Obtaining Binaural Room Impulse Responses From B-Format Impulse Responses Using Frequency-Dependent Coherence Matching

Fritz Menzer, Christof Faller, and Hervé Lissek

***Abstract*—Measuring binaural room impulse responses (BRIRs) for different rooms and different persons is a costly and time-consuming task. In this paper, we propose a method that allows to compute BRIRs from a B-format room impulse response (B-format RIR) and a set of head-related transfer functions (HRTFs). This enables to measure the room-related properties and head-related properties of BRIRs separately, reducing the amount of measurements necessary for obtaining BRIRs for different rooms and different persons to one B-format RIR measurement per room and one HRTF set per person. The BRIRs are modeled by applying an HRTF to the direct sound part of the B-format RIR and using a linear combination of the reflections part of the B-format RIR. The linear combination is determined such that the spectral and frequency-dependent interaural coherence cues match those of corresponding directly measured BRIRs. A subjective test indicates that the computed BRIRs are perceptually very similar to corresponding directly measured BRIRs.**

***Index Terms*—Acoustic reflection, B-format, binaural reverberation, binaural room impulse response (BRIR), diffuse sound, early reflection, head-related transfer function (HRTF), interaural coherence, late reverberation, linear decoding, room impulse response (RIR).**

I. Introduction

BINAURAL room impulse responses (BRIRs) are important tools for high-quality 3-D audio rendering [1]. BRIRs take into account both the properties of the listener (or dummy head) as well as the properties of the room in which the BRIRs have been recorded and give the listener the impression of being in the room and hearing a sound source in the position where the sound source used for the BRIR recording was placed. Head-related transfer functions (HRTFs) on the other hand are recorded in an anechoic environment and can be used to simulate listening to a loudspeaker in an anechoic environment. HRTFs completely lack room-related properties.

In this paper, we propose a method that allows to compute BRIRs using room impulse responses measured with a B-format microphone (B-format RIRs) and HRTF sets. This means that recording the listener-specific properties (HRTFs) is independent from recording room-specific properties (B-format RIRs). In particular, this very much simplifies the task of providing individualized BRIRs for a large number of different acoustic environments for many different persons—something which is relevant for providing high quality 3-D audio for a large user base.

Previously, [2] and [3] proposed a method which can generate RIRs for multi-channel loudspeaker setups with up to approximately 20 channels [4] from B-format RIRs. This method, called spatial impulse response rendering (SIRR), uses a decomposition into direct and diffuse parts. It distributes the direct part on the loudspeakers using vector base amplitude panning [5] and de-correlates the diffuse part to obtain several uncorrelated diffuse impulse responses.

The goal of the method proposed here is to generate BRIRs relative to any look direction of the head in a simple and robust way. Unlike SIRR, our technique cannot produce impulse responses for multi-channel loudspeaker systems. By applying the correct HRTFs to the impulse responses generated by SIRR it is possible to simulate the target loudspeaker setup in anechoic conditions and therefore generate an approximation of a BRIR. Thus, SIRR can be used to perform the same task as the method proposed in this paper. The proposed method is simpler than SIRR and eliminates the intermediate step of a multi-channel impulse response. This is very important because it also eliminates the need for a de-correlation method such as reverberators or phase randomization, which is necessary in a setup with more than two channels and which may introduce artifacts to the impulse response [4].

Given the B-format RIR of a specific room and an HRTF set, BRIRs individualized to the same listener as the HRTF set are generated as follows. The B-format RIR is separated in time into a direct sound part, and a reflections part, containing the early and late reflections of the RIR. The direct sound part of the BRIR is modeled by applying to the direct sound the HRTFs corresponding to the estimated direction of arrival. The reflections part of the BRIR is modeled as a linear combination of the late B-format signal channels such that the relevant spectral cues and perceptual spatial cues are the same as would be expected for a BRIR measured in the same room as the B-format

Manuscript received May 22, 2009; revised September 05, 2009; accepted November 26, 2009. Date of publication April 29, 2010; date of current version October 29, 2010. This work was supported by the Swiss National Science Foundation (SNSF) under Grant 200021-109406. This paper follows the concepts of reproducible research. The results presented in the paper are reproducible using the code and impulse responses available online at http://rr.epfl.ch/32. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

F. Menzer and C. Faller are with the Audiovisual Communications Laboratory (LCAV), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland (e-mail: fritz.menzer@epfl.ch; christof.faller@epfl.ch).

H. Lissek is with the Electromagnetics and Acoustics Laboratory (LEMA), Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland (e-mail: herve.lissek@epfl.ch).

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.2010.2049410

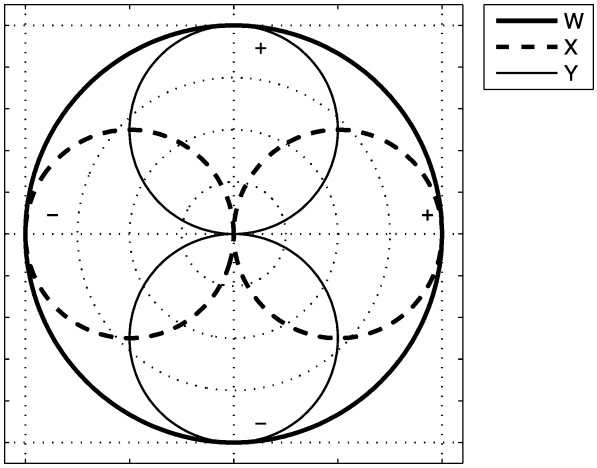

Fig. 1. Directional responses of the W, X, and Y channels of B-format in the horizontal plane (without $\sqrt{2}$ factors, i.e., all responses have a maximum of 1).

RIR was measured. The considered spatial cues are the left and right power spectra and the interaural coherence (IC) [6].

This paper is organized as follows. Section II describes the proposed method to compute BRIRs in detail. In Section III, the results produced by the proposed method are examined from a signal processing point of view, while a subjective test to evaluate the proposed method is described in Section IV. The conclusions are in Section V. Appendix A describes the room impulse measurements performed for the evaluation of the proposed method.

II. Obtaining BRIRs From B-Format Room Impulse Responses and HRTFs

A. B-Format Room Impulse Responses

A B-format room impulse response (B-format RIR) is a room impulse response measured with a B-format microphone [7], [8]. Ideally, it corresponds to a four-channel room impulse response measured with four coincident microphones: one omnidirectional microphone $(w(n))$ and three dipole microphones $(x(n), y(n), z(n))$, pointing in the X, Y, and Z directions of a Cartesian coordinate system. An example of the directional responses in the horizontal plane is shown in Fig. 1. Note that usually B-format is defined such that the dipoles have a gain which is $\sqrt{2}$ larger than the omnidirectional gain (not shown in the Figure, i.e., all directional responses have a maximum of 1).

