \\ \mathbf{C}^{(R)} \end{bmatrix} \mathbf{h}_q = \begin{bmatrix} \mathbf{d}_q \\ \vdots \\ \mathbf{d}_q \end{bmatrix} \tag{6}$$

where $\mathbf{C}^{(r)}$ denotes the rth realization of the channel matrix defined in (4). Both approaches will be considered in the next sections together with p-norm-based objective functions.

III. Spatial Sampling and System Perturbations

In this section, we will first describe the spatial sampling theorem and its implications on the behavior of system responses for arbitrary microphone locations within limited listening volumes. Then we will describe the changes of the acoustic channels due to microphone displacement in terms of stochastic perturbations.

A. Spatio-Temporal Sampling Principle of RIRs

In a listening room, the continuous-time RIR from one point $[x', y', z']$ to another point $[x, y, z]$ is denoted by $c_{x'y'z'xyz}(t)$. Throughout this section, we consider $[x', y', z']$ to be the position of the loudspeaker and $[x, y, z]$ to be the position of the microphone. The spatial coordinate z is not to be confused with the parameter of the z-transform.

Room impulse responses are not only functions of time but also heavily rely on the spatial positions of the speaker and/or microphone. For a given pair of loudspeaker and microphone positions, the time-domain sampling rate should be equal to or

higher than two times the highest frequency appearing in the considered signals, denoted by f_0. From the point of view of wave equations and for a given position of the loudspeaker, $c_{x'y'z'xyz}(t)$ is a band-limited space function for any time instance t. If $c_{x'y'z'xyz}(t)$ is band-limited to f_0 in the time-domain, 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 must satisfy [27]

$$f_{s.s.} \geq 2f_0/v. \tag{7}$$

In general we use

$$f_{s.s.} > f_{t.s}/v \tag{8}$$

where $f_{t.s}$ denotes the time-domain sampling frequency.

The space-domain sampling can generally be achieved in two ways by either moving a single microphone sequentially from one position to another inside the listening area, while keeping the loudspeaker at a fixed position outside the listening area or vice versa. Alternatively, an array of multiple microphones can be used to counteract the time-consuming process of sequential measures.

Denoting the discretized RIR by $c_{ijklmn}(t)$, with $[i,j,k]$ and $[l,m,n]$ being the positions of the loudspeaker and the microphone inside the listening area, respectively, we sample the whole listening area for each loudspeaker. If the spatial sampling condition (7) is satisfied, then the RIR caused by the loudspeaker located at $[i,j,k]$ can be reconstructed for any point $[x,y,z]$ inside the listening area by interpolation [23], [27]:

$$c_{ijkxyz}(t) = \sum_{l,m,n} \phi(x,y,z,l,m,n) \cdot c_{ijklmn}(t) \tag{9}$$

where the interpolating function is

$$\phi(x,y,z,l,m,n) \equiv \phi(x - l\Delta x, y - m\Delta y, z - n\Delta z) \tag{10}$$

for convenience. Δx, Δy and Δz are the spatial sampling periods in the three spatial dimensions. The interpolation function $\phi(x,y,z)$ is independent of the loudspeaker positions. For the three-dimensional case, the so-called sampling function can be used as an interpolating function [23]:

$$\phi(x,y,z) = \operatorname{sinc}\left(\frac{\pi x}{\Delta x}\right) \cdot \operatorname{sinc}\left(\frac{\pi y}{\Delta y}\right) \cdot \operatorname{sinc}\left(\frac{\pi z}{\Delta z}\right) \tag{11}$$

where $\operatorname{sinc}(a) = \sin(a)/a$. The frequency supporting domain is, in the three-dimensional case, a cube $([-f_{s.s.}/2, f_{s.s.}/2])^3$.

B. Basis for Spatially Robust Reshaping

Let $h_{ijk}(t)$ denote the prefilter for the loudspeaker at position $[i,j,k]$, then the global impulse response at position $[x,y,z]$ in the listening area is given by

$$\begin{aligned} g_{xyz}(t) &= \sum_{i,j,k} h_{ijk}(t) * c_{ijkxyz}(t) \\ &= \sum_{l,m,n} g_{lmn}(t) \cdot \phi(x,y,z,l,m,n) \end{aligned} \tag{12}$$

where $g_{lmn(t)} = \sum_{i,j,k} h_{ijk}(t) * c_{ijklmn}(t)$ is the overall RIR at position $[l,m,n]$ in the listening area.

For the investigation of the spatial robustness of RIR equalization, the spatial characteristics of RIRs must be taken into account [23]. The equalized RIRs should satisfy the following hypotheses:

1) For a given time instant t, $g_{lmn}(t)$ is a spatially stationary field or can at least be approximated as a stationary field with negligible error, which means that $\mathrm{E}\{g_{lmn}(t)\} = \mathrm{const}(t)$, and the correlation function

$$r_g(t,l,m,n,i,j,k) = \mathrm{E}\{g_{lmn}(t) g_{ijk}(t)\}$$

depends just on the differences $l-i$, $m-j$ and $n-k$, that is

$$\mathrm{E}\{g_{lmn}(t) g_{ijk}(t)\} = r_g(t, l-i, m-j, n-k).$$

2) The magnitude $|g_{lmn}(t)|$ is limited in the form

$$|g_{lmn}(t)| \leq g_0(t) \text{ for } t_1 < t < t_2$$

where $g_0(t) > 0$ is an upper bound. This implies that

$$r_g(t,0,0,0) = \mathrm{E}\{g_{lmn}^2(t)\} \leq \mathrm{E}\{g_0^2(t)\}, \quad t_1 < t < t_2.$$

A time interval (t_1, t_2) of interest for a desired transmission path would be, for example, from 4 ms after the main peak of $g_{lmn}(t)$ until $t \to \infty$, as this part of the impulse response is responsible for possibly audible reverberation [22]. In Section IV, we will treat the equalizer design problem by starting with a prescribed upper limit $g_0(t)$ for $t > t_1$ and try to find prefilter networks that push the global responses $g_{lmn}(t)$ under the limit $g_0(t)$. This process is referred to as reshaping the impulse response, rather than equalizing it.

Let us take a look at the statistical properties of the reshaped RIR at any given point in the listening area. If $x = l\Delta x$, $y = m\Delta y$, $z = n\Delta z$, then $g_{xyz}(t) = g_{lmn}(t)$. If $x \neq l\Delta x$, $y \neq m\Delta y$, $z \neq n\Delta z$, then $g_{xyz}(t)$ can be expressed as a linear combination of the sampled RIRs $g_{lmn}(t)$ as given in (12). The ensemble average of $g_{xyz}(t)$ is then

$$\begin{aligned} \mathrm{E}\{g_{xyz}(t)\} &= \sum_{l,m,n} \phi(x,y,z,l,m,n) \cdot \mathrm{E}\{g_{lmn}(t)\} \\ &= \mathrm{const}(t) \cdot \sum_{l,m,n} \phi(x,y,z,l,m,n). \end{aligned} \tag{13}$$

Considering the squared magnitudes of the global RIRs at a point in the listening area, we get

$$\begin{aligned} \mathrm{E}\{|g_{xyz}(t)|^2\} = \sum_{k_x,k_y,k_z=-\infty}^{\infty} & r_g(t,k_x,k_y,k_z) \\ & \cdot \Psi(x,y,z,k_x,k_y,k_z) \end{aligned} \tag{14}$$

where

$$\begin{aligned} \Psi(x,y,z,k_x,k_y,k_z) = \sum_{l,m,n} & \phi(x,y,z,l,m,n) \\ & \cdot \phi(x,y,z,l+k_x,m+k_y,n+k_z). \end{aligned} \tag{15}$$

It can be shown that the sampling function satisfies

$$\sum_{l,m,n=-\infty}^{\infty} \phi(x,y,z,l,m,n) = 1 \tag{16}$$

and

$$\Psi(x,y,z,k_x,k_y,k_z) = \begin{cases} 1, & \text{for } k_x,k_y,k_z = 0 \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

so we obtain $\mathrm{E}\{g_{xyz}(t)\} = \mathrm{const}(t)$ and

$$\mathrm{E}\left\{g_{xyz}^2(t)\right\} = r_g(t,0,0,0) \le \mathrm{E}\left\{g_0^2(t)\right\} \quad \text{for} \quad t_1 < t < t_2. \tag{18}$$

This means that for any given point $[x,y,z]$ inside the listening area, the global impulse response $g_{xyz}(t)$ will, on average, be limited by the same upper bound $g_0(t)$, by which all reshaped global impulse responses at the sampling points are limited. Thus, reshaping the impulse responses at the sampling points yields, on average, reshaping within the entire volume.

