Choice of Inverse Filter Design Parameters in virtual acoustic Imaging Systems*

TIMOS PAPADOPOULOS, *AES Member,* **AND PHILIP A. NELSON,** *AES Member*
(tp@isvr.soton.ac.uk) (p.a.nelson@soton.ac.uk)

Institute of Sound and Vibration Research, University of Southampton, Highfield, Southampton S017 1BJ, UK

The performance of the inverse filtering stage in binaural reproduction systems using loudspeakers is quantified objectively. The influence of the inverse filtering stage design parameters on the actual effectiveness of the inversion is examined and the optimal choice for those parameters determined. The analysis is supported with direct measurements of the inversion effectiveness and the conditions that allow the aimed inversion accuracy to be realistically achieved are discussed.

0 INTRODUCTION

The objective of binaural audio reproduction systems over loudspeakers (also known as transaural reproduction systems due to the early work of Cooper and Bauck [1], [2] on the subject) is to reproduce a pair of binaural signals at the ears of a listener positioned in a chosen listening space. With those signals copying the sound pressure signals that would have been created if the listener were present in a remote sound field, this audio reproduction method can immerse the listener perceptually into that remote sound field. The main challenge in the implementation of this principle is the design of a filtering stage that inverts the electroacoustic plant from the input to the loudspeakers to the output at the listener's ears, thus allowing the transparent delivery of the intended binaural material.

Since the introduction of this audio reproduction principle nearly 50 years ago [3], various designs have been described for the realization of this inverse stage. Damaske [4] described the "90° filter" design, in which delay and attenuation settings were tuned to achieve optimal reproduction of a laterally positioned virtual image before the filter was used for the reproduction of binaural material. Making use of the assumptions of left–right symmetry for the plant responses and the joint minimum-phase character of the sum and difference of the ipsilateral and contralateral plant terms, Cooper and Bauck [2] introduced the "shuffler" topology and demonstrated how it can be realized with recursive filters. An analysis is presented in that work on the inverse design optimization problem and on the suitability of different algebraic norms for this optimization. Extensions to the shuffler topology as well as symmetric and nonsymmetric hybrid designs that combine low-order recursive inversion up to 6 kHz and power transfer reproduction of the binaural signals at higher frequencies were investigated by Gardner [5], who implemented a system utilizing real-time tracking of a moving listener. Møller [6] described a factorization of the inverse matrix that contains only ratios of head-related transfer functions (HRTFs). Mouchtaris et al. [7] introduced a simplification to the common denominator recursive part of that design, making use of the natural head-shadowing effect. They described an adaptive inverse filtering implementation that takes into account the possibly nonminimum-phase characteristics of the denominator terms of the included HRTF ratios. The design considered in the present paper is that of full audio bandwidth mixed-phase inversion with regularization [8]. Applications of this design have been investigated with respect to the acoustic properties of the source/listener geometry (stereo dipole and optimal source distribution systems [9], [10]), the visual tracking of the listener and the online update of the inverse filters [11], the use of frequency-varying regularization and band-limited implementations with subband filtering [12], and the use of warped filters [13].

The basic tradeoff in the implementation of the design described in [8] is that of using low (or zero) regularization at the expense of long and computationally expensive inverse filters, or higher regularization resulting in lower order inverse filters but also suboptimal performance

*A preliminary version of this paper was presented at the Spring Conference of the Institute of Acoustics, Reading, UK, 2008 April 10–11. Manuscript received 2008 July 8; revised 2009 October 10 and December 29.

(at least as predicted in computer simulations). The increasing availability of cheap processing power and the versatility of the personal computer platform in implementing low-latency frequency-domain convolution algorithms should suggest a shift toward the former of these options. However, the performance shown in computer simulations is not necessarily replicated in the actual implementation of the inversion, especially when issues are taken into account such as the nonideal response of the loudspeakers used for the reproduction and the finite dynamic range of the audio reproduction setup. In this paper we use direct measurements of the inversion accuracy to examine that question.

A limited number of studies [5], [12], [14], [15] exist that examine the level of crosstalk cancellation and equalization as shown in measurements and compare it with what is expected from convolution simulations. Among these studies Gardner (examining a different inverse design than the one considered here) presents in [5] a detailed analysis with simulations but limited measurement/simulation comparisons. Bai and Lee [12] and Lentz et al. [14] show significant differences between measurement and simulation (in the former case up to 20 dB less measured crosstalk cancellation compared to the simulation) but offer limited explanations as to the actual reasons of the discrepancy. In a very recent study presenting such measurement/simulation comparisons Akeroyd et al. [15], attribute differences between measurement and simulation to the part of the plant occurring after the strictly anechoic first 5 ms. However, they do not show results concerning the inclusion of that part in the inversion or whether the reduced accuracy obtained from the inversion of only the strictly anechoic part of the plant will render the use of long inverse filters or the correction of the phase unnecessary. On the other hand, Haztiantoniou and Mourjopoulos [16] present results suggesting that the inversion effectiveness of long (more than 100-ms) single-channel loudspeaker-and-room responses cannot be predicted with convolution simulations (even in controlled experimental conditions). They attribute this mainly to "weak nonstationarity" of the plant and suggest the use of complex smoothing in the design of the inverse.

As will be shown with our measurement results, the part of the plant lasting for several milliseconds after the "strictly anechoic" 4 ms has a significant influence on the measured inversion results. Simulating the inversion by convolving the designed inverse with the strictly anechoic 4-ms plant model (as has been done in the past [7], [8], [17]–[19]) gives unrealistically optimistic results, even in anechoic conditions. When this part of the plant is not taken into account in the design of the inverse, the inversion accuracy is limited to a degree that practically negates the use of long inverse filters and the correction of the plant's phase characteristics. Improved results are obtained when this part of the plant is included in the determination of the inverse. Dynamic range limitations that arise in the inversion of such a longer model of the plant make suitable the use of regularization and the mixed-phase inversion requirement that it entails [20], [21].

1 EXPERIMENTAL METHOD

1.1 Terminology

The basic form of a binaural reproduction system using loudspeakers is described in Fig. 1. Using z transforms to denote the relevant signals and system responses, the objective is to determine the 2×1 vector $\boldsymbol{y}(z)$ of source input signals so that the pair of signals $\hat{\boldsymbol{s}}(z)$ reproduced at the positions of the listener's ears are equal to a given pair of binaural signals denoted by the 2×1 vector $\boldsymbol{s}(z)$. We denote with $\boldsymbol{C}(z)$ the 2×2 plant matrix, which contains the electroacoustic transfer functions $C_{ij}(z)$, each relating the electrical signal input $y_j(z)$ to the acoustic pressure output $\hat{s}_i$. With $\boldsymbol{H}(z)$ we denote the 2×2 matrix that transforms the binaural signals $\boldsymbol{s}(z)$ into the source input signals, $\boldsymbol{v}(z)$. With this arrangement the matrix $\boldsymbol{H}(z)$ that achieves transparent reproduction $\hat{\boldsymbol{s}}(z) = \boldsymbol{s}(z)$ is the inverse of the plant matrix $\boldsymbol{H}(z) = [\boldsymbol{C}(z)]^{-1}$. As shown by the results of the next section, the use of a regularized inverse [8], [21], [22] can prove advantageous over the exact inverse. The inverse transfer function matrix that we seek to implement is then given by the equation

$$\boldsymbol{H}(z, \beta) = [\boldsymbol{C}^{\mathrm{H}}(z)\boldsymbol{C}(z) + \beta \boldsymbol{I}]^{-1}\boldsymbol{C}^{\mathrm{H}}(z) \tag{1}$$

where the regularization parameter β is a real nonnegative scalar.

The elements of matrix $\boldsymbol{H}$ of Eq. (1) are rational transfer functions sharing a common denominator equal to the determinant $\det[\boldsymbol{C}^{\mathrm{H}}\,\boldsymbol{C} + \beta \boldsymbol{I}]$ which, for $\beta > 0$, has been

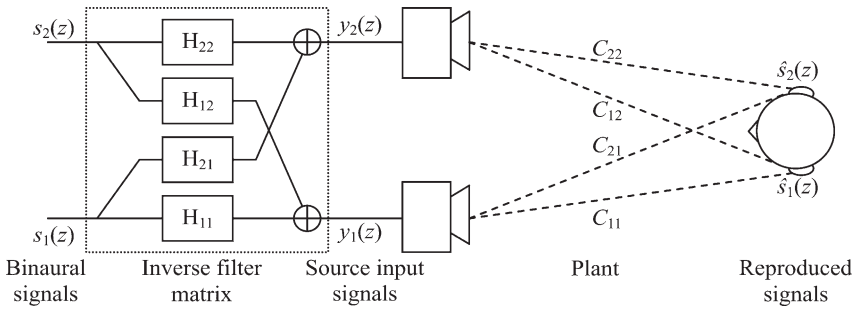

Fig. 1. Schematic of sound reproduction system under discussion.

found [20], [21] to be of mixed phase. A causal approximation of $\boldsymbol{H}$ with FIR filter elements can then be realized by introducing an adequate amount of modeling delay and truncating the corresponding infinitely long impulse responses to a finite length [8], [22]. In the following we denote with Δ the number of modeling delay samples and with L the total number of taps in the inverse filters. We denote with $\boldsymbol{X}(z)$ the 2×2 matrix containing the combined responses $X_{ij}(z)$ between the input s_j to the inverse filter network and the output $\hat{s}_i$ at the listener's ear position. The ideal form of combined response matrix will be $\boldsymbol{X}(z) = z^{-\Delta}\boldsymbol{I}$ with its diagonal elements $X_{11}(z)$ and $X_{22}(z)$ equal to pure delays $z^{-\Delta}$ and its off-diagonal elements $X_{12}(z)$ and $X_{21}(z)$ equal to zero.