Inspired by current models of reverberation [9], we consider room impulse responses to consist of a large peak corresponding to the direct sound as well as several delayed and filtered copies of this first peak, corresponding to the early reflections, and a diffuse reverberation tail, which may overlap with the early reflections.

B. B-Format RIR Separation

Since the direct sound is processed in a different way than the reverberation, it is necessary to separate the B-format RIR into these two parts.

The split point between the direct sound and the late RIR is determined as the lowest local minimum of the energy envelope of $w(n)$ in the 10 ms after the absolute maximum of the energy envelope of $w(n)$. An example of such a separation can be seen in Fig. 2. The 10-ms time interval after the direct sound was determined experimentally based on the RIRs at our disposal. For other rooms or other source and listener positions, there may be a need to slightly change the length of this interval in order to correctly separate the direct sound from the first reflection.

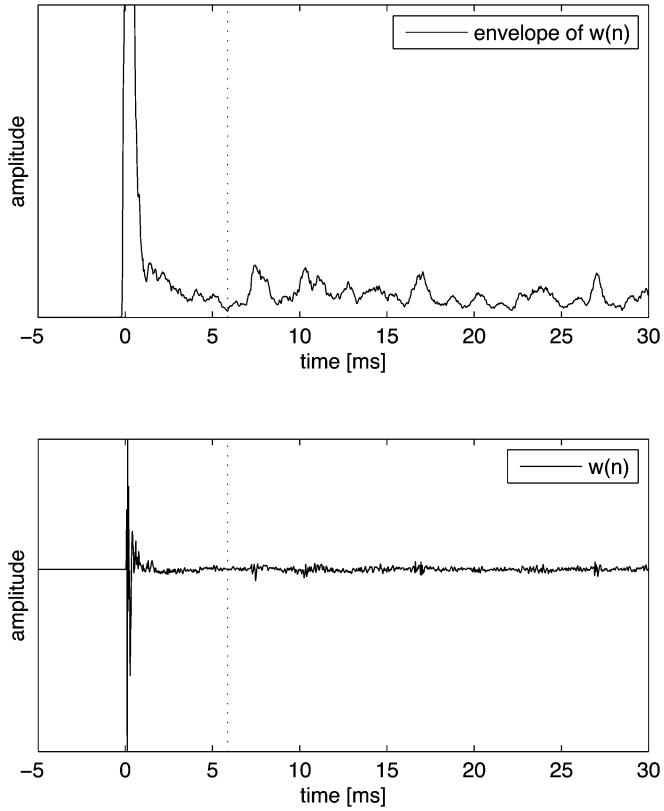

Fig. 2. Separation of the B-format RIR in direct sound and reflections parts. Top panel: envelope of $w(n)$ of B-format RIR. Bottom panel: $w(n)$ of B-format RIR. The separation is made at the lowest local minimum of the envelope of $w(n)$ in the first 10 ms after the direct sound.

As opposed to an earlier implementation of the proposed method [10] which, similar to SIRR, extracted the individual early reflections and convolved them with HRTFs corresponding to their directions of arrival, in this paper both early and late reflections are processed by a single frequency dependent linear B-format decoding described in Section II-D.

Two reasons led to this decision: when estimating the direction of arrival of the early reflections embedded in diffuse sound, errors are unavoidable. In practice, the linear decoding delivered better perceptual results than the directional modeling of the individual early reflections (i.e., the method presented here is both a simplification as well as an improvement compared to the method presented in [10]). Furthermore, as will be shown in Section III, the linear decoding method performs reasonably well even on the waveform level.

C. Modeling the Direct Sound

The early BRIR corresponding to the direct sound is generated as follows. For the direct sound in the B-format RIR the direction of arrival is estimated by

$$\phi = \arg(I_x + iI_y) \tag{1}$$

$$\psi = \arg\left(\sqrt{I_x^2 + I_y^2} + iI_z\right) \tag{2}$$

where I_x, I_y, and I_z are the components of the acoustic intensity vector $\vec{I}$ and are calculated as

$$I_x = \sum_{n\in D} x(n)w(n)$$
$$I_y = \sum_{n\in D} y(n)w(n)$$
$$I_z = \sum_{n\in D} z(n)w(n) \tag{3}$$

on the time interval D that corresponds to the direct sound.

Finally, the part of $w(n)$ corresponding to the direct sound is filtered with the HRTF closest to the estimated direction of arrival of the direct sound. Since the HRTF set used has resolution of 5° in the horizontal plane, for sources in the horizontal plane, the deviation from the estimated direction is 2.5° or less.

With respect to the direction of arrival estimate and the rendering of the direct sound, the presented method is equivalent to [2].

D. Modeling the Late BRIR

The late part of the BRIRs are obtained by linearly processing the late B-format RIR such that three conditions are fulfilled.

- The power spectra of the generated left and right late BRIR are the same as the power spectra of the true left and right BRIR.
- The coherence between the left and right generated late BRIRs is the same as the coherence between the true left and right late BRIRs at each frequency.
- At each frequency, the temporal envelope of the generated late BRIR is the same as for the true late BRIR.

In other words, the proposed method is designed to reproduce the energy decay relief for each channel as well as the frequency-dependent interaural coherence of the true late BRIR. The energy decay relief was introduced as an important perceptual cue for mono reverb by [11] and the frequency dependent interaural coherence has been shown to be a major cue for late binaural reverb [12].

In the following, we are computing the left and right true late BRIR power spectra and coherence as a function of frequency between the left and right late BRIR. Then, it is shown how to compute late BRIRs by linear B-format decoding from the B-format room impulse responses such that the power spectra and coherence are the same as in the true late BRIRs. The decay of the late BRIR is the same as the decay of the B-format RIR for each frequency. The linear B-format decoding is time-independent and therefore has no impact on the decay which thus will be automatically correct, implying that also the frequency dependent reverberation time of the generated BRIR will be correct.

All of the linear B-format decoding described hereafter was implemented using a fast Fourier transform (FFT), which is the natural choice since the B-format decoding is time-independent and frequency-dependent. However, alternative implementations, e.g., in STFT domain, are possible.

The proposed method for modeling the late BRIR is different from the diffuse sound rendering of SIRR because the late BRIR is calculated only by a linear decoding of the B-format RIR, with the aim of obtaining a BRIR with the correct interaural coherence directly, without using reverberators or other de-correlation techniques, which would be a possible source of artifacts.