C. Robust Crosstalk Canceller for the General MIMO Case Using Statistical Knowledge

In this section, we extend the robust 2×2 crosstalk canceller design from Kallinger and Mertins [7] to the general MIMO case and introduce a way to incorporate arbitrary weighting for the reverberation tail. For that, we briefly recapitulate the statistical properties of RIRs in the case of spatial deviation from a reference point. The problem of designing an equalizer for a reference location and then moving the microphone away from this position has been studied by Radlović *et al.* [8]. They formulated the following conditions under which the transfer function between a loudspeaker and a microphone is a stochastic one:

- The dimensions of the room must be large compared to the wavelengths of interest. This is true especially for speech signals transmitted in typical office rooms.
- Statistical assumptions can be met for frequencies above the Schroeder large-room frequency

$$f_{\mathrm{SL}} = 2000 \cdot \sqrt{\frac{\tau_{60}}{V}} \mathrm{Hz} \tag{19}$$

  where V is the volume of the room and τ_{60} is the reverberation time of the room. For example, a room with dimension $4\ \mathrm{m} \times 5\ \mathrm{m} \times 2.5\ \mathrm{m}$ and $\tau_{60} = 400\ \mathrm{ms}$ has $f_{SL} \approx 180\ \mathrm{Hz}$.
- All loudspeakers and microphones should have a distance of at least half a wavelength to adjacent walls.

Given these assumptions, Radlović *et al.* [8] defined the following frequency-dependent measure to express the error due to the displacement of the microphone position

$$F(\omega) = \mathrm{E}\left\{\left|\left[C(\omega) + P(\omega)\right] H(\omega) - 1\right|^2\right\}. \tag{20}$$

Here, $C(\omega)$, $P(\omega)$, and $H(\omega)$ are the Fourier transforms of the continuous-time acoustic impulse response $c(t)$, its perturbation $p(t)$ due to spatial movement, and the equalizer $h(t)$, respectively; $\omega = 2\pi f$ is the radial frequency. Assuming a perfect equalizer $H(\omega) = 1/C(\omega)$ and being in the far field in reverberant environments, the distance measure amounts to [8]

$$F(\omega) = \frac{\mathrm{E}\left\{\left|P(\omega)\right|^2\right\}}{\left|C(\omega)\right|^2} = 2 - 2\frac{\sin(\omega D/v)}{\omega D/v} \tag{21}$$

where D is the deviation of the microphone position from its reference location in meters, and v is the sound-propagation velocity (340 m/s). Thus,

$$\mathrm{E}\left\{\left|P(\omega)\right|^2\right\} = \left|C(\omega)\right|^2 \left(2 - 2\frac{\sin(\omega D/v)}{\omega D/v}\right). \tag{22}$$

Assuming band limited input signals with maximum radial frequency ω_c and sampling with frequency $f_s = 1/T \ge \omega_c/\pi$, the continuous-time impulse responses $c(t)$, $p(t)$, and $h(t)$ are replaced by their discrete-time equivalents $c(n)$, $p(n)$, and $h(n)$, respectively. Based on (22), the discrete-time autocorrelation sequence of the perturbation becomes

$$r_{pp}(n) = r_{cc}^E(n) * f(n) \tag{23}$$

where $r_{cc}^E(n) = c(n) * c(-n)$ and

$$f(n) = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} \left(2 - 2\frac{\sin(\omega D/v)}{\omega D/v}\right) e^{j\omega nT} d\omega. \tag{24}$$

The sequence $f(n)$ can be easily computed with sufficient accuracy by sampling $F(\omega)$ at discrete frequencies $\omega_k = k\omega_c/K$, with $k = -K, -K+1, \ldots, K$ and using the inverse discrete Fourier transform.

Since we aim at shaping the overall impulse responses, we assign to every impulse response

$$g_{mq}(n) = \sum_{\ell=1}^{N} \left[c_{m\ell}(n) + p_{m\ell}(n)\right] * h_{\ell q}(n) \tag{25}$$

from source q to microphone m a weighting with positive weights $w_{mq}(n)$. Assuming perfect equalization for the reference point, we are interested in the error

$$e_{mq}(n) = w_{mq}(n) \left[\sum_{\ell=1}^{N} p_{m\ell}(n) * h_{\ell q}(n)\right] \tag{26}$$

and its average power due to microphone movement. For a specific combination of source, speaker and microphone, the weighted error is given by

$$\hat{e}_{mnq}(n) = w_{mq}(n) \cdot \left[p_{m\ell}(n) * h_{\ell q}(n)\right]. \tag{27}$$

By collecting the weights $w_{mq}(n)$ into vectors $\mathbf{w}_{mq}$, all weights for source q can be expressed by the diagonal weighting matrix

$$\widetilde{\mathbf{W}}_q = \mathrm{diag}\left\{\left[\mathbf{w}_{1q}^T, \mathbf{w}_{2q}^T, \ldots, \mathbf{w}_{Mq}^T\right]\right\}. \tag{28}$$

The weights for the individual paths are given by

$$\mathbf{W}_{mq} = \mathrm{diag}\left\{\mathbf{w}_{mq}\right\}. \tag{29}$$

The weighted error that results from moving the microphones away from their reference positions for source q is given by

$$\tilde{\mathbf{e}}_q = \widetilde{\mathbf{W}}_q \mathbf{P} \mathbf{h}_q. \tag{30}$$

We now consider the mean squared error due to spatial movement:

$$O_q = \mathrm{E}\left\{ \|\widetilde{\mathbf{W}}_q \mathbf{P} \mathbf{h}_q\|_2^2 \right\} = \mathrm{E}\left\{ \sum_{m=1}^{M} \sum_{\ell=1}^{N} \|\mathbf{W}_{mq} \mathbf{P}_{m\ell} \mathbf{h}_{\ell q}\|_2^2 \right\}. \tag{31}$$

To maximize robustness in the least-squares sense, O_q should be minimized.

Assuming that the perturbations for different acoustic paths are uncorrelated, the expression for O_q can be simplified to

$$O_q = \sum_{m=1}^{M} \sum_{\ell=1}^{N} O_{m\ell q} \tag{32}$$

with

$$\begin{aligned} O_{m\ell q} &= \mathrm{E}\left\{ \|\mathbf{W}_{mq} \mathbf{P}_{m\ell} \mathbf{h}_{\ell q}\|_2^2 \right\} \\ &= \mathrm{E}\left\{ \mathbf{h}_{\ell q}^T \mathbf{P}_{m\ell}^T \mathbf{W}_{mq}^T \mathbf{W}_{mq} \mathbf{P}_{m\ell} \mathbf{h}_{\ell q} \right\} \\ &= \mathrm{E}\left\{ \mathbf{p}_{m\ell}^T \mathbf{H}_{\ell q}^T \mathbf{W}_{mq}^T \mathbf{W}_{mq} \mathbf{H}_{\ell q} \mathbf{p}_{m\ell} \right\} \end{aligned} \tag{33}$$

where $\mathbf{H}_{\ell q}$ is the convolution matrix made up of the filter $\mathbf{h}_{\ell q}$. Using $\mathrm{tr}\{\mathbf{AB}\} = \mathrm{tr}\{\mathbf{BA}\}$, we obtain

$$\begin{aligned} O_{m\ell q} &= \mathrm{E}\left\{ \mathrm{tr}\left\{ \mathbf{H}_{\ell q}^T \mathbf{W}_{mq}^T \mathbf{W}_{mq} \mathbf{H}_{\ell q} \mathbf{p}_{m\ell} \mathbf{p}_{m\ell}^T \right\} \right\} \\ &= \mathrm{tr}\left\{ \mathbf{H}_{\ell q}^T \mathbf{W}_{mq}^T \mathbf{W}_{mq} \mathbf{H}_{\ell q} \mathbf{R}_{pp}^{(m\ell)} \right\} \end{aligned} \tag{34}$$

where $\mathbf{R}_{pp}^{(m\ell)} = \mathrm{E}\{\mathbf{p}_{m\ell} \mathbf{p}_{m\ell}^T\}$ is the correlation matrix for the system perturbation. Given an average displacement D and the impulse response $c_{m\ell}(n)$, it can be set up as a Toeplitz matrix from the autocorrelation sequence $r_{pp}^{(m\ell)}(n)$, which can be computed similar to $r_{pp}(n)$ in (23).

Now, considering a decomposition of $\mathbf{R}_{pp}^{(m\ell)}$ into $\mathbf{R}_{pp}^{(m\ell)} = \mathbf{L}^{(m\ell)} \mathbf{L}^{(m\ell)^T}$, which may be obtained via a Cholesky or a singular value decomposition, we may rewrite $O_{m\ell q}$ as

$$O_{m\ell q} = \mathrm{tr}\left\{ \mathbf{L}^{(m\ell)^T} \mathbf{H}_{\ell q}^T \mathbf{W}_{mq}^T \mathbf{W}_{mq} \mathbf{H}_{\ell q} \mathbf{L}^{(m\ell)} \right\}. \tag{35}$$

With $\mathbf{l}_i^{(m\ell)}$ being the ith column of matrix $\mathbf{L}^{m\ell}$, denoting the corresponding convolution matrix by $\mathbf{L}_i^{(m\ell)}$, and using $\mathbf{H}_{\ell q} \mathbf{l}_i^{(m\ell)} = \mathbf{L}_i^{(m\ell)} \mathbf{h}_{\ell q}$, we finally obtain

$$O_{m\ell q} = \mathbf{h}_{\ell q}^T \mathbf{M}_{m\ell} \mathbf{h}_{\ell q} \tag{36}$$

where

$$\mathbf{M}_{m\ell} = \sum_{i=1}^{L_c} \mathbf{L}_i^{(m\ell)^T} \mathbf{W}_{mq}^T \mathbf{W}_{mq} \mathbf{L}_i^{(m\ell)}. \tag{37}$$

Equations (32) and (36) represent explicit expressions to measure the average quadratic error for source q. They are easily differentiated with respect to the sought filter coefficients $\mathbf{h}_{\ell q}$ and can therefore be efficiently used during filter design as additional cost terms that support spatial robustness. It should be noted that the computation of (37) can be time-consuming for long GIRs.