The choice of the plant matrix model used for the determination of the inverse as well as the convolution-based simulation of the inverse are important for the results presented in the following. We denote with $\hat{\boldsymbol{c}}_K$ models of the plant matrix of different lengths (to be discussed further) and with $\hat{\boldsymbol{h}}_K$ the inverse matrix determined by substituting $\hat{\boldsymbol{c}}_K$ in Eq. (1). We denote with $\boldsymbol{x}_{\text{meas}}$ the measured result of the matrix $\boldsymbol{X}$ and with $\boldsymbol{x}_{\text{conv}}$ the estimate of the inversion result obtained by the convolution product $\hat{\boldsymbol{c}}_K{}^{*}\hat{\boldsymbol{h}}_M$, where the indices K and M signify the fact that the model of the plant matrix used for the prediction of the inversion effectiveness can be different from the one used for the determination of the inverse.

1.2 Experimental Arrangement

Our experimental arrangement comprised a KEMAR (Knowles Electronics) dummy head fitted with the DB-60 and DB-61 (small) right and left pinna models and two 1/2-inch B&K 4134 microphones. The sources were two Celestion 1 bookshelf loudspeakers. These are two-way loudspeakers with vertically aligned midrange and tweeter units. The loudspeakers where positioned with the centers of their drivers at a 0.265-m horizontal distance. KEMAR was placed on the axis of symmetry, with its head at the same height as the two sources and with the center of its head 1.5 m from the face of the loudspeakers. This arrangement gives a total span of 10° between the lines joining the center of KEMAR's head and the two loudspeakers (stereo dipole geometry [9]). The input signals to the loudspeakers were amplified by a Yamaha H5000 audio power amplifier, and the signals from the microphones were conditioned by two B&K 2636 measuring amplifiers. The loudspeakers, measuring amplifiers and KEMAR were placed inside the large anechoic chamber of the Institute of Sound and Vibration Research (ISVR). The area between the loudspeakers and KEMAR as well as the area surrounding the arrangement by approximately 1 m was covered with an absorbing foam, which is normally used for the lining of the second anechoic chamber of ISVR. The acoustic properties of this foam lining were not specifically examined. KEMAR's ears were approximately 1.5 m above the chamber's grid floor and approximately 1.2 m above the top edges of the absorbing foam wedges. The synchronous output of the signals $\boldsymbol{y}$ and the capture of the signals $\hat{\boldsymbol{s}}$ at a 48-kHz sample rate and 24-bit word length was controlled by a Huron (version by Lake Technology Limited) digital audio workstation.

The impulse responses $c_{11}(n)$ and $c_{21}(n)$ were determined with a pseudorandom stimulus of pink spectrum as $y_1(n)$ and $y_2(n) = 0$ using the *tcdeconv* function of Huron's Matlab (Mathworks, Inc.) toolbox to operate on the measured signals $\hat{\boldsymbol{s}}$. The impulse responses $c_{12}(n)$ and $c_{22}(n)$ were determined with the same process but with the pink stimulus in place of $y_2(n)$ and $y_1(n) = 0$. The pink pseudorandom stimulus used for the measurement of the plant was 5.46 s long (2^{18} samples), and the microphone signals were averaged 10 times. Using Huron's real-time FIR filtering module we were able to implement the inverse filter matrix $\boldsymbol{H}$ on-line and directly measure the true elements of the combined response matrix $\boldsymbol{X}$ in the same way as those of the plant matrix $\boldsymbol{C}$. In particular X_{11} and X_{21} were measured by implementing H_{11} and H_{21} and determining the response between the input s_1 (with no input at s_2) and the outputs $\hat{s}_1$ and $\hat{s}_2$, respectively, and correspondingly for X_{12} and X_{22}. A pseudorandom stimulus having a pink spectrum was used again for this measurement, this time 2.73 s long (2^{17} samples), and the output signals $\hat{s}_1$ and $\hat{s}_2$ were again averaged 10 times. The computation for the determination of the impulse responses was carried out in Matlab's double-precision floating-point arithmetic.

All the measurements we present (for both $\boldsymbol{C}$ and $\boldsymbol{X}$) were made with the same microphone gain and power amplification settings. The level of power amplification setting is further discussed in the following. The measured responses are presented in the numeric values obtained from the Huron workstation (full scale in the range $[-1, 1]$). The stimulus signal s used for the measurements of the plant matrix $\boldsymbol{C}$ occupied this full range. In the measurements of the inverted responses $\boldsymbol{X}$, both the input $\boldsymbol{s}$ to the filter elements of $\boldsymbol{H}$ and the resulting output digital signals $\boldsymbol{y}$ have to be within this same range. Thus a scaling factor H_{sc} had to be applied to the (generally amplified) filter output signals $\boldsymbol{y}$ in order to normalize them to the available digitization range. As will be discussed further, this scaling reflects the loss of dynamic range incurred by the inversion.

2 RESULTS

2.1 Measured Plant Model

In Fig. 2 we plot the impulse response $c_{11}(n)$ of the plant matrix corresponding to our experimental arrangement. As with all time-domain responses presented in this study, the results are plotted on a logarithmic y axis as $20 \log_{10}|c_{ij}(n)|$. This conforms to the scaling of the human auditory system and allows a clearer depiction of small-scale effects that are audible but invisible on a linear scale. The strictly anechoic part of the response starts at the 260th sample and persists roughly until the 450th sample. We denote this strictly anechoic part of the response $\boldsymbol{c}(260 \leq n \leq 450)$ by $\hat{\boldsymbol{c}}_{190}$. Even though the plant is measured inside the anechoic chamber, the response can be seen to contain a series of individual reflections caused by

the equipment inside the chamber and the chamber's grid floor as well as secondary reflections from KEMAR to the loudspeakers and back to the ears. The time difference of the first secondary peak (appearing at approximately the 460th sample) and the response onset corresponds to the length difference between the path from the loudspeakers to the top of the foam wedges covering the floor to KEMAR's ears and the direct path from the loudspeakers to the ears. Similarly, the timing of the third secondary peak (appearing at approximately the 660th sample) corresponds to the distance from the loudspeakers to KEMAR, back to the loudspeakers, and back to KEMAR. Those reflections last up to approximately the 1000th sample of $\boldsymbol{c}(n)$, and they are followed by a slowly decaying tail, which reaches the measurement noise floor by the 3000th sample. We denote the part of the response $\boldsymbol{c}(260 \leq n \leq 1000)$ by $\hat{\boldsymbol{c}}_{740}$, the part $\boldsymbol{c}(260 \leq n \leq 2000)$ by $\hat{\boldsymbol{c}}_{1740}$, and the part $\boldsymbol{c}(1 \leq n \leq 3000)$ by $\hat{\boldsymbol{c}}_{3000}$. Along with the value of the regularization parameter β and the number of samples L and Δ used for the FIR realization of the inverse matrix $\boldsymbol{H}$, the choice of substituting the strictly anechoic model $\hat{\boldsymbol{c}}_{190}$ or a longer model in Eq. (1) for the determination of $\boldsymbol{H}$ constitutes a third parameter affecting different possible realisations of $\boldsymbol{H}$. As we will see in the following, this choice does indeed have a significant effect on the achieved effectiveness of the inversion and more importantly on the optimum tuning of the remaining parameters.

As part of our experimental arrangement it was necessary to choose the level of amplification of the pseudorandom stimulus before it was fed to the loudspeakers. As will become apparent in the following, this choice influences the dynamic range of the plant measurement as well as the dynamic range of the reproduction after the inversion. The level of amplification during the measurement of $\boldsymbol{c}(n)$ reflects a compromise between 1) the desired rise in the level of the measured responses above the noise floor when the stimulus is further amplified by the power amplifier and 2) the undesired excitation of significant nonlinear behavior in the audio reproduction chain (especially the loudspeakers) as the level of amplification is increased. In measurements with random or pseudorandom excitation of the type presented here, such nonlinear behavior appears in the results as an increase in the level of the noise floor. We tested a number of different amplification settings and chose the one that maximized the clearance between the highest value of the measured responses and the noise floor of the measurement [22].

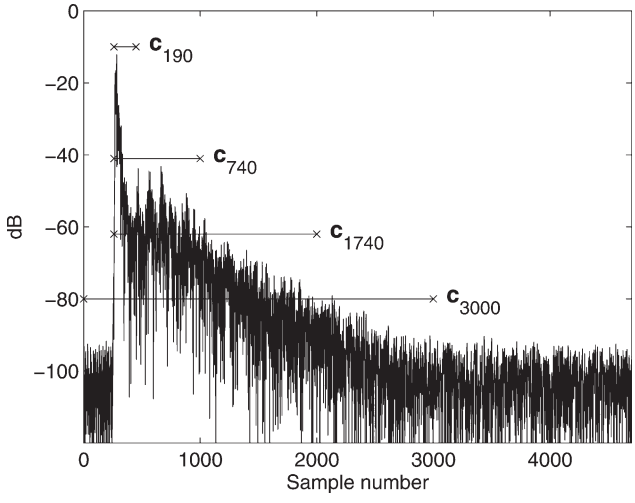

Fig. 2. Measured plant impulse response $c_{11}(n)$ (quantity $20 \log_{10}|c_{11}(n)|$ is plotted).

This noise floor can be seen in Fig. 2 to rise to approximately 95 dB below the [−1, 1] full-scale range of our measurement apparatus, which is of course far from the ideal −144-dB noise floor of the 24-bit quantization used in our arrangement. Even though analog-to-digital converters typically fail to achieve the full dynamic range of their bit-depth specification, such an increase in the noise floor level can only be attributed to nonlinear behavior of the reproduction system. This was corroborated by the fact that the level of the noise floor did not drop when the number of averages used in the capture of the microphone signals $\hat{\boldsymbol{s}}$ was increased. Nevertheless we note that the noise floor level in our arrangement matches the 96-dB theoretical optimum of the commonly used 16-bit arithmetic.

The dynamic range of our arrangement compares favourably with the two previously published works [23], [24] that present similar data. In the first [23], the dynamic range is estimated as the ratio of the total energy in 100 samples of the measured impulse response centered on the energetic part of the measurement to the energy in 100 samples of the initial delay that should ideally be zero. The quoted signal-to-noise ratio is 65 dB, whereas the same estimate in the left ipsilateral impulse response of the measurements presented here is 73 dB. In the second study [24] the dynamic range is estimated as the headroom between the transfer function of a representative measured response and the transfer function of a measurement with the output of the power amplifier connected to an 8-ohm resistor. The same procedure was followed with our arrangement and the results are presented in Fig. 3. A headroom of more than 60 dB between measurement and noise floor can be seen for the frequency range above 200 Hz, a result equivalent to the aforementioned study.