1) Computation of the True BRIR Parameters: In the following it is assumed that the late BRIR is ideally diffuse, i.e., sound arrives from all directions with the same power and sound arriving from each direction is independent of the sound arriving from all other directions. Further, diffuse sound is approximated by only considering directions for which HRTFs are available. The left and right HRTFs are denoted $L_i(\omega)$ and $R_i(\omega)$, where $i \in \{1, 2, \ldots, I\}$ is the direction index and I is the number of HRTFs in the set.

In the tests performed for this paper, an HRTF set with an angular resolution of 5° in the horizontal plane was used. In previous tests, the proposed method was applied using the CIPIC HRTF set [13] whose angular resolution in the horizontal plane varies between 5° and 20°.

Given these assumptions, the late omnidirectional impulse response can be written as

$$W_{\text{late}}(\omega) = \sum_{i=1}^{I} D_i(\omega) \tag{4}$$

where $D_i(\omega)$ is the diffuse sound arriving from the direction corresponding to index i.

Note that the assumption about diffuse sound implies that $\mathrm{E}\{|D_i(\omega)|^2\} = \mathrm{E}\{|D_k(\omega)|^2\}$ for all index pairs i and k, where $\mathrm{E}\{\cdot\}$ is expectation and $|\cdot|$ is the magnitude of a complex number. Also, the diffuse sound assumption implies that $\mathrm{E}\{D_i(\omega)D_k(\omega)\} = 0$ for $i \neq k$. Then with (4) it follows that the power spectrum of $D_i(\omega)$ is

$$\mathrm{E}\{|D_i(\omega)|^2\} = \frac{|W_{\text{late}}(\omega)|^2}{I} \tag{5}$$

where $W_{\text{late}}(\omega)$ is the spectrum of $w_{\text{late}}(n)$.

The late left and right BRIRs are

$$B_{L,\text{late}}(\omega) = \sum_{i=1}^{I} L_i(\omega)D_i(\omega)$$
$$B_{R,\text{late}}(\omega) = \sum_{i=1}^{I} R_i(\omega)D_i(\omega). \tag{6}$$

From (5) and (6) it follows that the BRIR power spectrum is

$$P_L(\omega) = \frac{|W_{\text{late}}(\omega)|^2}{I} \sum_{i=1}^{I} |L_i(\omega)|^2$$
$$P_R(\omega) = \frac{|W_{\text{late}}(\omega)|^2}{I} \sum_{i=1}^{I} |R_i(\omega)|^2 . \tag{7}$$

The magnitude of the coherence between the left and right BRIRs is

$$\Phi(\omega) = \frac{\left|\left\langle B_{L,\text{late}}(\omega)B^{\star}_{R,\text{late}}(\omega)\right\rangle\right|}{\sqrt{\left\langle |B_{L,\text{late}}(\omega)|^2\right\rangle \left\langle |B_{R,\text{late}}(\omega)|^2\right\rangle}} \tag{8}$$

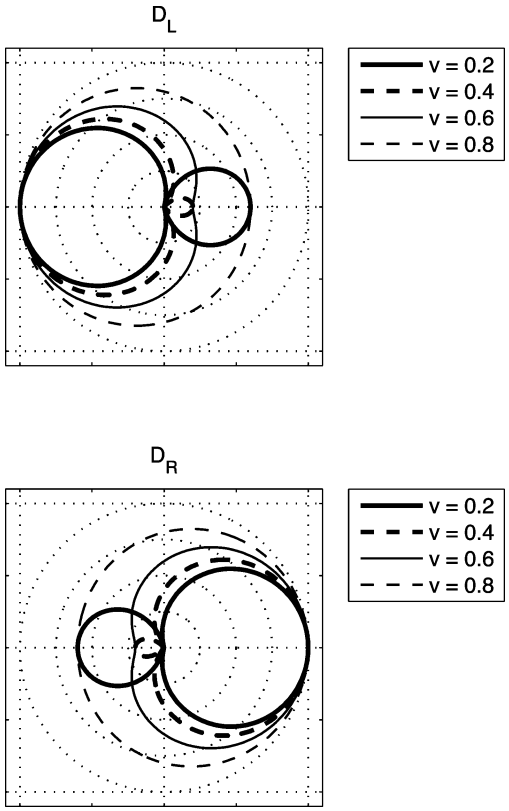

Fig. 3. Directional responses D_L and D_R for various B-format decoding constants v (normalized, on a linear scale).

where * denotes the complex conjugate of a complex number and $\langle x \rangle$ denotes the expected value of x. This is equivalent to

$$\Phi(\omega) = \frac{\left|\sum_{i=1}^{I} L_i(\omega) R_i^{\star}(\omega)\right|}{\sqrt{\sum_{i=1}^{I} |L_i(\omega)|^2 \sum_{i=1}^{I} |R_i(\omega)|^2}}. \tag{9}$$

In the following, late BRIRs are generated in a way that their left and right power spectrum is equal to (7) and their coherence is equal to (9).

Note that (7) and (9) imply a set of HRTFs for directions evenly spaced on a sphere around the head of the listener. If such a set is not available, it is necessary to weight each HRTF by the area on a unit sphere that represents all directions which would be quantized to the HRTF in question.

2) Computation of the Modeled BRIR: From the B-format late room impulse response signals, denoted $W_{\text{late}}(\omega)$, $X_{\text{late}}(\omega)$, $Y_{\text{late}}(\omega)$, and $Z_{\text{late}}(\omega)$, the left and right channels of the late BRIR, $B_{L,\text{late}}$, and $B_{R,\text{late}}$ are generated

$$\begin{aligned}\hat{B}_{L,\text{late}}(\omega) &= H_L(\omega)\left(v(\omega) W_{\text{late}}(\omega) + \frac{1 - v(\omega)}{\sqrt{2}} Y_{\text{late}}(\omega)\right)\\ \hat{B}_{R,\text{late}}(\omega) &= H_R(\omega)\left(v(\omega) W_{\text{late}}(\omega) - \frac{1 - v(\omega)}{\sqrt{2}} Y_{\text{late}}(\omega)\right)\end{aligned} \tag{10}$$

where $v(\omega)$ is a frequency dependent constant and $H_L(\omega)$ and $H_R(\omega)$ are real-valued filters that model the modification of the power spectrum imposed by the HRTF set. Note that the factor $1/\sqrt{2}$ is there to compensate the additional $\sqrt{2}$ gain in the B-format dipole gains.