IV. MIMO Crosstalk Cancellation and Impulse-Response Reshaping

The proposed design algorithm for MIMO crosstalk cancellation systems uses the p-norm based optimality criterion from [20] and extends it to multiple channels. As in [20], we take the average temporal masking threshold of the human auditory system into account and aim to push the reverberation tail under the masking limit.

It has been shown recently that it is necessary to also consider the frequency-domain representation of the global impulse responses or of the equalizers to circumvent for spectral distortions in the overall acoustic system [25]. For that purpose we adapt the p-norm based regularization term from [25] to the multichannel scenario, considered in this paper.

Since we are dealing with multichannel systems, we have to specify for each global impulse response $g_{mq}(n)$ whether it is a desired signal path or if it represents undesired crosstalk. Moreover, for the signal paths, there will be desired and unwanted parts of the impulse responses. The desired part of a GIR $g_{mq}(n)$ is denoted as

$$g_{mq}^{(d)}(n) = w_{mq}^{(d)}(n) \cdot g_{mq}(n) \tag{38}$$

whereas the unwanted part is denoted as

$$g_{mq}^{(u)}(n) = w_{mq}^{(u)}(n) \cdot g_{mq}(n). \tag{39}$$

If $g_{mq}(n)$ is a desired signal path, then the window $w_{mq}^{(d)}(n)$ cuts out the main peak of $g_{mq}(n)$ and the first few milliseconds after it, which corresponds to the direct sound path and some early reflections. For the unwanted part of a desired signal path, the window $w_{mq}^{(u)}(n)$ captures and weights the reverberation tail of $g_{mq}(n)$.

For a crosstalk path $g_{mq}(n)$, we have

$$w_{mq}^{(d)}(n) = 0 \tag{40}$$

as there is no *desired* part of the crosstalk component. The window $w_{mq}^{(u)}(n)$ for the unwanted part of a crosstalk path specifies the desired crosstalk attenuation and the shape of the crosstalk's reverberation tail.

To explain the choice of window functions, let us recall that the p-norm based approach in [20] was motivated by the fact that for $p \to \infty$, i.e., for $\|g_{mq}^{(u)}(n)\|_\infty = \max\{|g_{mq}^{(u)}(n)|\}$, the decay of $|g_{mq}^{(u)}(n)|$ is exactly determined by the window function $w_{mq}^{(u)}(n)$. This can be seen by considering the optimization problem

$$\underset{\mathbf{h}_q}{\text{minimize}}: \quad \log\left(\frac{\max_{m,n}\left\{ \left| g_{mq}^{(u)}(n) \right| \right\}}{\max_{m,n}\left\{ \left| g_{mq}^{(d)}(n) \right| \right\}} \right). \tag{41}$$

Similar to the stopband-behavior of FIR equiripple filter designs, the sequences $g_{mq}^{(u)}(n)$ will be limited by some constant

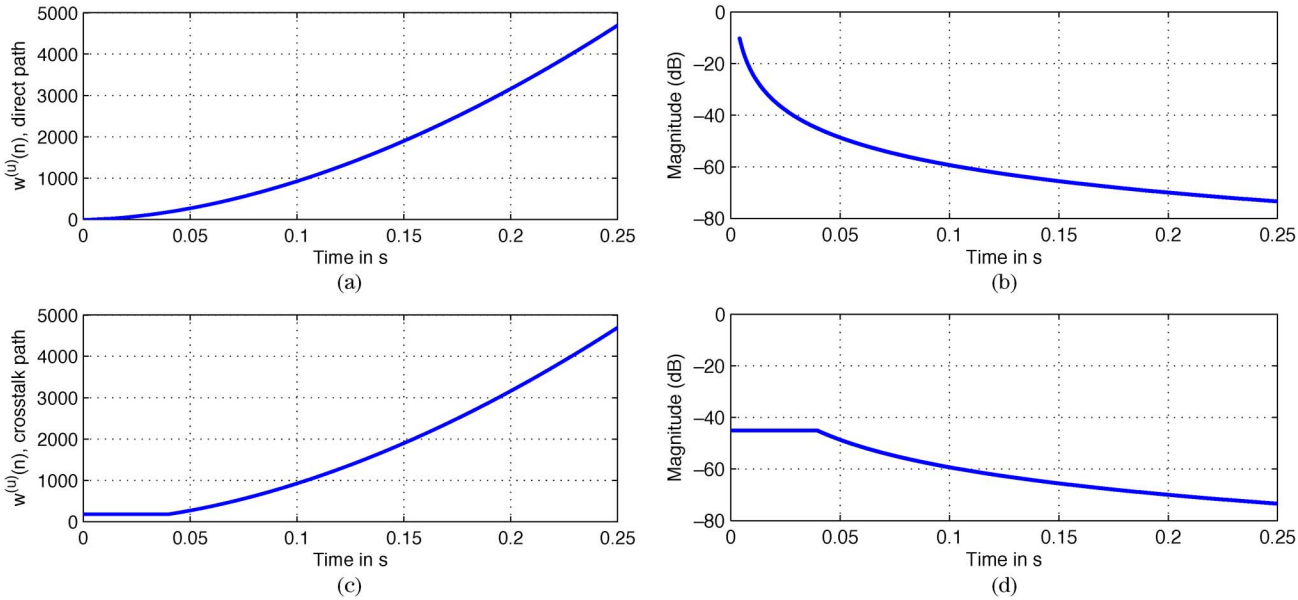

Fig. 2. Example of window functions. (a) The weighting window $l_m(n)$ for a signal path, plotted on a linear scale, and (b) its reciprocal, which approximates the temporal masking limit of the human auditory system, plotted on a logarithmic scale. (c) The corresponding weighting window for the crosstalk path with $w_0 = 180$, plotted on a linear scale, and (d) its reciprocal plotted on a logarithmic scale.

γ, and many values of $g_{mq}^{(u)}(n)$ will approach the limit γ. This means that

$$|g_{mq}(n)| \leq \gamma / w_{mq}^{(u)}(n) \quad \text{for} \quad n \geq n_1(m,q) \tag{42}$$

where we assume that the window $w_{mq}^{(u)}(n)$ is nonzero for $n \geq n_1(m,q)$. In other words, the unwanted part of $|g_{mq}(n)|$ is limited by the inverse of the window function $w_{mq}^{(u)}(n)$ times some constant γ. The sequence $\gamma / w_{mq}^{(u)}(n)$ is the equivalent to the limiting function $g_0(t)$ introduced in Section III. For signal paths, the indices $n_1(m,q)$ are chosen to represent about 4 ms after the main peak of $g_{mq}(n)$. For crosstalk paths, $n_1(m,q) = 0$.

Given the above considerations, the weighting function for the desired part of the direct signal path is defined as

$$w_{mq}^{(d)}(n) = \begin{cases} 1, & t_{0m} \cdot f_s \leq n < (t_{0m} + T_d) f_s \\ 0, & \text{otherwise} \end{cases} \tag{43}$$

where f_s is the sampling frequency, t_{0m} is the average time taken by the direct sounds from the N loudspeakers to the mth microphone, and T_d is chosen to be 4 ms.

In accordance with [20] we define the weighting window for the unwanted part of a direct signal path as

$$w_{mq}^{(u)}(n) = l_m(n) \tag{44}$$

with (45), shown at the bottom of the page, where $N_{1m} = t_{0m} \cdot f_s$, $N_2 = T_d \cdot f_s$, and $N_{0m} = (0.2\text{ s} + t_{0m}) f_s$. The reason why we define this window is that the term $1/l_m(n)$ approximates the average temporal masking limit of the human auditory system according to [22]. The masking curve starts with -10 dB at 4 ms after the direct sound impulse and then decays exponentially in the logarithmic domain to -70 dB at 200 ms after the direct pulse.

The weighting window for the unwanted part of a crosstalk path is defined as

$$w_{mq}^{(u)}(n) = \max\left\{w_0, \max_m \{l_m(n)\}\right\} \tag{46}$$

with $l_m(n)$ defined in (45) for the mth microphone. The value of w_0 directly captures the desired attenuation of the crosstalk component in comparison to the desired path. This can be seen using the same arguments as for the reverberation shaping based on the inequality (42). The maximum operators ensure that the tail of the crosstalk path does not exceed the reverberation tails of the desired signal paths. To illustrate this further, examples of the weighting windows for the undesired part of the *direct* and the *crosstalk* component, as well as their reciprocal values, are depicted in Fig. 2.