2.2 Accuracy of the Computer Simulation of the Inversion Process

Assuming that the whole reproduction system under discussion is linear and time invariant and that the true plant is modeled adequately by the chosen model of the plant matrix $\hat{\boldsymbol{c}}(n)$, the directly measured combined response matrix $\boldsymbol{x}_{\text{meas}}(n)$ corresponding to a specific realization of the inverse $\hat{\boldsymbol{h}}(n)$ will be equal to the convolution product $\boldsymbol{x}_{\text{conv}}(n) = \hat{\boldsymbol{c}}(n)^{*}\hat{\boldsymbol{h}}(n)$. We compare the time-domain and frequency-domain representations of the two quantities $\boldsymbol{x}_{\text{meas}}$ and $\boldsymbol{x}_{\text{conv}}$ in Figs. 4 and 5. The inverse filter matrix realization in this instance was $\hat{\boldsymbol{h}}_{190}(n)$. The regularization value was set to $\beta = 0$, the inverse filter length was $L = 4000$, and the modeling delay $\Delta = 2000$.

The scaling factor H_{sc} was equal to 1/130 (−42 dB). In Fig. 4 (a) and (b) the directly measured elements $x_{11,meas}(n)$ and $x_{12,meas}(n)$ are plotted. This result is compared to two instances of the convolution product $\boldsymbol{x}_{conv}(n)$. The first, in Fig.4 (c) and (d), depicts the elements $x_{11,conv}(n)$ and $x_{12,conv}(n)$ of the convolution product matrix $\boldsymbol{x}_{conv}(n) = \hat{\boldsymbol{c}}_{190}(n)^* \hat{\boldsymbol{h}}(n)$ delayed by 190 samples (note the difference in the y-axis scaling). The second, in Fig. 4(e) and (f), depicts the same elements of the convolution product matrix $\boldsymbol{x}_{conv}(n) = \hat{\boldsymbol{c}}_{3000}(n)^* \hat{\boldsymbol{h}}(n)$.

The convolution product $\hat{\boldsymbol{c}}_{190}(n)^* \hat{\boldsymbol{h}}(n)$ is seen to be practically equal to the ideal inversion result in both the time and frequency domains. This should be expected as the only difference between $\hat{\boldsymbol{h}}(n)$ and the exact inverse of $\hat{\boldsymbol{c}}_{190}(n)$ is the truncation of the infinitely long exact responses $h_{ij}(n)$ at their 2000th anticausal and 2000th causal samples. Hence the only imperfections in the predicted result $\hat{\boldsymbol{c}}_{190}(n)^* \hat{\boldsymbol{h}}(n)$ are the truncation effects that are visible 2000 samples either side of the single-sample spike in the time domain. Those truncation effects are further discussed in the following section.

The same result is demonstrated by the Fourier transforms $X_{11}(e^{j\omega})$, $X_{12}(e^{j\omega})$ of the elements $x_{11,conv}(n)$ and $x_{12,conv}(n)$ of the convolution product $\boldsymbol{c}_{190}(n)^* \hat{\boldsymbol{h}}(n)$. These are plotted in Fig. 5(b), which shows perfect equalization of the ipsilateral transmission path X_{11} and suppression of the crosstalk path X_{12} by more than 40 dB across the audible frequency range. Such simulation results have been commonly employed as estimates of the effectiveness of the inversion, especially in earlier sources in the literature [7], [8], [25]. Based on such an estimate, one would infer that such an inverse more than meets the requirement for the reproduction of the binaural signals (ipsilateral equalization within ±1 dB and crosstalk cancellation of more than 30 dB; see, for example, discussion in [7], [15] and references therein). Longer inverse filters would show further suppression of the time-domain truncation effects and, hence, further suppression of the crosstalk.

Contrary to that, the directly measured results $\boldsymbol{x}_{meas}$ of Fig. 4(a) and (b) and Fig. 5(a) yield a much more modest evaluation of the effectiveness of the inversion and reveal a number of deviations from the ideal form $z^{-\Delta}\boldsymbol{I}$. These elements of error are discussed in the next section. The point to be made here is that the simulation results obtained by $\boldsymbol{x}_{conv}(n) = \hat{\boldsymbol{c}}_{3000}(n)^* \hat{\boldsymbol{h}}(n)$ are in very good agreement with the directly measured results above the noise floor in the time domain and across the audio frequency in the frequency domain. The measurement noise floor is at the same level as the plant measurement results of Fig. 2, that is, the linearity characteristics of the system have remained the same as in the plant measurement, and the inverse filtering stage is performing according to its design. The actual performance, however, does not

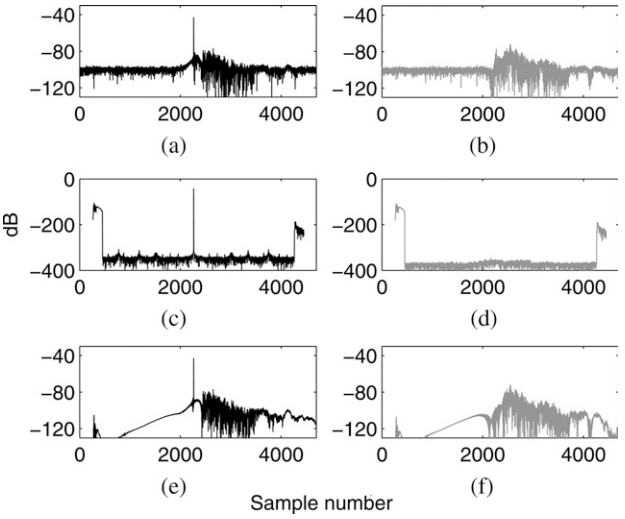

Fig. 4. Impulse response elements $x_{11}(n)$ and $x_{12}(n)$ of equalization matrix $\mathbf{x}$. (a), (b) Directly measured matrix $\boldsymbol{x}_{meas}$. (c), (d) Convolution simulation matrix $\boldsymbol{x}_{conv} = \hat{\boldsymbol{c}}_{190}{}^* \hat{\boldsymbol{h}}_{190}$. (e), (f) Convolution simulation matrix $\boldsymbol{x}_{conv} = \hat{\boldsymbol{c}}_{3000}{}^* \hat{\boldsymbol{h}}_{190}$. (a), (c), (e) — ipsilateral equalization elements $x_{11}(n)$, (b), (d), (f) — crosstalk elements $x_{12}(n)$.

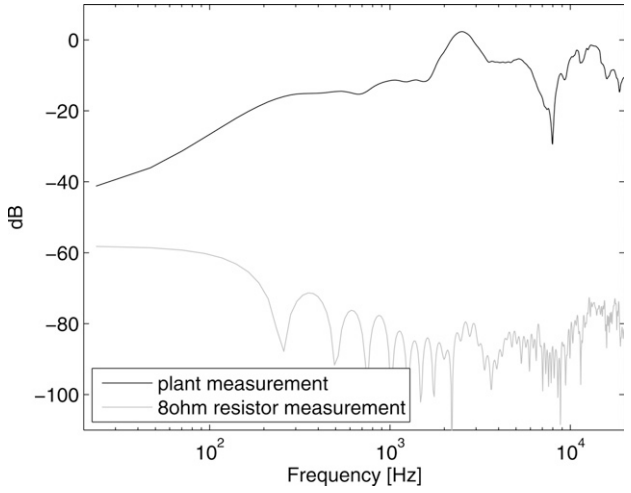

Fig. 3. Dynamic range between measured plant frequency response and noise floor obtained by means of a dummy load measurement.

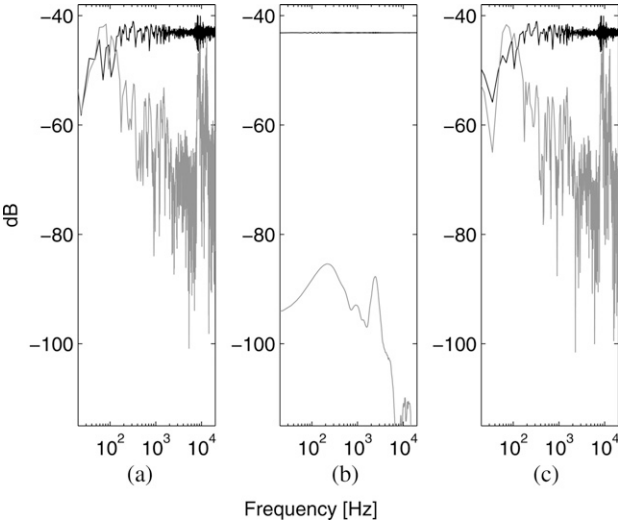

Fig. 5. Frequency responses $|X_{11}(e^{j\omega})|$ and $|X_{12}(e^{j\omega})|$ for the same cases as in Fig. 4.

surpass the aforementioned "target" for adequately transparent binaural reproduction and is misrepresented by the simulation $\hat{\boldsymbol{c}}_{190}{}^{*}\hat{\boldsymbol{h}}_{190}$. Except where otherwise stated, all the results presented in what follows are obtained by the estimate $\hat{\boldsymbol{c}}_{3000}(n)^{*}\hat{\boldsymbol{h}}(n)$.

2.3 Required Number of Taps for FIR Inverse Filters

The time-domain results of Fig. 4 show two types of deviation from the ideal delta-spike form in the ipsilateral equalization and from the perfect suppression in the crosstalk path. The first of these is visible as an abrupt rise in the decaying anticausal part of the response. This is due to the fact that the FIR filters realizing the inverse matrix are windowed versions of the infinitely long impulse response elements of the exact inverse matrix $\boldsymbol{H}$. The anticausal support of each element h_{ij} is truncated at the 2000th anticausal sample, and hence the truncation error effect appears in $\boldsymbol{x}(n)$ at 2000 samples before the position of the delta spike. Note that the corresponding truncation effect should also occur in forward time, that is, 2000 samples after the position of the delta spike, but this is not visible in the plot as it is swamped by the artifact associated with a second type of error demonstrated by the results.