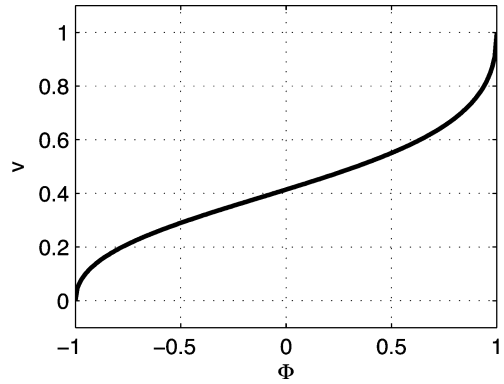

Fig. 4. B-format decoding constant v as a function of the coherence Φ.

First the constant $v(\omega)$ is determined. The directional responses of the two signals (11) are

$$\begin{aligned}D_L(\omega, \phi) &= H_L(\omega)\left(v(\omega) + (1 - v(\omega))\cos\phi\right)\\ D_R(\omega, \phi) &= H_R(\omega)\left(v(\omega) - (1 - v(\omega))\cos\phi\right).\end{aligned} \tag{11}$$

Fig. 3 shows a few example normalized directional responses for different B-format decoding constants v. As can be seen from Fig. 3, the directional response of the linear B-format decoding has its global maximum on the left side, i.e., corresponds to a microphone pointing to the left. The decoding for the right channel is the same as for the left channel, but mirrored.

From these directional responses the magnitude of the coherence for the generated BRIRs (11) can be determined, assuming diffuse sound[1]

$$\Phi(\omega) = \frac{\left|\int_{-\pi}^{\pi} D_L(\omega, \phi) D_R^{\star}(\omega, \phi) d\phi\right|}{\sqrt{\int_{-\pi}^{\pi} |D_L(\omega, \phi)|^2 d\phi \int_{-\pi}^{\pi} |D_R(\omega, \phi)|^2 d\phi}}. \tag{12}$$

By substituting (12) into (12), it can be shown that

$$\Phi(\omega) = \frac{v^2(\omega) + 2v(\omega) - 1}{3v^2(\omega) - 2v(\omega) + 1}. \tag{13}$$

Equation (13) is equivalent to the quadratic equation

$$(3\Phi(\omega) - 1)v^2(\omega) - 2(\Phi(\omega) + 1)v(\omega) + \Phi(\omega) + 1 = 0. \tag{14}$$

The solution of (14) which fulfills $v(\omega) \in [0,\ 1]$ is

$$v(\omega) = \frac{\Phi(\omega) + 1}{3\Phi(\omega) - 1} - \frac{\sqrt{4(\Phi(\omega) + 1)^2 - 4(3\Phi(\omega) - 1)(\Phi(\omega) + 1)}}{6\Phi(\omega) - 2}.$$

Fig. 4 shows the B-format decoding constant $v(\omega)$ as a function of the coherence $\Phi(\omega)$.

In addition to determining $v(\omega)$ in (11), the filters $H_L(\omega)$ and $H_R(\omega)$ need to be determined. From the condition that the

[1]For simplicity, a horizontal diffuse sound model is considered here. A 3-D diffuse sound model can be considered by integrating 3-D directional responses.

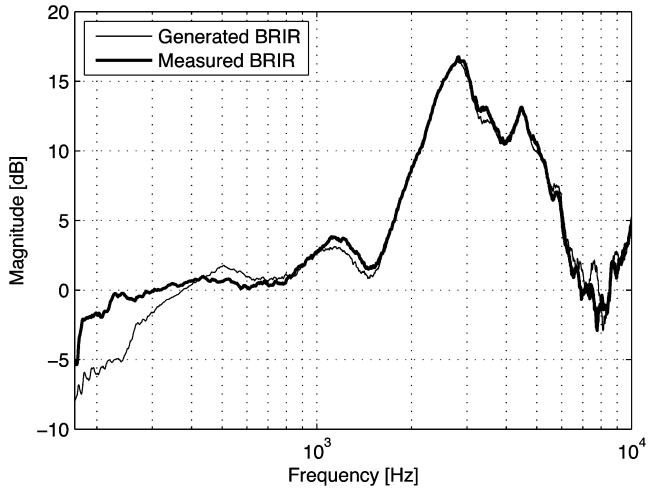

Fig. 5. Spectra of a measured and a generated left BRIR.

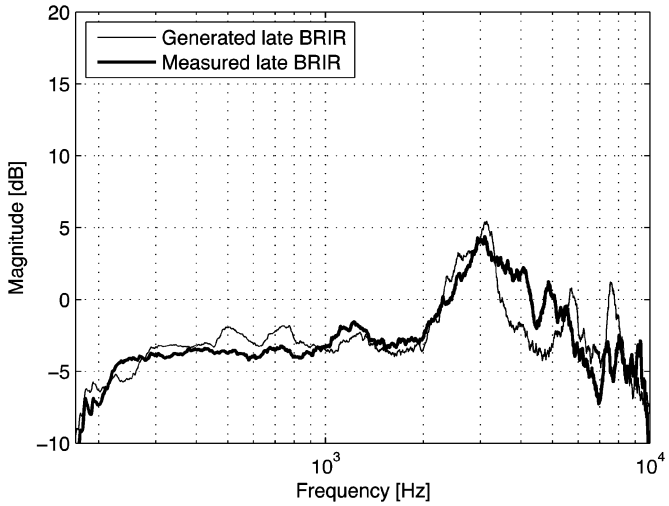

Fig. 6. Spectra of the reflections part of the same measured and generated left BRIR as in Fig. 5.

power spectra of (11) need to be equal to the desired power spectra (7), it follows that

$$H_L(\omega) = \frac{\sqrt{P_L(\omega)}}{\left|v(\omega)W_{\text{late}}(\omega) + \frac{1}{\sqrt{2}}(1 - v(\omega))Y_{\text{late}}(\omega)\right|}$$

$$H_R(\omega) = \frac{\sqrt{P_R(\omega)}}{\left|v(\omega)W_{\text{late}}(\omega) - \frac{1}{\sqrt{2}}(1 - v(\omega))Y_{\text{late}}(\omega)\right|}.$$

III. Signal-Level Evaluation

The proposed method was implemented in Matlab and was applied to a B-format RIR measured in a lecture hall at our university. We also measured in the same room and with the same loudspeaker setup a set of BRIRs (see Appendix A), from which we could also obtain a set of HRTFs for the same source directions by isolating the direct sound. Therefore, we could compare a measured BRIR with a BRIR generated from a B-format RIR and an HRTF set measured in the same room with the same loudspeaker setup and with the same microphone position. In