A. CTC Design for M Reference Listening Positions

In this subsection, we derive the algorithm for Q sources, N loudspeakers, and M microphones or listening positions. Using

$$l_m(n) = \begin{cases} 10^{\frac{3}{\log(N_{0m}/(N_{1m}+N_2))} \log\left(\frac{n}{N_{1m}+N_2}\right)+0.5}, & (N_{1m}+N_2) \leq n < L_g \\ 0, & \text{otherwise} \end{cases} \tag{45}$$

a p-norm-based objective function, this leads us to Q individual optimization problems given by

$$\underset{\mathbf{h}_q}{\text{minimize}}: \quad f_q(\mathbf{h}_q)+\alpha \cdot y_q(\mathbf{h}_q) \text{ s.t. } \mathbf{h}_q^T\mathbf{h}_q=1,\ q=1,\ldots,Q \tag{47}$$

where α is the weighting factor for the frequency domain based regularization term $y_q(\mathbf{h}_q)$ (see Section VI-D) and

$$f_q(\mathbf{h}_q) = \log\left(\frac{f_q^{(u)}(\mathbf{h}_q)}{f_q^{(d)}(\mathbf{h}_q)}\right) \tag{48}$$

with

$$f_q^{(d)}(\mathbf{h}_q) = \left\|\mathbf{g}_q^{(d)}\right\|_{p_d} \tag{49}$$

and

$$f_q^{(u)}(\mathbf{h}_q) = \left\|\mathbf{g}_q^{(u)}\right\|_{p_u}. \tag{50}$$

The log operation in (48) is used in view of obtaining a compact description for the gradient of the objective function. The vectors $\mathbf{g}_q^{(d)}$ and $\mathbf{g}_q^{(u)}$ are given by

$$\mathbf{g}_q^{(d)} = \left[\mathbf{g}_{1q}^{(d)^T}, \mathbf{g}_{2q}^{(d)^T}, \ldots, \mathbf{g}_{Mq}^{(d)^T}\right]^T \tag{51}$$

and

$$\mathbf{g}_q^{(u)} = \left[\mathbf{g}_{1q}^{(u)^T}, \mathbf{g}_{2q}^{(u)^T}, \ldots, \mathbf{g}_{Mq}^{(u)^T}\right]^T \tag{52}$$

where $\mathbf{g}_{mq}^{(d)}$ and $\mathbf{g}_{mq}^{(u)}$ contain the sequences $g_{mq}^{(d)}(n)$ and $g_{mq}^{(u)}(n)$, respectively. This means that, basically, all the global impulse responses $g_{mq}(n)$ are weighted according to their modes (signal-path or crosstalk-path) and stacked up to form the vectors $\mathbf{g}_q^{(d)}$ and $\mathbf{g}_q^{(u)}$.

The optimization of (47) is done by applying an iterative gradient-descent procedure. The learning rule reads

$$\mathbf{h}_q^{i+1} = \mathbf{h}_q^i - \mu^i \cdot \left(\nabla_{\mathbf{h}_q} f_q\left(\mathbf{h}_q^i\right) + \alpha \cdot \nabla_{\mathbf{h}_q} y_q\left(\mathbf{h}_q^i\right)\right) \tag{53}$$

with μ^i being an adaptive positive step-size parameter in iteration i. The fulfillment of the side condition is achieved by renormalizing the target vector $\mathbf{h}_q^{i+1}$ after every iteration i of the optimization procedure. The gradient of $f_q(\mathbf{h}_q)$ is formally given by

$$\nabla_{\mathbf{h}_q} f_q(\mathbf{h}_q) = \frac{1}{f_q^{(u)}(\mathbf{h}_q)} \nabla_{\mathbf{h}_q} f_q^{(u)}(\mathbf{h}_q) - \frac{1}{f_q^{(d)}(\mathbf{h}_q)} \nabla_{\mathbf{h}_q} f_q^{(d)}(\mathbf{h}_q). \tag{54}$$

The required individual gradients $\nabla_{\mathbf{h}_q} f_q^{(d)}(\mathbf{h}_q)$ and $\nabla_{\mathbf{h}_q} f_q^{(u)}(\mathbf{h}_q)$ will be given in Section IV-B, where they form the special case of (63) and (67) with $R = 1$.

B. Robust CTC Design Based on Multiple Realizations of the Channel

In this subsection, we derive the algorithm for the case in which we have R perturbed realizations of the channel matrix $\mathbf{C}^{(r)}, r = 1, \ldots, R$. The filters are designed in such a way that all realizations of the acoustic channels are reshaped jointly. The realizations usually result from multiple measurements of the acoustic channels in the vicinity of a given reference position. The corresponding global impulse responses, denoted by $\mathbf{g}_{mq}^{(r)}$, are computed via $\mathbf{g}_q^{(r)} = \mathbf{C}^{(r)}\mathbf{h}_q$. In accordance with [20], [23] and the previous section, we define the desired parts of the global impulse responses as

$$g_{mq}^{(r,d)}(n) = w_{mq}^{(d)}(n) \cdot g_{mq}^{(r)}(n). \tag{55}$$

The unwanted parts are given by

$$g_{mq}^{(r,u)}(n) = w_{mq}^{(u)}(n) \cdot g_{mq}^{(r)}(n). \tag{56}$$

The windows for the desired and unwanted parts are defined as before (i.e., (40) and (43) for the desired and (44) and (46) for the undesired parts).

As in Section IV-A, where the design was based on a single set of microphone positions, one ends up with Q individual optimization problems given by

$$\underset{\mathbf{h}_q}{\text{minimize}}: \quad f_q(\mathbf{h}_q)+\alpha \cdot y_q(\mathbf{h}_q) \text{ s.t. } \mathbf{h}_q^T\mathbf{h}_q=1,\ q=1,\ldots,Q \tag{57}$$

where α is the weighting factor for the frequency domain-based regularization term $y_q(\mathbf{h}_q)$ (see Section IV-D) and

$$f_q(\mathbf{h}_q) = \log\left(\frac{f_q^{(u)}(\mathbf{h}_q)}{f_q^{(d)}(\mathbf{h}_q)}\right) \tag{58}$$

with

$$f_q^{(d)}(\mathbf{h}_q) = \left\|\mathbf{g}_q^{(d)}\right\|_{p_d} \tag{59}$$

and

$$f_q^{(u)}(\mathbf{h}) = \left\|\mathbf{g}_q^{(u)}\right\|_{p_u} \tag{60}$$

where

$$\mathbf{g}_q^{(d)} = \left[\mathbf{g}_{1q}^{(1,d)^T}, \ldots, \mathbf{g}_{Mq}^{(1,d)^T}, \ldots, \mathbf{g}_{1q}^{(R,d)^T}, \ldots, \mathbf{g}_{Mq}^{(R,d)^T}\right]^T \tag{61}$$

and

$$\mathbf{g}_q^{(u)} = \left[\mathbf{g}_{1q}^{(1,u)^T}, \ldots, \mathbf{g}_{Mq}^{(1,u)^T}, \ldots, \mathbf{g}_{1q}^{(R,u)^T}, \ldots, \mathbf{g}_{Mq}^{(R,u)^T}\right]^T. \tag{62}$$

Thus, all realizations of the global impulse responses are weighted and stacked up to form the vectors $\mathbf{g}_q^{(d)}$ and $\mathbf{g}_q^{(u)}$, with each one consisting of $M \cdot R$ impulse responses of length L_g.

The optimization is, again, carried out by utilizing a gradient descent procedure with renormalizing the target vector $\mathbf{h}_q$ after every iteration of the optimization procedure. The involved gradients are derived in the following equations. The gradient required in Section IV-A is given by the special case in which R equals one.

The gradient for $f_q^{(d)}(\mathbf{h}_q)$ is calculated as

$$\nabla_{\mathbf{h}_q} f_q^{(d)}(\mathbf{h}_q) = \zeta_q^{(d)}(\mathbf{h}_q) \cdot \nabla_{\mathbf{h}_q} \phi_{f_q^{(d)}}(\mathbf{h}_q) \tag{63}$$

where

$$\zeta_q^{(d)}(\mathbf{h}_q) = \left(\sum_{m=1}^{M} \sum_{r=1}^{R} \sum_{n=0}^{L_g-1} \left| g_{mq}^{(r,d)}(n) \right|^{p_d} \right)^{\frac{1}{p_d}-1} \tag{64}$$

and

$$\nabla_{\mathbf{h}_q} \phi_{f_q^{(d)}}(\mathbf{h}) = \begin{bmatrix} \sum_{m=1}^{M} \sum_{r=1}^{R} \left(\mathbf{C}_{m1}^{(r)} \right)^T \mathbf{b}_{mq}^{(r,d)} \\ \vdots \\ \sum_{m=1}^{M} \sum_{r=1}^{R} \left(\mathbf{C}_{mN}^{(r)} \right)^T \mathbf{b}_{mq}^{(r,d)} \end{bmatrix} \tag{65}$$

with $\mathbf{b}_{mq}^{(r,d)}$ given by

$$\mathbf{b}_{mq}^{(r,d)} = \operatorname{diag}\left\{ \operatorname{sign}\left\{ \mathbf{g}_{mq}^{(r,d)} \right\} \right\} \operatorname{diag}\left\{ \mathbf{W}_{mq}^{(d)} \right\} \cdot \left| \mathbf{g}_{mq}^{(r,d)} \right|^{(p_d-1)}. \tag{66}$$