This second type of error appears as a broad rise in both responses $x_{11}(n)$ and $x_{12}(n)$ that starts 190 samples after the delta spike and is superimposed on the subsequent decay of the responses. This second type of error is due to the fact that the model of the plant used for the determination of the inverse only models the part of the plant from its onset at the 260th sample and up to the 450th sample. Hence the part of $\boldsymbol{c}$ after these 190 samples is not inverted and appears in the combined responses $\boldsymbol{x}(n)$ as an artifact starting 190 samples after the delta spike in the ipsilateral path and at the same time in the crosstalk path. To describe this analytically one can write the plant $\boldsymbol{c}$ as in the equation

$$\boldsymbol{C}(z) = z^{-260}\boldsymbol{C}_{190}(z) + z^{-450}\boldsymbol{C}_{\text{tail}}(z) \tag{2}$$

where the responses in both matrices $\boldsymbol{c}_{190}$ and $\boldsymbol{c}_{\text{tail}}$ appear with their onset at sample $n = 0$.

When $\boldsymbol{c}$ is convolved with a given realization of the inverse matrix, say $\boldsymbol{h}_1$, the resulting matrix $\boldsymbol{x}$ will thus be

$$\begin{aligned}\boldsymbol{X}(z) &= \boldsymbol{C}(z)\boldsymbol{H}_1(z) \\ &= z^{-260}\boldsymbol{C}_{190}(z)\boldsymbol{H}_1(z) + z^{-450}\boldsymbol{C}_{\text{tail}}(z)\boldsymbol{H}_1(z).\end{aligned} \tag{3}$$

It is then easy to observe that since $\boldsymbol{h}_1$ is determined as the truncated and delayed by 2000 samples inverse of $\boldsymbol{c}_{190}$ (that is, $\boldsymbol{C}_{190}(z)\boldsymbol{H}_1(z) = z^{-2000}\boldsymbol{I}$), $x_{11}(n)$ will be the superposition of a delta spike delayed by 2260 samples and the artifact described by the second term of Eq. (3), which is what is depicted in Fig. 4.

The first type of error described here has been studied in a single-channel equalization context in [26] where, in accordance with the theory of temporal masking, it was found that such truncation effects appearing before the equalization spike can result in severe deterioration of the perceived quality of the reproduced material, even at levels much lower than the equalized signal. In the results presented in Fig. 4 the 2000-sample length of the anticausal part of the filters realizing the inverse matrix $\boldsymbol{H}$ has been so chosen that the truncation effect appears at a level well below the noise floor of our measurement arrangement. As was discussed, this noise floor level practically coincides with the −96-dB minimum obtainable quantization noise floor in standard audio reproduction equipment. On the other hand, the situation is quite different for the truncation effect error in the forward-time part of the response $x_{11}(n)$. First of all, due to temporal masking, truncation effects appearing after the delta spike have much less influence on the perceived quality of the reproduced material. But more importantly, any truncation effect of this type is physically covered by the error caused by the omission of $\boldsymbol{c}_{\text{tail}}$ from the inversion. In fact it can be shown [22] that in most cases both the time-domain and the frequency-domain results of $\boldsymbol{X}$ are literally indistinguishable if the forward-time part of $\boldsymbol{H}$ is truncated to any length higher than 190 samples. In any such case the truncation error is lower than the error due to the omission of $\boldsymbol{c}_{\text{tail}}$ from the inversion. Therefore even in this controlled anechoic reproduction environment and with no suboptimality imposed on the inversion by means of regularization, the level of accuracy that can be realistically achieved by the inversion of only the strictly anechoic part of the plant before other elements of error become more significant is reached by a realization with roughly $L = 2000$ taps per element of $\boldsymbol{H}$.

A complementary description of the effectiveness of the inverse filtering stage is given by the frequency responses X_{11} and X_{12} plotted in Fig. 5. In those plots the ipsilateral equalization can be seen to be effective to a margin of ±2 dB in the frequency region above 200 Hz with the exception of a slightly higher deviation in the region of 8–10 kHz. The same holds for the suppression of the crosstalk, which is between 15 and 20 dB for the same frequency region but with the crosstalk transmission increasing in the region from 9–11 kHz. The inversion ineffectiveness in the region below 200 Hz is due to the combined effect of the first ill-conditioned frequency of the stereo dipole geometry at 0 Hz and the reduced response of the loudspeakers in this region. The poorly controlled region around 10 kHz is a combined result of the low signal-to-noise ratio of the plant response due to the pinna notch [27] and of the second ill-conditioned frequency of the stereo dipole geometry at 11 kHz [9], [10]. The left-ear pinna notch can be seen in Fig. 3 to appear at 8 kHz. Similar notches at slightly higher frequencies are present in the right-ear responses [22]. In these two regions (below 200 Hz and 10 kHz) the error due to the mismatch between the plant that is inverted $(\hat{\boldsymbol{c}}_{190})$ and the true plant is amplified. It should be noted that the ill-conditioning at 0 Hz appears in all axisymetric two-source geometries [9], [10] but is aggravated at close spacing geometries such as the stereo dipole. Wider geometries are less susceptible to this problem, but at the expense

of inversion errors due to the second ill-conditioning frequency appearing at lower frequencies [9], [28].

The results of this section described the main sources of error in the implementation of the inverse filtering stage of the binaural audio reproduction system under discussion. These results show that when only the strictly anechoic part of the plant is inverted (3–5-ms part of the plant describing solely the reflection/diffraction from the listener's head and torso [23], [24], [29]), the use of a longer, more accurate FIR realization of the inverse matrix $\boldsymbol{H}$ cannot, in itself, alleviate the imperfections discussed here.

2.4 Use of Regularization

One issue that is made evident from the results of the previous section — and one that very often goes unnoticed — is the loss of dynamic range incurred by the inverse filtering stage. Noting again that both the plant and the inversion measurements presented here were made with the same level of power amplification and microphone gain settings, it becomes evident that the loss of dynamic range incurred by the inversion is shown by the difference in the level of the frequency responses obtained in the plant measurement (Fig. 3) and the level of the ipsilateral term of the combined matrix $\boldsymbol{X}$ when the inversion is in place, as shown in Fig. 5. With audio signals (such as music) occupying most of the available dynamic range of the digital (and also the electronic and electromechanical) devices in the audio reproduction chain, a loss of more than 20 dB as suggested by comparison of the aforementioned figures should be expected to result in degradation of the perceived quality of the material presented. As will be shown with the results of this section, this loss of dynamic range becomes more severe when the inversion of longer models of the plant is attempted, but can be addressed successfully with the use of regularization.

In Fig. 6 we plot the inversion results obtained when $\boldsymbol{c}_{190}$ is inverted with different values of regularization, $\beta = 0$, $\beta = 10^{-4}$, and $\beta = 10^{-2}$. The effect of regularization is then clearly seen in these results as that of effectively redistributing the total available sound reproduction power away from the "power needy" frequency regions and into an increase of the reproduction level (and thus the dynamic range) of the frequency regions the control of which is less demanding in acoustic power. The horizontal lines in Fig. 6 show an estimate of the reproduction level corresponding to each different regularization value that is obtained using the following analysis, leading to Eq. (10).

One estimate of the loss of dynamic range (LDR) in the digital reproduction system can be given by the scaling needed in order to keep the two-channel signal $\boldsymbol{y}$ in the same range as the input signals $\boldsymbol{s}$. Using standard norm notation [30], [31] and given that $\max_i[\max_n[|s_i(n)|]] = \|\boldsymbol{s}(n)\|_\infty = 1$ we will have

$$\begin{aligned}\mathrm{LDR} &= \max_{i=1,2}\big[\max_n(|y_i(n)|)\big] \\ &= \|\boldsymbol{y}(n)\|_\infty = \|\boldsymbol{h}(n)^*\boldsymbol{s}(n)\|_\infty. \end{aligned} \tag{4}$$

Clearly, the value of $\|\boldsymbol{y}(n)\|_\infty$ in Eq. (4) will depend not only on the inverse filter $\boldsymbol{h}(n)$ but also on the specific form and the statistics of the input signal $\boldsymbol{s}(n)$, which are not known. It is straightforward to see [32] that for a single-channel filter $h(n)$ of N coefficients the worst case of amplification of the output $y(n)$ for any input signal $s(n)$ which is normalized to unity amplitude $\max_n[|s(n)|] = \|s(n)\|_\infty = 1$ will be bounded by

$$\begin{aligned}\max_{\|s(n)\|_\infty=1}\big[\|y(n)\|_\infty\big] &= \max_{\|s(n)\|_\infty=1}\big[\|h(n)^*s(n)\|_\infty\big] \\ &= \|h(n)\|_1 = \sum_{n=0}^{N-1}|h(n)|. \end{aligned} \tag{5}$$

The bound in Eq. (5) is genuine and can be seen [32] to be obtained for an input containing a segment $s(n)$,

$$\begin{aligned} s(n) &= \mathrm{sign}[h(n)] \\ &= \begin{cases} 1, & \text{if } h(n) \geq 0 \\ -1, & \text{if } h(n) < 0 \end{cases}, \quad 0 \leq n \leq N-1. \end{aligned} \tag{6}$$

The extension of Eqs. (5) and (6) to two channels is straightforward but not followed here as such an estimate of the loss of dynamic range would be unrealistically stringent given that a signal matched to a filter such as that of Eq. (6) cannot be expected to occur in practice. A more realistic estimate of the required amplification (and hence loss of dynamic range) between the input $\boldsymbol{s}$ and output $\boldsymbol{y}$ of the inverse filtering stage can instead be provided by the ratio of the energy norms $\|\boldsymbol{y}(n)\|_2/\|\boldsymbol{s}(n)\|_2$, which can be shown [30], [31] to be bounded by

$$\frac{\|\boldsymbol{y}(n)\|_2}{\|\boldsymbol{s}(n)\|_2} \leq \|\boldsymbol{H}(e^{j\omega})\|_\infty = \max_i\{\max_\omega[\sigma_{Hi}(e^{j\omega})]\} \tag{7}$$

where $\sigma_{Hi}(e^{j\omega})$, $i = 1, 2$, are the singular values of $\boldsymbol{H}(e^{j\omega})$. Writing the singular value decomposition of the