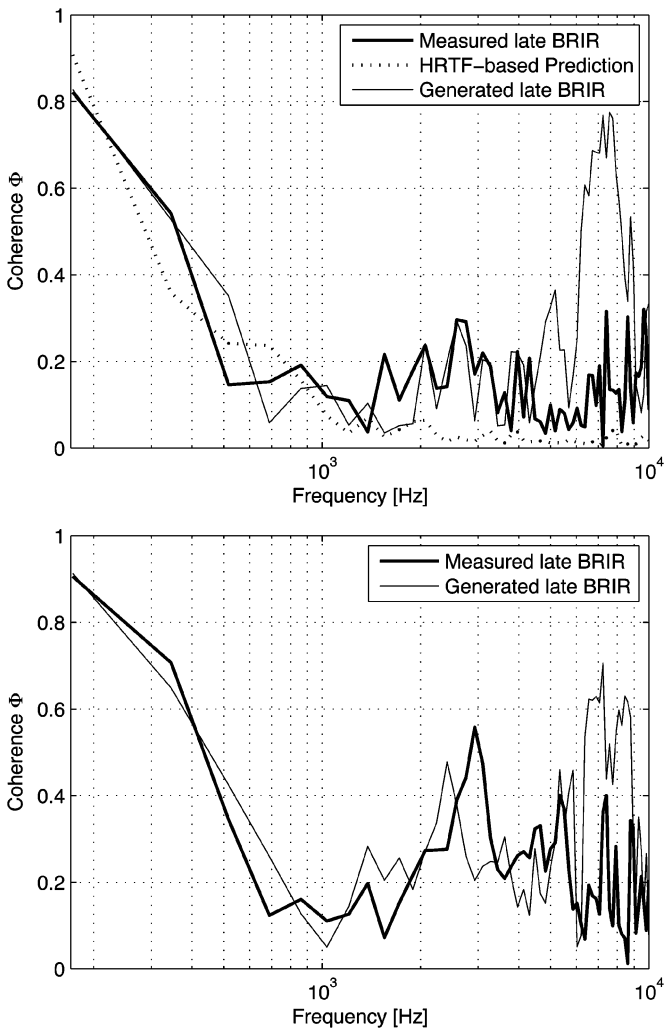

Fig. 7. Interaural coherence of the reflections part of the same measured and generated BRIRs as in Fig. 5 Top panel: interaural coherence for the diffuse reflections part only, not taking into account the first 150 ms of the BRIR. The dotted line shows the HRTF-based prediction of the coherence of diffuse sound recorded with the artificial head. Bottom panel: interaural coherence for the entire reflections part, starting 6 ms after the direct sound. Note that because the coherence analysis for this figure was done in a short-time Fourier domain, the frequency resolution of the coherence is smaller than the frequency resolution of the power spectra shown in Fig. 5.

the following, all data shown is for BRIRs with azimuth 0° and elevation 0° (i.e., the sound source is directly in front of the listener).

The power spectra and coherence of the measured BRIR and the generated BRIR are shown in Figs. 5–7, respectively. Fig. 5 compares the spectra of the entire BRIRs. The good match between the two BRIRs above 300 Hz is due to the fact that the direct sound, which contains most of the energy, is similar for the measured and for the generated BRIR. For simplicity, only the left channel is shown. However, one can observe a deviation of about 5 dB around 200 Hz. It may be that the separation of the direct sound from the rest of the BRIR is not well adapted to low frequencies, where artifacts may occur because of the abrupt transition from the HRTF-based direct sound processing to linear decoding of the late tail.

However, to evaluate the performance of the linear decoding of the reflections part of the B-format RIR, the spectra of the

reflections parts of the BRIRs must be compared, as in Fig. 6. The spectrum of the reflections part generated with the linear B-format decoding matches the measured BRIR up to 3 kHz, but above this frequency deviations of 5 dB and more occur. One possible source of these errors is that at high frequencies the directional responses of the Soundfield microphone used for the B-format RIR measurements start to deviate from the ideal responses [14].

The coherence of the measured and the generated BRIR are shown in Fig. 7. The top panel shows the interaural coherence for the late reverb tail, from 150 ms after the direct sound, as well as the HRTF-based prediction of the interaural coherence for diffuse sound. In this case, the assumption of a perfectly diffuse sound in the late BRIR is approximately verified and all three curves match well up to 4 kHz, giving evidence that the proposed method for interaural coherence matching works as intended.

The bottom panel of Fig. 7 shows the interaural coherence for the entire reflections part of the measured BRIR and the generated BRIR. Even though the assumption of a perfectly diffuse sound is not verified for single early reflections, the linear decoding technique based on this assumption produces a reverberation with a qualitatively similar interaural coherence.

It can be noticed that above 4 kHz and especially above 6 kHz, the coherence of the generated BRIR is generally too high. Again, imperfections of the Soundfield microphone may be the source of these errors.

In order to compare the proposed method with a more conventional way of generating BRIRs from a B-format RIR, a simple B-format RIR decoding with multiple directional RIRs obtained by simulating cardioid directional microphones was implemented. The directional RIRs were convolved with HRTFs for the corresponding directions in order to obtain a simulated BRIR. In particular, simulated BRIRs with three and four cardioid responses with elevation 0° and azimuths 0°, 120°, and 240° and 0°, 90°, 180°, and 270°, respectively, were calculated, where 0° corresponds to the azimuth direction of the direct sound. Informal listening showed that higher numbers of cardioids lead to less natural sounding BRIRs, therefore only the aforementioned 3- and 4-cardioid BRIRs were used for further investigations.

Fig. 8 shows the coherence for the late tail of the 3- and 4-cardioid BRIRs and for the reference BRIR (all starting from 150 ms after the direct sound, for fair comparison with the top panel in Fig. 7). The coherence of the cardioid BRIRs is generally too high, and does not follow the curve of the coherence of the measured BRIR above 1 kHz.

Fig. 9 shows the directional responses as described in (12) used for the B-format decoding generating the late BRIRs. For simplicity only the responses for the left channel are shown.

The measured and modeled BRIRs are shown in Fig. 10. As can be seen in the zoomed portion of the waveform, the early reflections are reproduced well, despite the fact that only the linear B-format decoding was applied and no HRTF for the specific direction of the early reflection was used. The good result can be explained because the linear decoding uses directional responses with maxima to the left for the left channel and to the right for the right channel, as can be seen in Fig. 3. This is

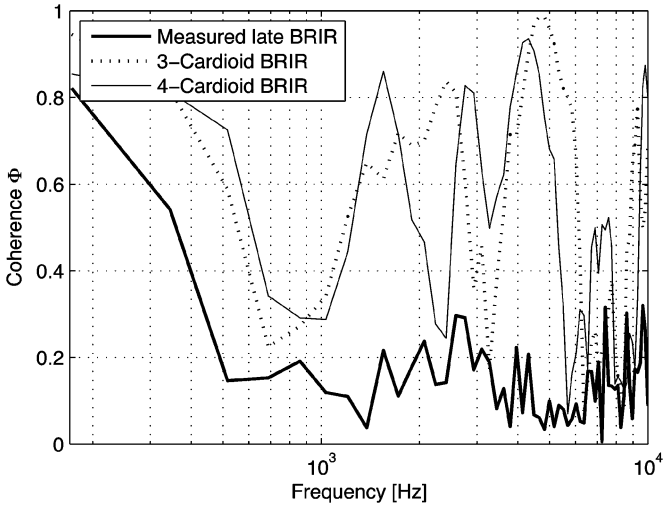

Fig. 8. Coherence of the measured BRIR and two different BRIRs based on cardioid response decodings of the B-format BRIR. In order to be able to assume a diffuse sound, the first 150 ms of the impulse response are not taken into account. The coherence of the cardioid BRIRs is generally too high and does not follow well the coherence of the measured BRIR.