The gradient for the undesired part $f_q^{(u)}(\mathbf{h}_q)$ is calculated as

$$\nabla_{\mathbf{h}_q} f_q^{(u)}(\mathbf{h}_q) = \zeta_q^{(u)}(\mathbf{h}_q) \cdot \nabla_{\mathbf{h}_q} \phi_{f_q^{(u)}}(\mathbf{h}_q) \tag{67}$$

where

$$\zeta_q^{(u)}(\mathbf{h}_q) = \left(\sum_{m=1}^{M} \sum_{r=1}^{R} \sum_{n=0}^{L_g-1} \left| g_{mq}^{(r,u)}(n) \right|^{p_u} \right)^{\frac{1}{p_u}-1} \tag{68}$$

and

$$\nabla_{\mathbf{h}_q} \phi_{f_q^{(u)}}(\mathbf{h}_q) = \begin{bmatrix} \sum_{m=1}^{M} \sum_{r=1}^{R} \left(\mathbf{C}_{m1}^{(r)} \right)^T \mathbf{b}_{mq}^{(r,u)} \\ \vdots \\ \sum_{m=1}^{M} \sum_{r=1}^{R} \left(\mathbf{C}_{mN}^{(r)} \right)^T \mathbf{b}_{mq}^{(r,u)} \end{bmatrix} \tag{69}$$

with $\mathbf{b}_{mq}^{(r,u)}$ given by

$$\mathbf{b}_{mq}^{(r,u)} = \operatorname{diag}\left\{ \operatorname{sign}\left\{ \mathbf{g}_{mq}^{(r,u)} \right\} \right\} \operatorname{diag}\left\{ \mathbf{W}_{mq}^{(u)} \right\} \cdot \left| \mathbf{g}_{mq}^{(r,u)} \right|^{(p_u-1)}. \tag{70}$$

Finally, the gradient of $f_q(\mathbf{h}_q)$ reads

$$\nabla_{\mathbf{h}_q} f_q(\mathbf{h}_q) = \frac{1}{f_q^{(u)}(\mathbf{h}_q)} \nabla_{\mathbf{h}_q} f_q^{(u)}(\mathbf{h}_q) - \frac{1}{f_q^{(d)}(\mathbf{h}_q)} \nabla_{\mathbf{h}_q} f_q^{(d)}(\mathbf{h}_q). \tag{71}$$

The algorithm can be implemented computationally efficient by exploiting the Toeplitz structure of the convolution matrices $\mathbf{C}_{m\ell}$ and utilizing the FFT and IFFT to calculate the corresponding matrix-vector multiplications in the Fourier domain. The Hadamard product can be used to lower the computational effort in calculating the vectors $\mathbf{b}_{mq}^{(r,d)}$ and $\mathbf{b}_{mq}^{(r,u)}$.

C. Robust CTC Design Based on Statistics of Room Impulse Responses

Designing robust prefilters based on multiple realizations of the channel matrix is quite time consuming and requires measurements of the different RIRs. As a remedy, we present an algorithm that yields spatial robustness by incorporating the statistical knowledge about room impulse responses into the optimization problem. As described in Section III-C, the perturbation of an acoustic channel $c_{m\ell}(n)$ is modeled by an additive stochastic system $p_{m\ell}(n)$ that describes statistically the perturbation in the case of spatial mismatch of a microphone m to its reference position.

A straightforward way to incorporate the stochastic perturbation in the prefilter design would be to extend the design criterion (47) by the expectation operator as follows:

$$\begin{aligned} \underset{\mathbf{h}_q}{\text{minimize :}} \quad & \log\left(\frac{\bar{f}_q^{(u)}(\mathbf{h}_q)}{f_q^{(d)}(\mathbf{h}_q)} \right) + \alpha \cdot y_q(\mathbf{h}_q) \\ \text{s.t. } & \mathbf{h}_q^T \mathbf{h}_q = 1,\ q = 1, \ldots, Q \end{aligned} \tag{72}$$

where

$$\bar{f}_q^{(u)}(\mathbf{h}_q) = \mathrm{E}\left\{ \left\| \mathbf{g}_q^{(u)} \right\|_{p_u} \right\} \tag{73}$$

and

$$f_q^{(d)}(\mathbf{h}_q) = \left\| \mathbf{g}_q^{(d)} \right\|_{p_d}. \tag{74}$$

As in (47), α is the weighting factor for the regularization term $y_q(\mathbf{h}_q)$.

For the reason of simplicity, the stochastic component is considered for the undesired part only. The required weighted stochastic global responses are given by

$$\mathbf{g}_q^{(u)} = \widetilde{\mathbf{W}}_q (\mathbf{C} + \mathbf{P}) \mathbf{h}_q \tag{75}$$

where $\widetilde{\mathbf{W}}$ is defined in (28) and the individual weights $w_{mq}(n) = w_{mq}^{(u)}(n)$ are given by (44) and (46), respectively. Instead of aiming at minimizing (72) directly, based on the Minkowski inequality in the form

$$\| \widetilde{\mathbf{W}}_q \mathbf{C} \mathbf{h}_q + \widetilde{\mathbf{W}}_q \mathbf{P} \mathbf{h}_q \|_{p_u} \leq \| \widetilde{\mathbf{W}}_q \mathbf{C} \mathbf{h}_q \|_{p_u} + \| \widetilde{\mathbf{W}}_q \mathbf{P} \mathbf{h}_q \|_{p_u} \tag{76}$$

we might try to minimize the upper bound by replacing $\bar{f}_q^{(u)}(\mathbf{h}_q)$ in (73) with

$$\| \widetilde{\mathbf{W}}_q \mathbf{C} \mathbf{h}_q \|_{p_u} + \mathrm{E}\left\{ \| \widetilde{\mathbf{W}}_q \mathbf{P} \mathbf{h}_q \|_{p_u} \right\}.$$

However, since little is known about the exact probability density functions of the perturbation $\mathbf{P}$, we resort to a quadratic cost function for the perturbation term and consider the optimization problem

$$\underset{\mathbf{h}_q}{\text{minimize :}} \quad f_q(\mathbf{h}_q) + \alpha \cdot y_q(\mathbf{h}_q) \text{ s.t. } \mathbf{h}_q^T \mathbf{h} = 1,\ q = 1, \ldots, Q \tag{77}$$

where

$$f_q(\mathbf{h}_q) = \log\left(\frac{\tilde{f}_q^{(u)}(\mathbf{h}_q)}{f_q^{(d)}(\mathbf{h}_q)} \right) \tag{78}$$

with

$$\tilde{f}_q^{(u)}(\mathbf{h}_q) = \| \widetilde{\mathbf{W}}_q \mathbf{C} \mathbf{h}_q \|_{p_u} + \beta \cdot \mathrm{E}\left\{ \| \widetilde{\mathbf{W}}_q \mathbf{P} \mathbf{h}_q \|_2 \right\} \tag{79}$$

instead, where β is some appropriate positive weight.[1]

The objective function $f_q(\mathbf{h}_q)$ according to (77) is minimized by applying the gradient-descent procedure (53). The gradient reads

$$\nabla_{\mathbf{h}_q} f_q(\mathbf{h}_q) = \frac{1}{\tilde{f}_q^{(u)}(\mathbf{h}_q)} \nabla_{\mathbf{h}_q} \tilde{f}_q^{(u)}(\mathbf{h}_q) - \frac{1}{f_q^{(d)}(\mathbf{h}_q)} \nabla_{\mathbf{h}_q} f_q^{(d)}(\mathbf{h}_q). \tag{80}$$

with

$$\nabla_{\mathbf{h}_q} \tilde{f}_q^{(u)}(\mathbf{h}_q) = \nabla_{\mathbf{h}_q} f_q^{(u)}(\mathbf{h}_q) + \beta \cdot \nabla_{\mathbf{h}_q} f_q^{(P)}(\mathbf{h}_q) \tag{81}$$

where $\nabla_{\mathbf{h}_q} f_q^{(d)}(\mathbf{h}_q)$ and $\nabla_{\mathbf{h}_q} f_q^{(u)}(\mathbf{h}_q)$ are given in (63) and (67) with R set to one. The remaining part of $\nabla_{\mathbf{h}_q} \tilde{f}_q^{(u)}(\mathbf{h}_q)$ can be derived by exploiting the theory developed in Section III-C. With

$$f_q^{(P)}(\mathbf{h}_q) = \mathrm{E}\left\{ \|\widetilde{\mathbf{W}}_q \mathbf{P} \mathbf{h}_q\|_2 \right\} = \left(\sum_{m=1}^{M} \sum_{\ell=1}^{N} \mathbf{h}_{\ell q}^T \mathbf{M}_{m\ell} \mathbf{h}_{\ell q} \right)^{\frac{1}{2}} \tag{82}$$

and $\mathbf{M}_{m\ell}$ given in (37), the gradient for the regularization term $f_q^{(P)}(\mathbf{h}_q)$ becomes

$$\nabla_{\mathbf{h}_q} f_q^{(P)}(\mathbf{h}_q) = \frac{1}{2} \cdot \left(\sum_{m=1}^{M} \sum_{\ell=1}^{N} \mathbf{h}_{\ell q}^T \mathbf{M}_{m\ell} \mathbf{h}_{\ell q} \right)^{-\frac{1}{2}} \times \begin{bmatrix} \sum_{m=1}^{M} \left(\mathbf{M}_{m1} + \mathbf{M}_{m1}^T \right) \mathbf{h}_{1q} \\ \vdots \\ \sum_{m=1}^{M} \left(\mathbf{M}_{mN} + \mathbf{M}_{mN}^T \right) \mathbf{h}_{Nq} \end{bmatrix}. \tag{83}$$

D. Frequency Domain-Based Regularization Term

In [25], we proposed to jointly optimize the time- and frequency-domain representations of an impulse response in order to achieve a *good* overall reshaping without degrading the perceived quality due to high spectral peaks in the overall system. In this section we extend the p-norm based optimality criterion that is used as a regularization term to the multichannel scenario considered in this paper.