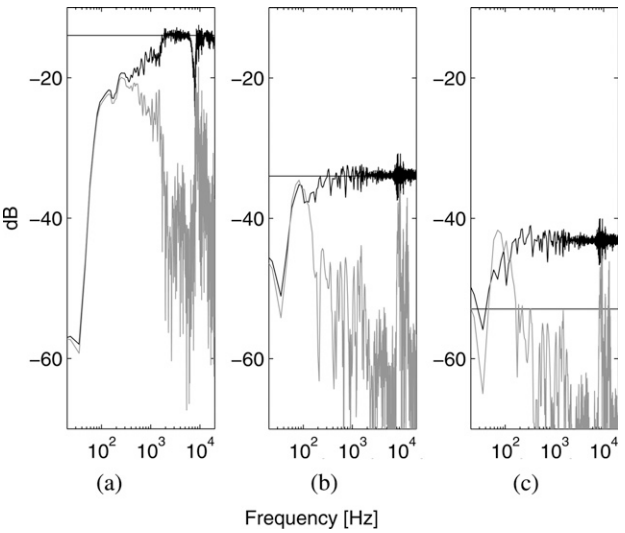

Fig. 6. Frequency responses $|X_{11}(e^{j\omega})|$ (black line) and $|X_{12}(e^{j\omega})|$ (grey line) obtained with $\hat{\boldsymbol{h}}_{190}$ for different regularization values. (a) $\beta = 10^{-2}$. (b) $\beta = 10^{-4}$. (c) $\beta = 0$. Horizontal lines show loss of dynamic range as estimated by Eqs. (7) and (10).

plant matrix $\boldsymbol{C}(e^{j\omega})$ with $\sigma_i(e^{j\omega})$, $i = 1, 2$, its singular values

$$\boldsymbol{C}(e^{j\omega}) = \boldsymbol{U}(e^{j\omega}) \begin{bmatrix} \sigma_1(e^{j\omega}) & 0 \\ 0 & \sigma_2(e^{j\omega}) \end{bmatrix} \boldsymbol{V}^H(e^{j\omega}) \quad (8)$$

it follows straightforwardly [33] that the SVD of the regularized inverse $\boldsymbol{H}(e^{j\omega}, \beta)$ of Eq. (1) will be given by

$$\boldsymbol{H}(e^{j\omega}, \beta) = \boldsymbol{V}(e^{j\omega}) \begin{bmatrix} \frac{\sigma_1(e^{j\omega})}{\sigma_1^2(e^{j\omega})+\beta} & 0 \\ 0 & \frac{\sigma_2(e^{j\omega})}{\sigma_2^2(e^{j\omega})+\beta} \end{bmatrix} \boldsymbol{U}^H(e^{j\omega}) \quad (9)$$

thus determining the singular values $\sigma_{Hi}(e^{j\omega})$ of Eq. (7), which are therefore given by

$$\sigma_{Hi}(e^{j\omega}) = \frac{\sigma_i(e^{j\omega})}{\sigma_i^2(e^{j\omega}) + \beta}. \quad (10)$$

The maximum value of Eq. (10) over frequency and over the index i for $\beta = 0$, $\beta = 10^{-4}$, and $\beta = 10^{-2}$ is plotted in Fig. 6 with a horizontal line, and it can be seen to agree perfectly with the actual reproduction level for the cases where $\beta > 0$ and in order of magnitude for the case $\beta = 0$. It is also easy to see that in the regularized case, $\beta > 0$, the maximum value of Eq. (10) can be estimated by simply maximizing the function of Eq. (11) over $x > 0$ and hence without the need to perform the singular value decomposition of the plant matrix $\boldsymbol{c}$,

$$f(x) = \frac{x}{x^2 + \beta} \leq \frac{1}{2\sqrt{\beta}}. \quad (11)$$

A similar estimate has been suggested previously [10], but without the supporting analysis presented here. Note that the maximum of $f(x)$ will only be obtained at $x = \sqrt{\beta}$, which practically means that for vanishingly small values of regularization, the maximum will only be obtained if the Fourier transform of the plant matrix $\boldsymbol{C}(e^{j\omega})$ has a singular value that is correspondingly small for some frequency ω, $\sigma_{Hi}(e^{j\omega}) = \sqrt{\beta}$. If this is not the case, then the estimate of Eq. (11) would prove too strict. Note further that the actual amplification through $\boldsymbol{H}$ (and hence the loss of dynamic range) will also depend on the frequency distribution and the statistics of the input signal $\boldsymbol{s}$. In the results presented here the input was predefined to have most of its energy content concentrated at low frequencies as well as a rather low crest factor. Even though this is not uncharacteristic of typical audio signals such as speech or music, a complete investigation of the exact loss of dynamic range in such cases is not presented here. This analysis does, however, provide the basis for a simple strategy for the choice of the regularization value β, something that has been considered previously [8] to be largely a matter of trial and error. That is, with knowledge of the plant matrix $\boldsymbol{C}$ and with no need for prior computation of the inverse matrix $\boldsymbol{H}$, one can estimate the loss of dynamic range (at least in order of magnitude) for zero regularization by means of Eqs. (7) and (8). If this turns out to be more than can be accommodated by the available dynamic range of the reproduction system (as is usually the case), then the value of β that accomplishes the required increase in the level of the reproduced signals can be determined by the use of Eq. (11).

Returning to the results of Fig. 6, we see that by setting $\beta = 10^{-4}$ we reduce the loss of dynamic range by approximately 10 dB without any sacrifice in terms of the achieved effectiveness of the inversion when compared to the nonregularized case, which was discussed in the previous section. On the other hand, a further increase of the regularization value to $\beta = 10^{-2}$ results in a further 20-dB reduction in the loss of dynamic range but also in a deterioration of the control in the low-frequency and 10-kHz regions. In that case only the region of 1–8 kHz is effectively controlled. This would render such a choice unsuitable for most types of audio material. Hence the optimal choice of the regularization value will ultimately depend on the frequency content of the input signals and the information they carry. The length L of the inverse filters required to adequately suppress the truncation effects as described in the previous section is reduced in these two cases to $L = 1500$ taps per element of $\boldsymbol{h}$ for the $\beta = 10^{-4}$ case and to $L = 700$ taps per element of $\boldsymbol{h}$ for the $\beta = 10^{-2}$ case.

It may be worth reiterating the fact that these results correspond to the ideal reproduction scenario whereby the anechoic plant corresponding to the individual listener, the specific reproduction system, and the exact geometry is used for the determination of the inverse. Such controlled reproduction conditions can only realistically be expected to hold when the plant is measured in situ. In such a case one would have knowledge of the whole length of the plant model matrix $\boldsymbol{c}$ and could thus use that for the determination of the inverse. This is demonstrated in Fig. 7, where we plot the results obtained with $\hat{\boldsymbol{h}}_{740}$ and for the same three cases of

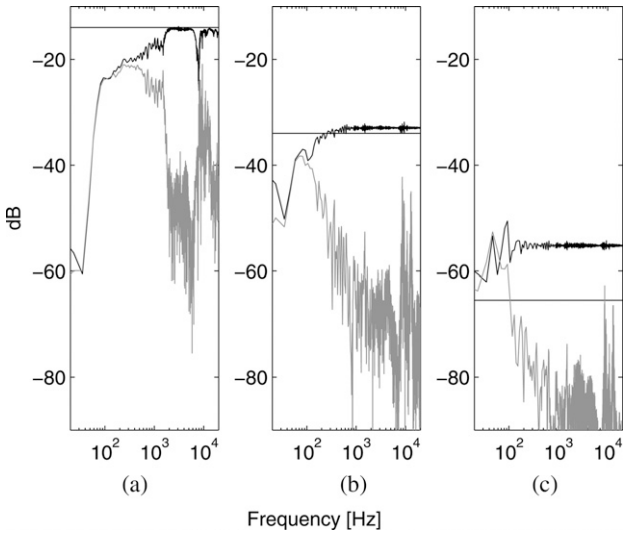

Fig. 7. Frequency responses $|X_{11}(e^{j\omega})|$ (black line) and $|X_{12}(e^{j\omega})|$ (grey line) obtained with $\hat{\boldsymbol{h}}_{740}$ for different regularization values. (a) $\beta = 10^{-2}$. (b) $\beta = 10^{-4}$. (c) $\beta = 0$. Horizontal lines show loss of dynamic range as estimated by Eqs. (7) and (10).

regularization value as before ($\beta = 0$, $\beta = 10^{-4}$, and $\beta = 10^{-2}$). The results obtained with the regularization set to $\beta = 10^{-4}$ [plotted in Fig. 7(b)] show a tighter ipsilateral equalization curve and increased crosstalk cancellation compared to the corresponding case of Fig. 6(b). They can be seen to come close to the ideal of completely transparent reproduction with an ipsilateral equalization error of less than ± 1 dB and increased crosstalk cancellation reaching nearly 30 dB in the region of 1–8 kHz. The crosstalk cancellation gradually decreases below 1 kHz but remains effective down to approximately 100 Hz. Note that in this case the required length L of the inverse filters rises to a few thousand taps ($L = 5000$, $\Delta = 2000$). The introduction of regularization has limited the inversion of the low-frequency end, but has preserved the dynamic range of the inversion to the same level as the $\hat{\boldsymbol{h}}_{190}$ case. On the other hand the nonregularized case [shown in Fig. 7(c) and obtained with the inverse filter length set to 10 000 taps] shows a higher level of reproduction at the low-frequency end. But this is at the cost of an overall reduction in the level of reproduction of more than 40 dB compared to the level of reproduction without the inversion (as was shown in Fig. 3). This would practically mean that users would have to increase the volume during the reproduction with the inversion by 40 dB in order to achieve the same level of reproduction they would when listening to nonbinaural material. The requirement for such a dynamic headroom is an issue that has to be considered for the practical implementation of this inversion case. We note that the use of loudspeakers with better low-frequency extension than the bookshelf units used in this study should give better results in this respect, as they would require less amplification for the correction of their (better) low-frequency response to flat response.