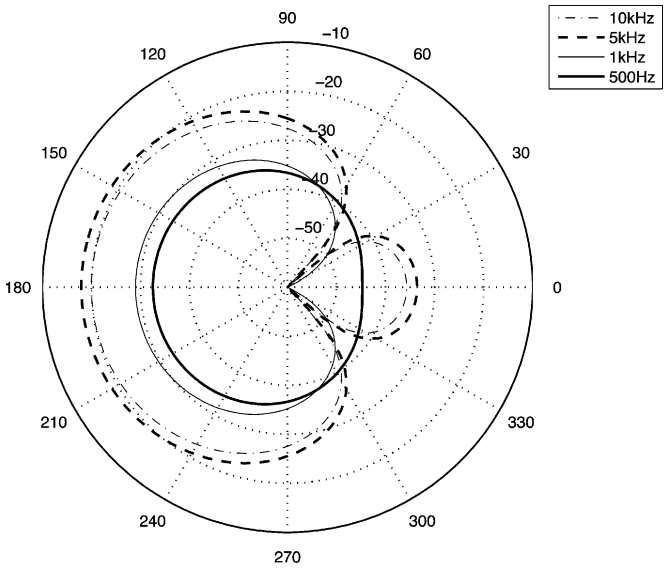

Fig. 9. Directional responses of the linear decoding of the late B-format RIR for the left channel, for different frequencies, in decibels.

similar to the directional responses of the ears, which have their maxima between 60° and 90° to the left and to the right of the median plane [15], [16].

IV. Subjective Evaluation

A subjective test was conducted to show that the proposed method produces high-quality BRIRs comparable to recorded BRIRs and that the proposed method performs better than a conventional method to obtain BRIRs from B-format RIRs (linear cardioid decoding of the B-format and convolution with HRTFs applied). Informal listening showed that the decoding with three cardioids performed better than the decoding with four cardioids. In order to reduce the number of stimuli, only the decoding with three cardioids was used in the subjective test. We have asked both experienced listeners and naive listeners to take part in our subjective test.

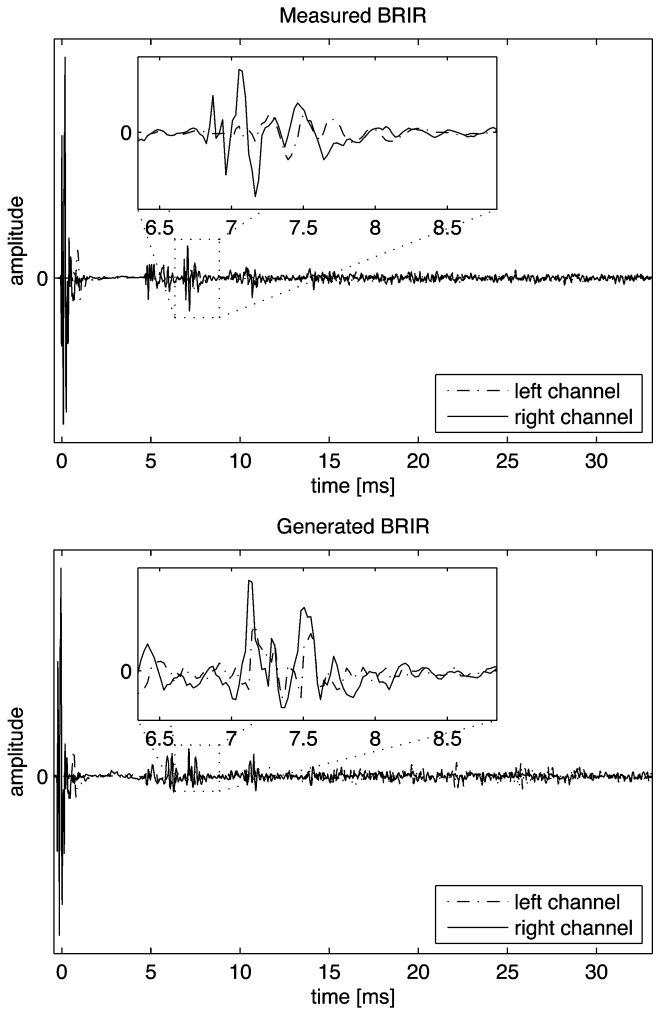

Fig. 10. Waveforms of measured and generated BRIRs with zoom on an early reflection from the left. Top panel: measured BRIR. Bottom panel: generated BRIR. As can be seen from the zoomed early reflection, the linear B-format decoding produces approximately correct level differences.

A. Stimuli

In order to test the different BRIRs in different conditions, we applied the BRIRs to six different speech excerpts and six different dry recordings of musical instruments. The length of the speech excerpts was between 4 and 7 seconds and the length of the music recordings was between 3 and 4 seconds. BRIRs for the azimuth angles -30°, 0°, and 30° and elevation 0° were used. Furthermore, two excerpts of stereo music recordings were presented using the -30° and 30° BRIRs simultaneously. A list of all excerpts is given in Table I.

We chose the sounds with the aim of using natural sounds similar to those that may be used in potential 3-D audio applications. Speech and music seemed reasonable choices in this context.

Each excerpt was convolved with four different "BRIRs" for the assigned direction: the measured BRIR, the generated BRIR, the 3-cardioid BRIR, and a colored (low-pass-filtered) HRTF.

B. Subjects and Test Setup

We asked nine persons to participate in the test. Five of the subjects were experienced listeners (including two of the authors) and four of them were naive listeners. They carried out the test with an automated subjective test software. The subjects used high-quality headphones (Sennheiser HD 600 and Sennheiser HD 25). The listeners were instructed to set the volume level to their preferred level.