The proposed regularization term is defined as

$$y_q(\mathbf{h}_q) = \|\mathbf{a}_{qf}\|_{p_f} \tag{84}$$

where the vector $\mathbf{a}_{qf}$ is made up by the concatenation of the discrete Fourier transforms (DFTs) of all available realizations of the GIRs for the qth source. Using this optimality criterion one demands the GIRs to not contain any high spectral peaks.

To derive the gradient for (84), we reformulate the regularization term in matrix-vector notation. We define a matrix $\widetilde{\mathbf{C}}$ as

$$\widetilde{\mathbf{C}} = \begin{bmatrix} \widehat{\mathbf{C}}^{(1)} \\ \widehat{\mathbf{C}}^{(2)} \\ \vdots \\ \widehat{\mathbf{C}}^{(R)} \end{bmatrix} \quad \text{with} \quad \widehat{\mathbf{C}}^{(r)} = \begin{bmatrix} \mathbf{C}_{11}^{(r)} & \cdots & \mathbf{C}_{1N}^{(r)} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{M1}^{(r)} & \cdots & \mathbf{C}_{MN}^{(r)} \end{bmatrix} \tag{85}$$

that contain the individual convolution matrices for each acoustic channel. Furthermore, we define a block-diagonal matrix $\widetilde{\mathbf{F}} = \mathrm{diag}\{\mathbf{F}, \mathbf{F}, \ldots, \mathbf{F}\}$ in which $\mathbf{F}$ is a DFT matrix of compatible size such that products $\mathbf{F}\mathbf{C}_{ik}^{(r)}$ can be taken.

These definitions allow us to rewrite the regularization term as

$$y_q(\mathbf{h}_q) = \|\mathbf{a}_{qf}\|_{p_f} = \|\widetilde{\mathbf{F}}\widetilde{\mathbf{C}}\mathbf{h}_q\|_{p_f}. \tag{86}$$

The gradient for the regularization term is calculated as

$$\nabla_{\mathbf{h}_q} y_q(\mathbf{h}_q) = \zeta_q(\mathbf{h}_q) \cdot \Re\left\{ (\widetilde{\mathbf{F}}\widetilde{\mathbf{C}})^H \mathbf{b}_q f \right\} \tag{87}$$

with $\zeta_q(\mathbf{h}_q)$ given by

$$\zeta_q(\mathbf{h}_q) = \left(\sum_{k=0}^{L_a-1} |a_{qf}(k)|^{p_f} \right)^{\frac{1}{p_f}-1} \tag{88}$$

$\mathbf{b}_{qf}$ given by

$$\mathbf{b}_{qf} = \mathrm{diag}\left\{ \mathrm{sign}\{\mathbf{a}_{qf}\} \right\} \cdot |\mathbf{a}_{qf}|^{p_f-1}. \tag{89}$$

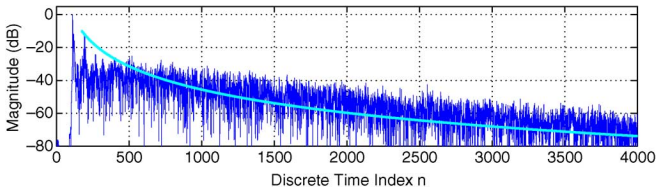

Fig. 3. Magnitude $|c_{m\ell}(n)|$ of one of the measured room impulse responses. The cyan (light gray) curve is the compromise temporal masking limit of the human auditory system.

V. Experiments and Results

For the experiments we measured room impulse responses in an office room of size 6.85 m × 5.3 m × 3 m. The reverberation time was estimated as $\tau_{60} = 0.4$ s. We used four Klein + Hummel M52 loudspeakers as sound sources. They had a distance of 1.6 m to the back wall, 1.5 m to the ground and a spacing of 40 cm between them. For the recordings we used a Cortex MK-2 dummy head with MK250 microphone capsules inside its ears, with the ears placed at a height of 1.6 m above the floor and mounted on a linear stage with a high positioning accuracy. Measurements were taken around two reference listening positions with a distance of 80 cm between them, both 2.2 m away from the loudspeakers, facing directly toward them. Using two reference positions allows us to present results not only for one listener, but also to include the case in which the CTC problem has to be solved simultaneously for two listeners.

The room impulse responses were measured using the exponential sine-sweep method from [28] at a sampling rate of 48 kHz and were then downsampled to 16 kHz. The lengths of the room impulse responses were limited to $L_c = 4000$ taps. To get different realizations of the acoustic channels from the four loudspeakers to the microphone positions, we moved the head within a 2 cm × 2 cm × 2 cm volume around the respective reference positions with a spatial sampling distance of 1 cm on every axis, resulting in 27 realizations of each channel. Besides that, we measured 40 more realizations of the channels by placing the dummy head at 40 positions inside the listening areas, but not on the reference positions. The prefilters were designed using

[1] If the perturbation is assumed to be Gaussian distributed, then it would in fact be possible to obtain an analytic expression for $\mathrm{E}\{\|\widetilde{\mathbf{W}}_q \mathbf{P} \mathbf{h}_q\|_p\}$, but computing the gradient with respect to $\mathbf{h}_q$ would still be cumbersome.

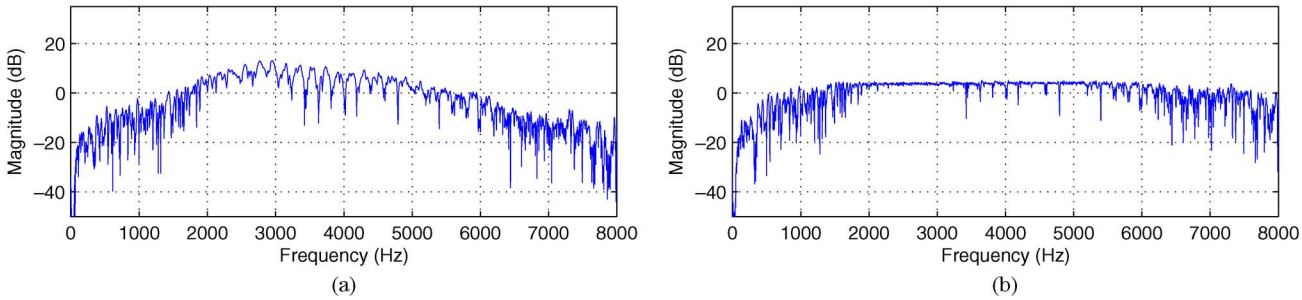

Fig. 4. (a) Frequency response of the reshaped direct sound path by optimizing only the p-norm of the time-domain representation of the GIRs. (b) Frequency response of the reshaped direct sound path by jointly optimizing the p-norm of the time- and the frequency-domain representation of the GIRs ($\alpha = 10$).

either one or all 27 impulse responses from the reference positions and were then applied to the 40 test positions to measure the performance in the case of spatial mismatch.

For illustration purposes, the energy decay behavior of one of the measured RIRs is shown in Fig. 3 together with the average temporal masking limit according to [22].

For a quantitative description of the achieved dereverberation we use a normalized version of the *perceivable reverberation quantization measure* introduced in [25]. This measure captures the average magnitude of the impulse response taps that overshoot the temporal masking limit on a logarithmic scale and is above -60 dB compared to the direct sound. We denote this measure by nPRQ. It is calculated as

$$\mathrm{nPRQ} = \begin{cases} \frac{1}{\|\mathbf{g}_E\|_0} \cdot \sum_{n=N_0}^{L_g-1} g_E(n), & \text{for } \|\mathbf{g}_E\|_0 > 0 \\ 0, & \text{otherwise} \end{cases} \quad (90)$$

with Equation (90), shown at the bottom of the page, and $\|\mathbf{g}_E\|_0$ denoting the ℓ_0 pseudo norm, which counts the number of nonzero elements of a vector. If the RIR is completely reshaped, then either no time coefficient exceeds the temporal masking limit or the energy of all exceeding coefficients is below -60 dB; in both cases $\mathrm{nPRQ} = 0$. Otherwise, if filter taps are above the masking limit, it measures the average overshot in dB.

For all experiments the lengths of the prefilters were chosen to be $L_h = 5000$ taps. As in [24], the parameters p_d and p_u were selected as $p_d = 20$ and $p_u = 10$. Similar to [25] p_f was selected as $p_f = 8$ for the frequency-domain based regularization term. Moreover, a value of $w_0 = 180$ was used, which means that if the objective function (47) with $p_d, p_u \to \infty$ amounts to $f_q(\mathbf{h}_q) = 0$, then the crosstalk component will be $20 \cdot \log_{10}(180) = 45.1$ dB below the direct component. To measure the performance of the crosstalk cancellation, we compare the magnitude of the main peak of the desired signal path to the magnitude of the main peak of the crosstalk path. We refer to this measure as the *direct signal to crosstalk ratio* (DSCR). The value for the weighting factor α was chosen empirically to be $\alpha = 10$ so that the frequency responses of the overall systems had an acceptable shape. To demonstrate the effect of the regularization term, we exemplarily depict the frequency responses of reshaped GIRs of the direct sound path with $\alpha = 0$ and $\alpha = 10$ in Fig. 4.