Finally, in Fig. 8(a) we compare the directly measured results obtained with $\hat{\boldsymbol{h}}_{1740}$, namely, an inverse

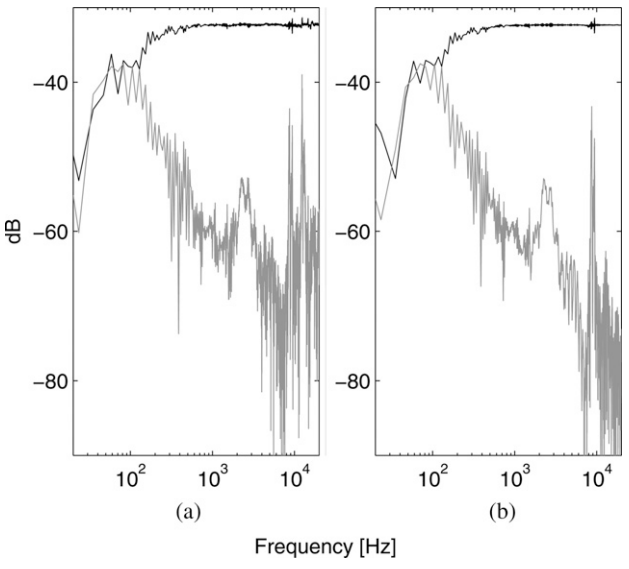

Fig. 8. Frequency responses $|X_{11}(e^{j\omega})|$ (black line) and $|X_{12}(e^{j\omega})|$ (grey line) obtained with $\hat{\boldsymbol{h}}_{1740}$ for $\beta = 10^{-4}$. (a) Directly measured results. (b) Results of convolution $\hat{\boldsymbol{c}}_{3000}^{*}\hat{\boldsymbol{h}}_{1740}$.

computed on the basis of a plant model extending from the 260th until the 2000th sample (approximately 35-ms duration). In Fig. 8(b) we plot the simulation results $\boldsymbol{x}_{\text{conv}} = \hat{\boldsymbol{c}}_{3000}^{*}\hat{\boldsymbol{h}}_{1740}$. In this case the regularization is set to the value of $\beta = 10^{-4}$ and the inverse filter length used for the measurement and the simulation is set to 4000 taps (the truncation effects appear at the level of the measurement noise floor). The scaling factor H_{sc} was equal to 1/38 (−32 dB). The results obtained by the inversion of this longer model of the plant do not offer significant improvement over the previously presented $\hat{\boldsymbol{h}}_{740}$ case. But the good agreement between measurement and simulation up to above 10 kHz in the inversion of such a longer model of the plant suggests that the inversion of in situ measured plant models that describe the acoustics due to surfaces and objects close to the listener (but not necessarily the long decay of the room) should be free of the problems raised in [16]. However, the noticeable discrepancy between measurement and simulation above 10 kHz (a region that is of limited importance for the application at hand) suggests that the inversion of such a length of the plant could be more susceptible to nonlinear effects (possibly atmospheric attenuation) or time variance because of changes in environmental parameters such as temperature [34]. The use of regularization allows the inversion to be realized at a level that preserves the dynamic range of the reproduction as much as the nonregularized inversion of $\hat{\boldsymbol{h}}_{190}$.

2.5 Modeling the Plant with Equalized HRTFs

As was mentioned, in order to obtain inversion results such as those presented so far, one would have to be in a position to measure the plant in situ. A simplification that is usually adopted in practice in order to sidestep such a requirement is to model the plant matrix $\boldsymbol{C}$ that is used for the determination of the inverse in Eq. (1) with equalized HRTFs rather than with the exact plant responses that contain the response of the individual reproduction system. In such an implementation the user would have his or her individual HRTF data set measured in a dedicated facility, and this data set would be equalized with respect to the response of the measurement apparatus. Evidently when binaural material is reproduced over the user's own audio reproduction system, an inverse determined on the basis of his or her HRTFs can, at best, control the part of the plant response that is particular to his or her HRTFs, but not the characteristics or imperfections that are particular to his or her audio reproduction system. In this section we give results that quantify the influence of such a simplification in the design of the inverse matrix $\boldsymbol{H}$.

The difference between the raw measured plant responses and the corresponding HRTF is exemplified in Fig. 9. In that figure we plot the C_{11} element of the plant matrix $\boldsymbol{C}$ as described in Fig. 1 and the corresponding left-ear HRTF equalized with respect to the free-field response from the left source to the center of KEMAR's head with no head present. As is typically the case with publicly available equalized HRTFs (compact data set in

the MIT HRTF database [23], CIPIC database [29]) as well as previous studies on human HRTFs [24], the example presented here is based on HRTFs determined on the basis of the strictly anechoic plant model $\boldsymbol{c}_{190}$. As can be seen in the figure, the low-frequency rolloff of the plant response is solely due to the low-frequency rolloff of the free-field response of the corresponding loudspeaker. Hence when this response is deconvolved from the plant response, the HRTF's frequency content converges to 0 dB at low frequencies. Such low-frequency behavior is common in any properly measured HRTF and reflects the fact that, at very long wavelengths, the disturbance of the sound field by the presence of the listener's head and torso is negligible.

In Fig. 10 we show the results obtained for $X_{11}(f)$ and $X_{12}(f)$ when the inverse is computed by substituting a version of $\boldsymbol{C}$ containing the corresponding HRTFs in Eq. (1). Two cases are considered, corresponding to regularization values of $\beta = 0$ and $\beta = 1$. The effectiveness of the inversion is very similar in the two cases with the high-regularization case having the advantage of a smaller loss of dynamic range and requiring just 600 taps for the adequate suppression of the time-domain truncation effects. The corresponding number of taps for the nonregularized example was 1500. No improvement was obtained by using longer inverse filters in any of the two cases.

In both cases the shape of the free-field response of the loudspeaker can be seen to have passed through the inversion unaltered, and the level of crosstalk cancellation is significantly reduced in the ill-conditioned frequency regions (below 200 Hz and around 11 kHz) compared to the results of the previous sections (Fig. 5). At these frequency regions the error due to the mismatch between the plant that is inverted (equalized HRTF modeled plant) and the true plant is amplified.

The results of this section suggest that the issue regarding the control of the low-frequency region does not really apply when one settles for inverting a plant modeled with HRTFs rather than the plant containing the responses of the individual reproduction system. When such a model of the plant is used, the low-frequency response of the true plant will be misrepresented in most practical cases where the loudspeakers used have limited low-frequency extension. Hence an attempt to control this frequency range would be futile. The results of Fig. 10 suggest that in such a case one should get perceptibly equivalent results by bypassing the low-frequency part of the inverse filtering stage and reproducing the low-frequency content of the input without modification.

2.6 Correction of the Phase

A last point to be addressed regarding the properties of the inversion has to do with the phase of the inversion. In all the results presented, the inverse matrix $\boldsymbol{H}$ was implemented so that both the magnitude and the phase characteristics of Eq. (1) are satisfied (up to a constant delay of Δ samples). However, if one chooses to ignore the correction of the phase of the ipsilateral transmission paths, a causal and stable implementation that requires no modeling delay can be implemented by replacing the common denominator of Eq. (1) with its minimum-phase approximation. We note that such an option cannot be combined straightforwardly with the use of regularization because, as has been shown previously [20], [21], the introduction of regularization replaces minimum-phase zeros of the denominator polynomial with pairs of zeros, one inside and one outside the unit circle. We also note that minimum-phase inverse designs have been proposed and used previously (including the recursive realization for the shuffler topology described by Cooper and Bauck [2]).

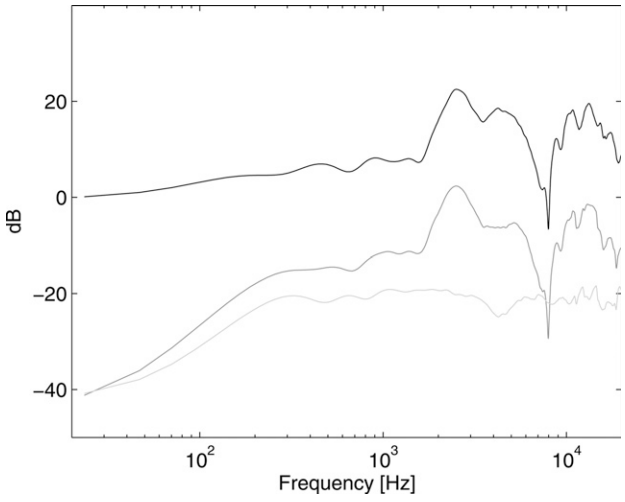

Fig. 9. Plant raw response $C_{11}(e^{j\omega})$ (dark grey line), free-field response from left source to center of head with no head present (light grey line), and resulting HRTF (black line).

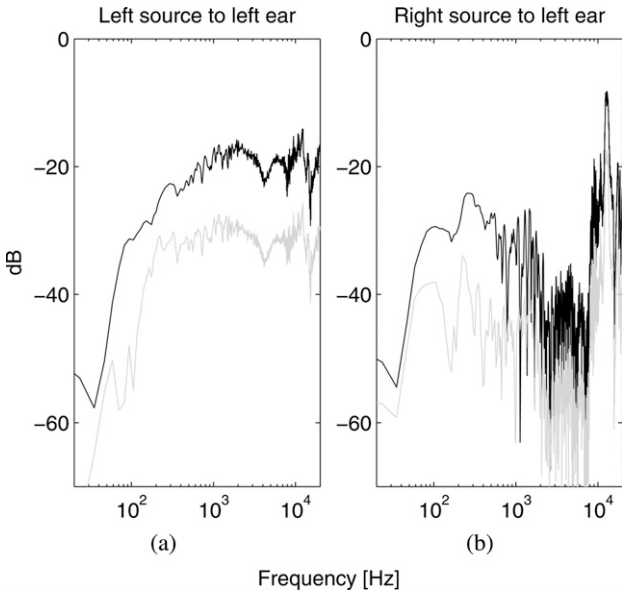

Fig. 10. Frequency responses $|X_{11}(e^{j\omega})|$ and $|X_{12}(e^{j\omega})|$ when inverse is computed on the basis of HRTFs with regularization set to $\beta = 0$ (grey line) and $\beta = 1$ (black line). (a) Left source to left ear. (b) Right source to left ear.