TABLE I
LIST OF AUDIO EXCERPTS FOR SUBJECTIVE TEST. BOLD FACE FONT INDICATES THAT AN ITEM WAS USED AS A TRAINING ITEM

Excerpt	BRIR angle(s)
English speech, male	0°
English speech, female	30°
French speech, male	30°
French speech, female	-30°
German speech, male	-30°
German speech, female	0°
Electric guitar	30°
Pop Drum	-30°
Oud	0°
Synthesizer	-30°
Shaker	0°
Electric bass	30°
Irish folk (instrumental)	-30°, 30°
Choir	-30°, 30°

C. Test Method

A MUSHRA [17] type subjective test using a relative grading scale was conducted. The subjects were asked to grade the similarity between the reference (the recorded BRIR) and the other BRIRs relative to three difference aspects: spatial aspects, coloration, overall similarity. A hidden reference was used to test the reliability of the subjects, as well as an "anchor" which consisted of the HRTF, and was expected to obtain marks close to "very different."

Fig. 11 shows the graphical user interface of the subjective test software. The subjects were presented with four play buttons and four sliders to judge the stimuli. Furthermore, there was a play button and a frozen slider for the reference. The subjects could switch between the stimuli at any time while the sound instantly faded from one BRIR to the other.

The test software showed written instructions on the computer screen before the test started. The test contained the 14 excerpts listed in Table I, three of which were used as training items (one speech excerpt, one instrument excerpt, and one stereo music excerpt). The excerpt and method order were randomized differently for each subject.

The duration of the test session varied between the listeners due to the freedom to repeat the stimuli as often as requested. Typically the test duration was between 30 min and 1 h.

D. Results

The results averaged over all subjects and 95% confidence intervals are shown in Fig. 12 (single-instrument music), Fig. 13 (speech), and Fig. 14 (stereo music). As can be seen from these

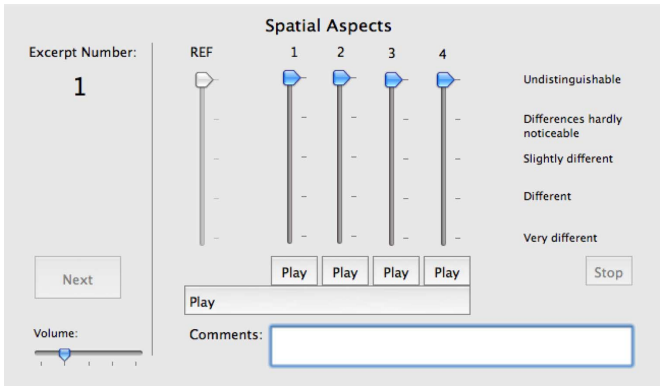

Fig. 11. Graphical user interface of the subjective test software. The frozen slider to the right corresponds to the reference while the four sliders to the left correspond to the other methods (including the hidden reference).

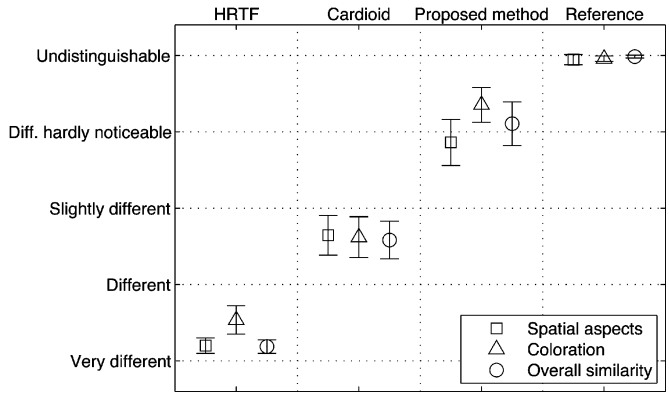

Fig. 12. Average results for all subjects for the single instrument music stimuli, showing 95% confidence intervals.

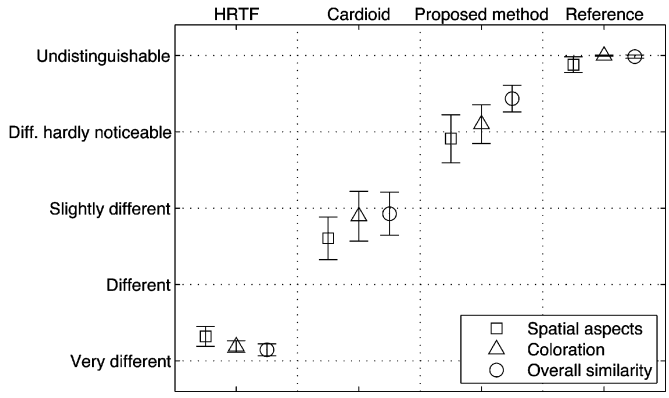

Fig. 13. Average results for all subjects for the speech stimuli, showing 95% confidence intervals.

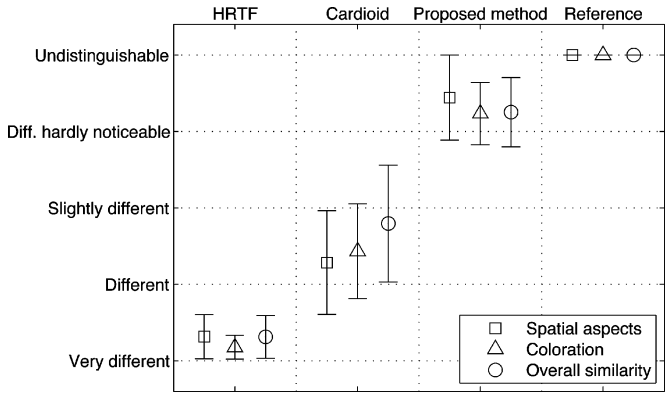

Fig. 14. Average results for all subjects for the stereo music stimuli, showing 95% confidence intervals.

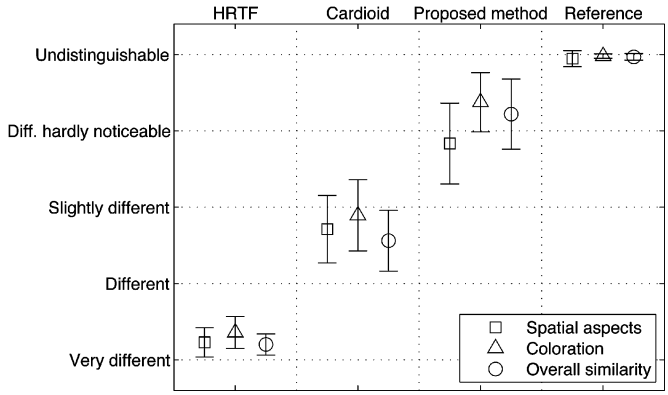

Fig. 15. Average results for all subjects for the pop drum sample and the shaker sample, showing 95% confidence intervals.