First, we applied the algorithm from Section IV-A to design the prefilters. To simplify the explanation, this method will be referred to as Algorithm A in the following. Correspondingly, the algorithms from Sections IV-B and IV-C will be called Algorithms B and C, respectively. The nPRQ and DSCR measures obtained with Algorithm A for different scenarios are listed in Table I. For comparison purposes, we also considered the least-squares design criteria ($p_d = p_u = 2$). When minimizing the least-squares optimality criterion, we utilize the postfiltering approach from [19] to compensate for spectral distortions. The postfilters were designed based on the average autocorrelation sequence of all available reference global impulse responses. The length of the postfilters was chosen empirically to generate an acceptable overall frequency response. We exemplarily show the effect of the postfiltering method in Fig. 5.

When considering just one dummy head and two loudspeakers, the original DSCR without any prefiltering was 6.1 dB, and the nPRQ measure was 8.4 dB. By applying the prefilters designed with Algorithm A, the DSCR could be enhanced to 41.6 dB, and the nPRQ measure could be reduced to 0.6 dB for the signal path. The fact that the nPRQ measure is greater than zero means that the room reverberation is so strong that prefilters with 5000 taps are still too short to push the reverberation tail completely under the masking limit. However, given the p-norm design criterion, the tail follows the desired decay. This can be observed in Fig. 6, which depicts the obtained overall responses. Besides the shaping of the decay for the desired part, the figure also shows that the reverberation tail of the crosstalk component does not exceed the tail of the desired component. Considering four loudspeakers and one dummy head, the nPRQ measure could be reduced to zero, and the DSCR between the two ears was enhanced to 51.7 dB,

$$g_{\mathrm{E}}(n) = \begin{cases} 20 \cdot \log_{10}\left(|g(n)| \cdot w_u(n)\right), & \text{for } |g(n)| > \max\left[\frac{1}{w_u(n)}, -60\ \mathrm{dB}\right] \\ 0, & \text{otherwise} \end{cases} \quad (91)$$

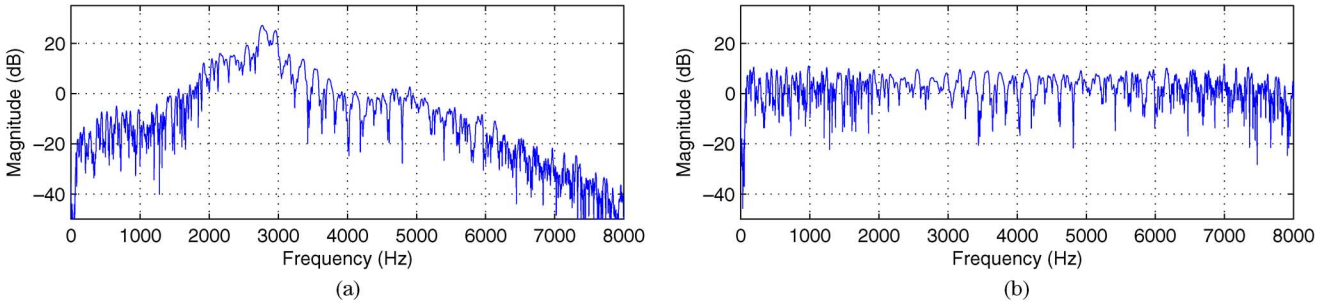

Fig. 5. (a) Frequency response of the reshaped direct sound path by optimizing the least-squares optimality criterion. (b) The frequency response after applying the postfilter method from [19]; the length of the linear prediction-error filter was 40 taps.

TABLE I
VALUES FOR nPRQ AND DSCR AT THE REFERENCE POSITIONS BEFORE AND AFTER APPLYING ALGORITHM A

No. speakers	2		4		4		4	
No. microphones	2		2		4 (Head 1)		4 (Head 2)	
Measure [dB]	nPRQ	DSCR	nPRQ	DSCR	nPRQ	DSCR	nPRQ	DSCR
untouched	8.4	6.1	8.4	6.1	8.4	6.1	10.3	10.7
Alg. A (p-norm)	0.6	41.6	0.0	51.7	4.2	36.7	0.1	43.7
Alg. A (2-norm)	4.6	18.7	1.6	35.6	5.6	23.2	4.6	25.7

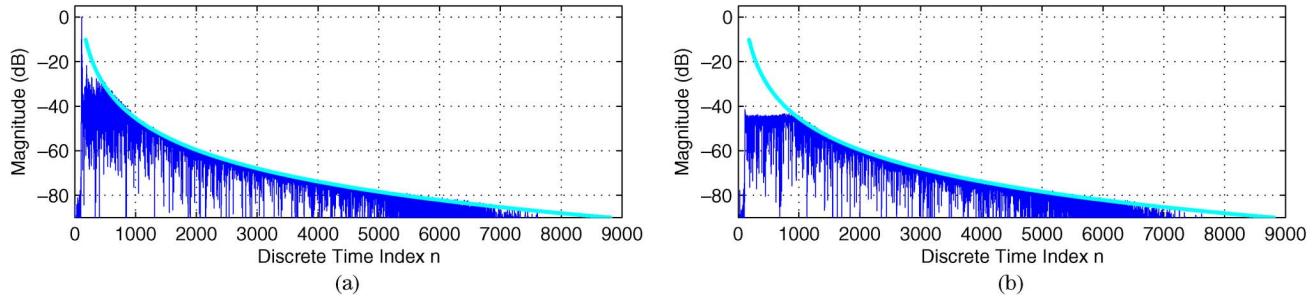

Fig. 6. Reshaped room responses for the 2 × 2 case. (a) Signal path. (b) Crosstalk path. The cyan (light gray) line is the temporal masking limit.

TABLE II
AVERAGE nPRQ AND DSCR VALUES FOR THE DIFFERENT ALGORITHMS IN THE PRESENCE OF SPATIAL MISMATCH

No. speakers	2		4		4		4	
No. microphones	2		2		4 (Head 1)		4 (Head 2)	
Measure [dB]	nPRQ	DSCR	nPRQ	DSCR	nPRQ	DSCR	nPRQ	DSCR
Alg. A (p-norm)	8.2	25.5	9.1	22.7	9.6	23.6	7.2	28.9
Alg. B (p-norm)	9.2	26.8	4.6	32.2	10.8	24.2	4.7	32.5
Alg. C (p-norm)	7.5	25.7	7.2	23.1	11.3	19.7	5.7	28.8
Alg. A (2-norm)	11.2	17.7	13.6	18.1	12.6	19.9	10.8	23.8
Alg. B (2-norm)	6.6	20.4	8.4	14.7	8.2	14.1	8.7	18.0
Alg. C (2-norm)	11.8	22.5	13.0	17.3	10.7	19.5	9.2	21.2

which is above the design specification of 45.1 dB. In this case, the design goal could be reached with 5000-tap prefilters, and all responses stay below their prescribed limits. With four loudspeakers and four microphones, the DSCR could be kept around the 40-dB mark, and nPRQ measures of 0.1 and 4.2 dB were obtained. The results for the least-squares approach were generally inferior, especially in terms of crosstalk cancellation.

To investigate the robustness of the different algorithms, the prefilters were tested on spatial positions that were not used in the filter design. For Algorithms A and C, the design was based on the reference position only, whereas for Algorithm B the design was based on the 27 reference room impulse responses. In all cases, the prefilters were then applied to the 40 test impulse responses that were measured between the 27 reference positions. We then calculated the average DSCR and nPRQ measures over the 40 reshaped realizations. For Algorithm C, we assumed an average displacement of $D = 2$ cm to the reference position. The value of the regularization factor β in (79) has been found empirically as $\beta = 10^{-4}$, which worked well for all setups.

The nPRQ and DSCR results are listed in Table II together with those for the corresponding least-squares designs. As one can see, when a spatial mismatch occurs, the average nPRQ with Algorithm A is in the same range as without applying the reshaping filters. However, it is important to note that the average DSCR measure could still be enhanced. With Algorithm B, the values for the nPRQ and DSCR measures could be improved for all setups. The comparison of Algorithm C with Algorithm A shows that it can effectively improve the spatial robustness. Importantly, with Algorithm C the performance of Algorithm B

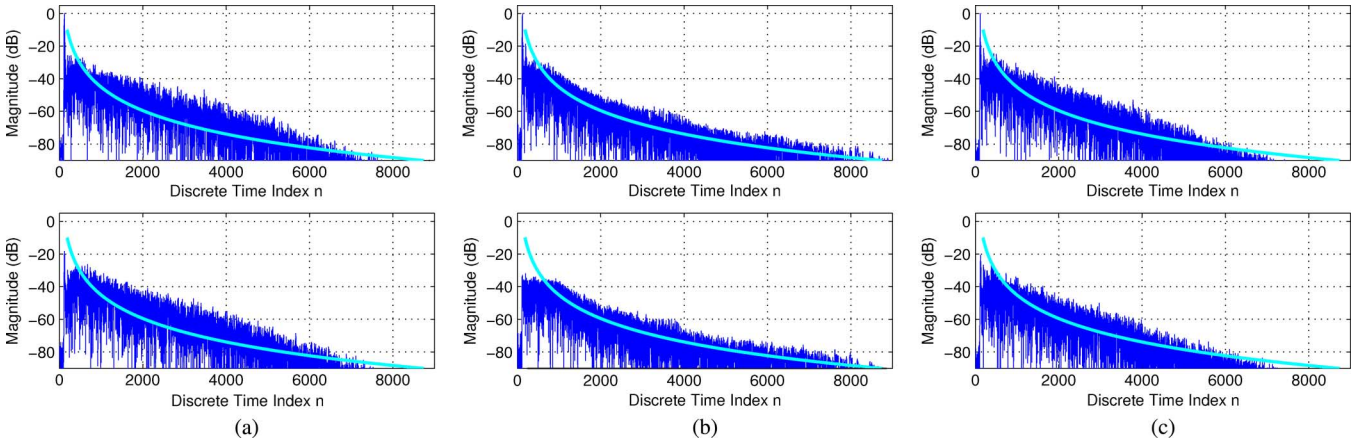

Fig. 7. Signal path (upper row) and the crosstalk path (lower row) when applying the reshaping filters designed with Algorithm A (a), Algorithm B (b) and Algorithm C (c) with spatial mismatch. The actual RIRs were not part of the design process. The cyan (light gray) line is the temporal masking limit.

is almost reached, but it should be noted that Algorithm B used a 27-fold number of measurements, which is difficult to obtain in practice.