Taking then the nonregularized case, instead of implementing the exact inverse

$$\boldsymbol{H}(e^{j\omega}) = \frac{1}{\det[\boldsymbol{C}(e^{j\omega})]}\,\mathrm{adj}\left[\boldsymbol{C}(e^{j\omega})\right] = \frac{1}{\det[\boldsymbol{C}(e^{j\omega})]}\begin{bmatrix} C_{22}(e^{j\omega}) & -C_{12}(e^{j\omega}) \\ -C_{21}(e^{j\omega}) & C_{11}(e^{j\omega}) \end{bmatrix} \quad (12)$$

with adj denoting the adjoint of a matrix, one can choose to implement the minimum-phase approximation

$$\boldsymbol{H}_{\min}(e^{j\omega}) = \frac{1}{\mathrm{minphase}\{\det[\boldsymbol{C}(e^{j\omega})]\}}\,\mathrm{adj}\left[\boldsymbol{C}(e^{j\omega})\right]. \quad (13)$$

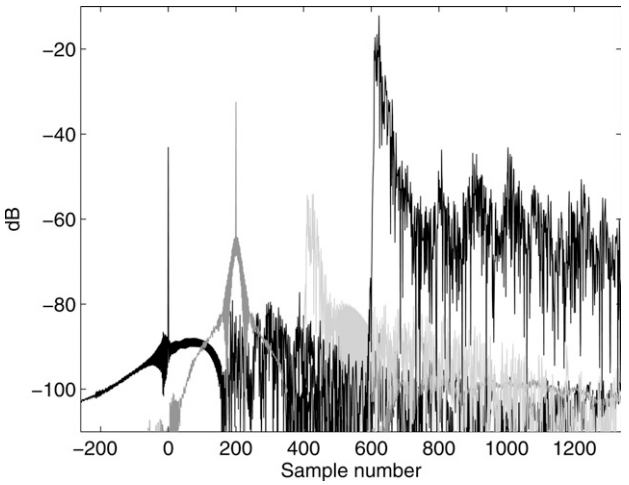

Fig. 11. Impulse response $x_{11}(n)$ obtained with three different realizations of inverse $\boldsymbol{H}$. 1) $\hat{\boldsymbol{h}}_{190}$ with $\beta = 0$ (black line, delta spike centered at $n = 0$); 2) $\hat{\boldsymbol{h}}_{1740}$ with $\beta = 10^{-4}$ (dark grey line, delta spike centered at $n = 200$); 3) by minimum-phase approximation described in text (light grey line, response onset positioned at $n = 400$). Impulse response $c_{11}(n)$ of noninverted plant (black line, response onset positioned at $n = 600$).

It then follows directly that the magnitude of the elements of the resulting combined response matrix $\boldsymbol{X}_{\min}(e^{j\omega}) = \boldsymbol{C}(e^{j\omega})\,\boldsymbol{H}_{\min}(e^{j\omega})$ is equal to the magnitude of the elements of the original combined response matrix $\boldsymbol{X}(e^{j\omega}) = \boldsymbol{C}(e^{j\omega})\,\boldsymbol{H}(e^{j\omega})$. In other words, the level of the crosstalk transmission paths $|X_{ij}(e^{j\omega})|$, $i \neq j$, as well as the magnitude of the equalized ipsilateral paths $|X_{ij}(e^{j\omega})|$, $i = j$, achieved by $\boldsymbol{H}_{\min}(e^{j\omega})$ will be the same as that achieved by $\boldsymbol{H}(e^{j\omega})$, with the only difference between the two being the phase of the ipsilateral paths $\angle X_{ij}(e^{j\omega})$, $i = j$. The question then is the degree to which the phase characteristics of the ipsilateral equalization are distorted by such an approximation.

In Fig. 11 we plot the results obtained for the element $x_{11}(n)$ of the combined response matrix $\boldsymbol{x}$ with three different realizations of the inverse $\boldsymbol{h}$, 1) $\hat{\boldsymbol{h}}_{190}$ with $\beta = 0$ [the same case as in Fig. 5 (c)], 2) $\hat{\boldsymbol{h}}_{1740}$ with $\beta = 10^{-4}$ [the same case as in Fig. 8(b)], and 3) the minimum-phase approximation to the inverse of Eq. (13). The minimum-phase impulse responses were implemented up to a length of 5000 taps (lengths from 2000 to 10 000 taps gave identical results). In Fig. 11 we also plot the unprocessed plant impulse response $c_{11}(n)$, and Fig. 12 shows the group delay (excluding the initial delay) of the same four responses. As was discussed, all three first cases are associated with a flat-magnitude frequency response, but as can be seen in Fig. 11, only the first two (corresponding to the mixed-phase inversion) result in a delta spike in the time domain whereas the one corresponding to the minimum-phase inversion is smeared in time in a manner similar to the nonequalized ipsilateral plant response.

As can be seen in Fig. 12, the group delay of all four cases is very similar, with the only exception of the regularized mixed-phase inversion result of $\hat{\boldsymbol{h}}_{1740}$, which is slightly tighter around the initial delay value and seems to extend down to lower frequencies. It should of course be noted that the unprocessed plant case presented here is not directly comparable to the three other cases, which include

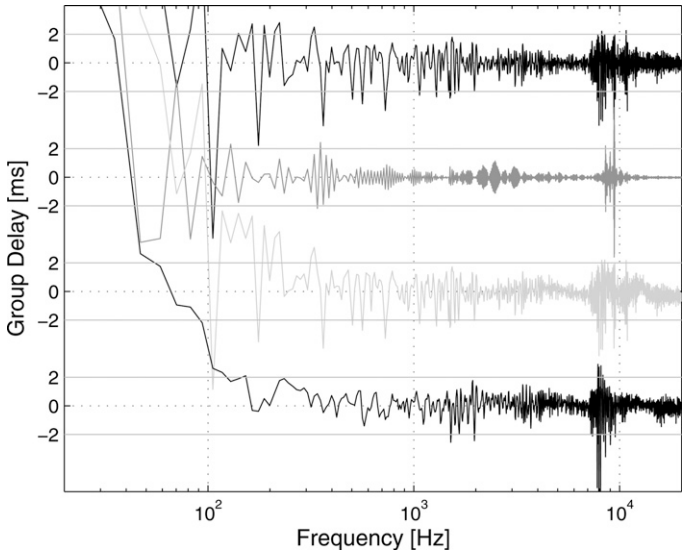

Fig. 12. Group delay of $X_{11}(e^{j\omega})$ and $C_{11}(e^{j\omega})$ for same cases as in Fig. 11 plotted from top to bottom with 8-ms vertical offset. [Case (1) on top followed downward by cases (2), (3), and (4)]. Zero values on y axis correspond to onset time of each response.

the inverse filtering stage. This is because in the former case the transmission path to the left ear amounts to the superposition of the ipsilateral path C_{11} and the crosstalk path C_{12}, whereas in the latter cases the crosstalk is suppressed. Nevertheless the results show that only when $\boldsymbol{c}_{1740}$ is available does the mixed-phase inverse achieve some improvement compared to the minimum-phase approximation in correcting the phase response of the ipsilateral transmission path. It is not clear to us whether such an improvement would indeed be beneficial.

3 DISCUSSION

We presented a detailed description of the inversion results that are realistically achievable by the mixed-phase inverse design of [8] for a closely spaced loudspeaker geometry. We showed that an accurate model of the inversion process can be obtained by assuming linear and time-invariant operation of the system. For that to hold true, we showed that the plant model used in the convolution simulation of the inversion has to contain the whole time response down to the level of the noise floor.

The directly measured inversion results that we presented verified that in completely controlled and anechoic conditions of implementation, ipsilateral equalization within ± 2 dB and crosstalk suppression of 15–20 dB are readily achievable for nearly all of the audio bandwidth. However, those results required knowledge of the plant with regard to the individual listener, the specific audio reproduction system, and the exact source/listener geometry. In current practical implementations in which the inverse is determined on the basis of generic HRTFs (not describing the characteristics of the individual listener, the audio reproduction setup, and the acoustics of the area around the listener) such a condition is severely violated. The actual inversion results are degraded significantly in that case (see, for example, discussion in [15] and references therein).

The aforementioned requirement for knowledge of the true plant to be inverted is practically equivalent to the ability to measure the plant in situ. We showed that when such a measurement is possible, the inversion results can be improved by including in the determination of the inverse the minor reflections that are bound to exist even in anechoic conditions of reproduction. In such a case we showed that the ipsilateral path equalization error can be reduced to less than ± 1 dB, and the crosstalk suppression can be increased to up to 30 dB for nearly the entire audio range. We showed directly measured results verifying that.

These latter inversion results represent a realistic estimate of the inverse design under discussion, with further improvement hindered by the loss of dynamic range incurred by the inversion. We showed how this loss of dynamic range can be moderated by the use of regularization without any significant loss in the accuracy of the inversion, and we presented a simple form of an algorithm for the automatic selection of the appropriate value of the regularization parameter. Our results made clear that even in those controlled implementation conditions, which lend themselves to the highest possible inversion accuracy, the required number of taps for the FIR realization of the inverse filters is not prohibitively high. These range from a few hundreds to a few thousands per element of $\boldsymbol{H}$ in all regularized cases.

It is a well-known fact [9], [10] that crosstalk cancellation using closely spaced loudspeakers poses increased amplification requirements for the low-frequency region. We argued that when equalized HRTFs are used for the determination of the inverse, the low-frequency part of the true plant response (containing the loudspeaker response) will be misrepresented in many practical applications of the system. Hence the increased amplification requirement for inverting the low-frequency region would not be justified. As was made apparent by our regularized inversion results, such an arrangement would allow all the available acoustic control power to be used for the inversion of the frequency region above 200 Hz and hence ease the dynamic range requirement of the inversion. In such an event, then, perceptually equivalent results should be obtained by the use of what is commonly referred to as a 2.1 system, with the input signals' frequency content below 200 Hz bypassing the inverse filtering and being reproduced unprocessed.

Our results concerning the phase of the inversion challenge the justification of correcting the phase of the strictly anechoic part of the plant. However, our results showed that regularization has a beneficial effect in the mitigation of the loss of dynamic range incurred when a longer model of the plant is inverted. This would be the case when the plant measurement describes the acoustics of surfaces positioned close to the reproduction geometry (that is, a desktop environment). Such plants would 1) be increasingly nonminimum phase due to strong reflections appearing at a later part of the plant responses [35], and 2) need to be inverted up to a higher length, thus making the use of regularization and the mixed-phase inversion it entails suitable.

4 CONCLUSIONS

The results presented in this paper focused on the working assumption that an exact plant model is available that describes the characteristics of the individual listener and the audio reproduction setup. We showed that the inversion of such a plant, which achieves measurements of the "target" of ± 1-dB equalization and 30-dB crosstalk cancellation required for binaural reproduction over loudspeakers, requires a more careful choice of the inverse design parameters than previously thought. The requirement for knowledge of the true plant can be met by implementing the system using in situ plant measurements. Given the versatility and widespread availability of PC platforms, such an option is feasibly implementable in computer multimedia applications. We believe that such a practical mode of application of the transaural reproduction principle should be further investigated.

5 ACKNOWLEDGMENT

The authors gratefully acknowledge the comments and suggestions by two anonymous reviewers, which resulted

in significant improvements in the presentation of this paper.

6 REFERENCES

[1] J. Bauck and D. H. Cooper, "Generalized Transaural Stereo and Applications," *J. Audio Eng. Soc.*, vol. 44, pp. 683–705 (1996 Sept.).

[2] D. H. Cooper and J. L. Bauck, "Prospects for Transaural Recording," *J. Audio Eng. Soc.*, vol. 37, pp. 3–19 (1989 Jan./Feb.).

[3] B. S. Atal and M. R. Schroeder, "Apparent Sound Source Translator," U.S. patent 3 236 949 (1966).

[4] P. Damaske, "Head-Related Two-Channel Stereophony with Loudspeaker Reproduction," *J. Acoust. Soc. Am.*, vol. 50, pp. 1109–1115 (1971).

[5] W. G. Gardner, *3-D Audio Using Loudspeakers* (Kluwer Academic, Boston, MA, 1998).

[6] H. Møller, "Fundamentals of Binaural Technology," *Appl. Acoust.*, vol. 36, pp. 171–218 (1992).

[7] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis, "Inverse Filter Design for Immersive Audio Rendering over Loudspeakers," *IEEE Tran. Multimedia*, vol. 2, pp. 77–87 (2000).

[8] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna Bustamante, "Fast Deconvolution of Multichannel Systems Using Regularization," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 189–194 (1998).

[9] O. Kirkeby, P. A. Nelson, and H. Hamada, "The "Stereo Dipole" — A Virtual Source Imaging System Using Two Closely Spaced Loudspeakers," *J. Audio Eng. Soc.*, vol. 46, pp. 387–395 (1998 May).

[10] T. Takeuchi and P. A. Nelson, "Optimal Source Distribution for Binaural Synthesis over Loudspeakers," *J. Acoust. Soc. Am.*, vol. 112, pp. 2786–2797 (2002).

[11] P. Mannerheim and P. A. Nelson, "Virtual Sound Imaging Using Visually Adaptive Loudspeakers," *Acta Acustica/Acustica*, vol. 94, pp. 1024–1039 (2008).

[12] M. R. Bai and C. C. Lee, "Development and Implementation of Crosstalk Cancellation System in Spatial Audio Reproduction Based on Subband Filtering," *J. Sound Vib.*, vol. 290, pp. 1269–1289 (2006).

[13] O. Kirkeby, P. Rubak, L. G. Johansen, and P. A. Nelson, "Implementation of Cross-Talk Cancellation Networks Using Warped FIR Filters," presented at the AES 16th International Conference on Spatial Sound Reproduction, Rovaniemi, Finland, 1999 Apr. 10–12.

[14] T. Lentz, I. Assenmacher, and J. Sokoll, "Performance of Spatial Audio Using Dynamic Cross-Talk Cancellation," presented at the 119th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 53, p. 1202 (2005 Dec.), convention paper 6541.

[15] M. A. Akeroyd, J. Chambers, D. Bullock, A. R. Palmer, A. Q. Summerfield, P. A. Nelson, and S. Gatehouse, "The Binaural Performance of a Crosstalk Cancellation System with Matched or Mismatched Setup and Playback Acoustics," *J. Acoust. Soc. Am.*, vol. 121, pp. 1056–1069 (2007).

[16] P. D. Hatziantoniou and J. N. Mourjopoulos, "Errors in Real-Time Room Acoustics Dereverberation," *J. Audio Eng. Soc.*, vol. 52, pp. 883–899 (2004 Sept.).

[17] O. Kirkeby and P. A. Nelson, "Digital Filter Design for Inversion Problems in Sound Reproduction," *J. Audio Eng. Soc.*, vol. 47, pp. 583–595 (1999 July/Aug.).

[18] T. Takeuchi, P. A. Nelson, and H. Hamada, "Robustness to Head Misalignment of Virtual Sound Imaging Systems," *J. Acoust. Soc. Am.*, vol. 109, pp. 958–971 (2001).

[19] O. Kirkeby, P. A. Nelson, F. Orduna Bustamante, and H. Hamada, "Local Sound Field Reproduction Using Digital Signal Processing," *J. Acoust. Soc. Am.*, vol. 100, pp. 1584–1593 (1996).

[20] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna Bustamante, "Fast Deconvolution of Multi-Channel Systems Using Regularization," Tech. Rep. 255, ISVR, University of Southampton, Southampton, UK (1996 Apr.).

[21] P. A. Nelson, "Active Control of Acoustic Fields and the Reproduction of Sound," *J. Sound Vib.*, vol. 177, pp. 447–477 (1994).

[22] T. Papadopoulos, "Inverse Filtering in Virtual Acoustic Imaging Systems," Ph.D. dissertation, University of Southampton, Southampton, UK (2006).

[23] B. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone," Tech. Rep. 280, MIT Media Lab, Cambridge, MA (1994 May).

[24] H. Møller, M. F. Sorensen, D. Hammershoi, and C. B. Jensen, "Head-Related Transfer Functions of Human Subjects," *J. Audio Eng. Soc.*, vol. 43, pp. 300–321 (1995 May).

[25] J. Rose, P. Nelson, B. Rafaely, and T. Takeuchi, "Sweet Spot Size of Virtual Acoustic Imaging Systems at Asymmetric Listener Locations," *J. Acoust. Soc. Am.*, vol. 112, pp. 1992–2002 (2002).

[26] S. G. Norcross, G. A. Soulodre, and M. C. Lavoie, "Subjective Investigations of Inverse Filtering," *J. Audio Eng. Soc.*, vol. 52, pp. 1003–1028 (2004 Oct.).

[27] E. A. Lopez-Poveda and R. Meddis, "A Physical Model of Sound Diffraction and Reflections in the Human Concha," *J. Acoust. Soc. Am.*, vol. 100, pp. 3248–3259 (1996).

[28] M. R. Bai, C. W. Tung, and C. C. Lee, "Optimal Design of Loudspeaker Arrays for Robust Cross-talk Cancellation Using the Taguchi Method and the Genetic Algorithm," *J. Acoust. Soc. Am.*, vol. 117, pp. 2802–2813 (2005).

[29] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF Database," in *Proc. 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (2001).

[30] S. J. Elliott, *Signal Processing for Active Control* (Academic Press, San Diego, CA, 2001).

[31] S. Skogestad and I. Postlethwaite, *Multivariable Feedback Control: Analysis and Design* (Wiley, Chichester, UK, 1996).

[32] L. B. Jackson, *Digital Filters and Signal Processing* (Kluwer, Boston, MA, 1986).

[33] P. A. Nelson and S. H. Yoon, "Estimation of Acoustic Source Strength by Inverse Methods: Part I: Conditioning of the Inverse Problem," Tech. Rep. 278, ISVR, University of Southampton, Southampton, UK (1998 Oct.).

[34] G. Elko, E. Diethorn, and T. Gansler, "Room Impulse Response Variation due to Thermal Fluctuation and Its Impact on Acoustic Echo Cancellation," presented at the International Workshop on Acoustic Echo and Noise Control (IWAENC2003) (Kyoto, Japan, 2003 Sept.).

[35] S. T. Neely and J. B. Allen, "Invertibility of a Room Impulse Response," *J. Acoust. Soc. Am.*, vol. 66, pp. 165–169 (1979).

THE AUTHORS

T. Papadopoulos

Timos Papadopoulos received a degree in electrical and computer engineering from the National Technical University of Athens, Greece, in 1996, and M.Sc. and Ph.D. degrees from the Institute of Sound and Vibration Research, University of Southampton, Southampton, UK, in 2000 and 2006, respectively. His Ph.D. research involved the design and experimental evaluation of multichannel inverse filters for virtual acoustic imaging systems, including recursive filters realized in forward and backward time.

During his Ph.D. work he worked as an audio signal processing researcher in academic and industrial projects funded by Samsung, the Institute of Hearing Research UK, the Department of Education Studies of the University of Southampton, as well as various audio companies. In 2005 he obtained a research fellowship at ISVR, and he has since then been working on the topic of mammalian echolocation. His research interests are in the areas of audio signal processing, three-dimensional audio reproduction systems, and bioacoustics.

Dr. Papadopoulos is a member of the Audio Engineering Society.

P. A. Nelson

Philip A. Nelson is deputy vice-chancellor of the University of Southampton, Southampton, UK, with particular responsibility for research and enterprise. He previously served as director of the University's Institute of Sound and Vibration Research and holds the post of professor of Acoustics. He has personal research interests in the fields of acoustics, vibrations, fluid dynamics, and signal processing, and is the author or coauthor of 2 books, over 100 papers in refereed journals, 30 granted patents, and over 200 other technical publications.

Professor Nelson is a Fellow of the Royal Academy of Engineering, a Chartered Engineer, a Fellow of the Institution of Mechanical Engineers, a Fellow of the Institute of Acoustics, a Fellow of the Acoustical Society of America, and a member of the Audio Engineering Society. He is the recipient of both the Tyndall and the Rayleigh medals of the Institute of Acoustics, and he served as president of the International Commission for Acoustics from 2004 to 2007.