Comparing the results from the p-norm to the 2-norm based method in Table II it can be seen that the p-norm based approach yields, in general, better results. However, in some cases the 2-norm based approach results in a different tradeoff between dereverberation and crosstalk cancellation.

To give a visual impression of the effect of reshaping in the presence of spatial mismatch, Fig. 7 depicts the reshaping results for a realization of the channel matrix for the 2×4 setup with a small displacement with prefilters designed with Algorithms A and B. It can be clearly seen that Algorithm B performs better in terms of crosstalk cancellation and reshaping.

VI. Conclusion

We presented a unified framework that covers two demanding auditory objectives, namely dereverberation by RIR reshaping and crosstalk cancellation with arbitrary speaker and microphone setups. Furthermore, we explicitly considered the problem of degraded perceived quality due to high spectral peaks in the overall systems. It was shown that, according to the spatial sampling theorem of RIRs, one can achieve effective reshaping and crosstalk cancellation in a limited listening area by designing the equalizers based on a set of spatially sampled RIRs. Moreover, a spatially robust design method that incorporates statistical knowledge of RIR behavior has been introduced. This method requires only RIR measurements at the reference positions and almost reaches the performance of the spatial-sampling method at a fraction of the measurement effort.

References

[1] P. Damaske, "Head-related two-channel stereophony with loudspeaker reproduction," *J. Acoust. Soc. Amer. (JASA)*, vol. 50, pp. 1109–1115, 1971.

[2] C. Bourget and T. Aboulnasr, "Inverse filtering of room impulse response for binaural recording playback through loudspeakers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Adelaide, Australia, Apr. 1994, vol. 3, pp. 301–304.

[3] O. Kirkeby, P. A. Nelson, H. Hamada, and F. O. na Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 189–194, Mar. 1998.

[4] Y. Kahana, P. A. Nelson, and S. Yoon, "Experiments on the synthesis of virtual acoustic sources in automotive interiors," in *Proc. AES 16th Int. Conf.: Spatial Sound Reproduction*, Roveniemi, Finland, Apr. 1999, vol. 15, pp. 218–232.

[5] 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.

[6] D. B. Ward, "Joint least squares optimization for robust acoustic crosstalk cancellation," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 211–215, Feb. 2000.

[7] M. Kallinger and A. Mertins, "A spatially robust least squares crosstalk canceller," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 15–20, 2007, vol. 1, pp. 177–180.

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

[9] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, no. 1, pp. 165–169, Jul. 1979.

[10] S. J. Elliott and P. A. Nelson, "Multiple-point equalization in a room using adaptive digital filters," *J. Audio Eng. Soc.*, vol. 37, no. 11, pp. 899–907, Nov. 1989.

[11] D. D. Falconer and F. R. Magee, "Adaptive channel memory truncation for maximum likelihood sequence estimation," *Bell Syst. Tech. J.*, vol. 52, no. 9, pp. 1541–1562, Nov. 1973.

[12] P. J. W. Melsa, R. C. Younce, and C. E. Rohrs, "Impulse response shortening for discrete multitone transceivers," *IEEE Trans. Commun.*, vol. 44, no. 12, pp. 1662–1672, Dec. 1996.

[13] R. K. Martin, D. Ming, B. L. Evans, and C. R. Johnson, Jr., "Efficient channel shortening equalizer design," *J. Appl. Signal Process.*, vol. 13, pp. 1279–1290, Dec. 2003.

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

[15] W. Zhang, A. W. H. Khong, and P. A. Naylor, "Adaptive inverse filtering of room acoustics," in *Proc. 42nd Asilomar Conf. Signals, Syst., Comput.*, 2008, pp. 26–29.

[16] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, 2010.

[17] *ISO Norm 3382: Acoustics—Measurement of the Reverberation Time of Rooms with Reference to other Acoustical Parameters*, ISO Norm 3382, Int. Org. for Standardization, 1997.

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

[19] M. Kallinger and A. Mertins, "Room impulse response shortening by channel shortening concepts," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Oct. 30–Nov. 2 2005, pp. 898–902.

[20] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 249–259, Feb. 2010.
[21] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*, 3rd ed. New York: Springer, 2007.
[22] L. D. Fielder, "Practical limits for room equalization," in *Proc. AES 111th Conv.*, 2001, pp. 1–19.
[23] T. Mei and A. Mertins, "On the robustness of room impulse response reshaping," in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.
[24] J. O. Jungmann, R. Mazur, M. Kallinger, and A. Mertins, "Robust combined crosstalk cancellation and listening-room compensation," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA 2011)*, Mohonk, New Paltz, NY, Oct. 2011.
[25] 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. EUSIPCO 2011*, Barcelona, Spain, Aug. 2011.
[26] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.
[27] T. Ajdler, L. Sbaiz, and M. Vetterli, "The plenacoustic function and its sampling," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 3790–3804, Oct. 2006.
[28] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Proc. 122nd AES Convention*, Vienna, Austria, May 5–8, 2007.

Jan Ole Jungmann (S'11) received the B.Sc. and M.Sc. degrees in informatics from the University of Lübeck, Lübeck, Germany, in 2006 and 2009, respectively. He is currently pursuing the Ph.D. degree at the Institute for Signal Processing, University of Lübeck.

His current research interests are digital signal and audio processing, with a special focus on listening room compensation and crosstalk cancellation.

Radoslaw Mazur (S'09–M'11) was born in Wroclaw, Poland, in 1976. He received the Diplominformatiker degree from the University of Oldenburg, Oldenburg, Germany, in 2004 and the Dr.-Ing. degree in computer science from the University of Lübeck, Lübeck, Germany, in 2010.

He was an Assistant Researcher in the Department of Physics, University of Oldenburg, from 2004 to 2006, and then joined the University of Lübeck. The current research interests are digital signal and audio processing, with a special focus on blind source separation.

Markus Kallinger (M'06) received the Dipl.-Ing. degree in electrical engineering from the University of Ulm, Ulm, Germany, in 1999 and the Dr.-Ing. degree in electrical engineering from the University of Bremen, Bremen, Germany, in 2006.

From 1999 to 2004, he was a Research Assistant in the Department of Communications Engineering, University of Bremen. From 2004 to 2006, he was a Research Fellow at the Faculty of Mathematics and Science, University of Oldenburg, Oldenburg, Germany. Since 2007, he has been with the Audio Department, Fraunhofer IIS, Erlangen, Germany. His research interests include speech and audio processing, spatial audio coding, psychoacoustics, quality measures, adaptive filters, and digital audio effects.

Tiemin Mei received the B.S. degree in physics from Sun Yat-sen University, Guangzhou , China, in 1986, the M.S. degree in biophysics from China Medical University, Shenyang, China, in 1991, and the Ph.D. degree in signal and information processing from Dalian University of Technology, Dalian, China, in 2006.

From 2007 to 2010, he was with the Institute for Signal Processing, University of Lübeck, Lübeck, Germany, and from 2004 to 2005 with the School of Electrical Computer and Telecommunications Engineering, the University of Wollongong, Australia. He has been a member of academic staff at Shenyang Ligong University, Shenyang, China, since 1996. His research interests include stochastic signal processing, speech and audio processing, and image processing.

Alfred Mertins (M'96–SM'03) received the Dipl.-Ing. degree from the University of Paderborn, Paderborn, Germany, in 1984 and the Dr.-Ing. degree from the Hamburg University of Technology, Hamburg, Germany, in 1991, both in electrical engineering.

From 1986 to 1991, he was a Research Assistant at the Hamburg University of Technology, and from 1991 to 1995 he was a Senior Scientist at the Microelectronics Applications Center Hamburg. From 1996 to 1997, he was with the University of Kiel, Kiel, Germany, and from 1997 to 1998 with the University of Western Australia, Perth. In 1998, he joined the University of Wollongong, where he was at last an Associate Professor of Electrical Engineering. From 2003 to 2006, he was a Professor in the Faculty of Mathematics and Science at the University of Oldenburg, Germany. In November 2006, he joined the University of Lübeck, Lübeck, Germany, where he is a Professor and Director of the Institute for Signal Processing. His research interests include speech, audio, and image processing, wavelets and filter banks, pattern recognition, and digital communications.