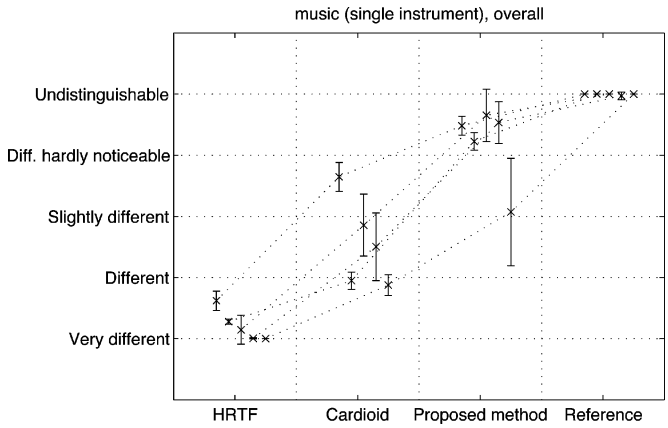

Fig. 16. Individual results for the experienced listeners for the overall similarity aspect of the single instrument samples, showing 95% confidence intervals.

graphs, the proposed method produces BRIRs that are significantly more similar to the reference BRIR than the cardioid-based BRIRs in all cases.

The average rating for the overall similarity of the proposed method with the reference was in all of the cases between "indistinguishable" and "differences hardly noticeable." We conclude that for the average listener our method produces BRIRs that are hard to distinguish from measured BRIRs.

The samples that were used for the listening test not always covered the whole frequency range. In particular, the speech samples had most of their energy below 2 kHz. Two of the musical instrument samples had spectra extending to 10 kHz (and above): the pop drum sample and the shaker sample. Because there was a strong deviation in the coherence above 4 kHz (see Fig. 7), special attention was paid to these two samples. The averaged results for these two samples only are shown in Fig. 15. For the proposed method, the results did not deviate significantly from the averaged results for all the musical instrument samples shown in Fig. 12. It may be concluded that the observed deviation of the coherence above 4 kHz does not significantly influence the perception of the wide-band sounds convolved with BRIRs generated with the proposed method.

When comparing the results of the individual listeners, we observed that the main difference between the different listeners

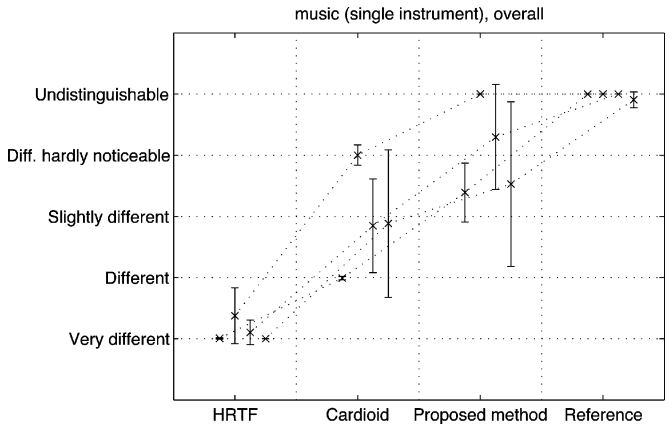

Fig. 17. Individual results for the naive listeners for the overall similarity aspect of the single instrument samples, showing 95% confidence intervals.

(a) (b)

Fig. 18. (a) Loudspeaker setup with Soundfield ST350 at the microphone position. (b) Loudspeaker setup with KEMAR artificial head and torso at the microphone position.

was in their overall sensitivity to deviations from the reference BRIR. Some listeners judged the proposed method as almost undistinguishable and the cardioid based method slightly different, while others found the proposed method to be slightly different and the cardioid-based method different to very different. The main difference between the experienced listeners and the naive listeners was that the naive listeners' results tended to have bigger confidence intervals. The means were similar for both groups of listeners. The individual results for the overall similarity aspect of the single instrument samples are shown in Fig. 16 (for the experienced listeners) and in Fig. 17 (for the naive listeners).

V. Conclusion

A technique was proposed to process B-format room impulse responses (RIRs) and head-related transfer functions (HRTFs) to obtain a set of binaural room impulse responses (BRIRs), individualized to the same head and torso as the used HRTFs. This enables conversion of different HRTF sets to BRIR sets for different rooms with only a need for measuring each room with a B-format microphone. The synthesis of the BRIRs is done differently for direct sound and diffuse sound. The direct sound is extracted from a B-format RIR and its direction of arrival is estimated. It is then filtered with the HRTF corresponding to its direction of arrival to generate the direct sound in the BRIR. The late (diffuse) BRIRs are generated by using a linear combination of the B-format signals, chosen at each frequency such that the spectral and interaural cues are the same as for the true BRIRs.

The BRIRs generated with the proposed method were compared to measured reference BRIRs. The comparison has shown that with respect to the spectra and the frequency-dependent interaural coherence, the BRIRs generated with the proposed method are very close to the reference BRIR up to 3 kHz. Also the waveforms of the early reflections are relatively similar, which can be explained because the linear decoding method uses directional responses similar to the directional responses of the human ear. Therefore, even though the linear decoding is based on the assumption that the B-format recording contains only perfectly diffuse sound, i.e., a hypothesis which is true for late reflections, but not for the early reflections, it approximates the ILD of the early reflections in the measured BRIR.

There are known limitations of the method proposed in this paper. The coherence of the generated BRIR does not match the reference BRIR above approximately 4 kHz. There is some coloration around 200 Hz which may be due to the abrupt transition between the direct sound processing and the linear diffuse sound decoding. A better method of separating the direct sound from the rest of the B-format RIR could help solving this issue.

A subjective test was performed, using as the reference a BRIR measured in a lecture hall. This test showed that the differences in spatial aspects and coloration, and the overall similarity of the generated BRIRs to the reference BRIRs are hardly noticeable. The proposed method also performed significantly better than a conventional method of generating BRIRs from B-format RIRs using cardioid responses extracted from the B-format RIR to which the corresponding HRTFs were applied.

Appendix

Room Impulse Measurements: All room impulse measurements for this research were conducted in a lecture hall at our university (ELA 2) which is 10 m wide, 14 m long, and whose floor ascends in steps towards the back of the room. The loudspeakers and the microphones were placed in the front of the room, where the height is 4 m (see Fig. 18).

For all measurements, the same microphone position and the same loudspeaker setup were used. Seven loudspeakers were placed in a vertical plane pointing towards the microphone position. Their elevation angles and distances relative to the microphone position are shown in Table II.

All D/A and A/D conversions were done with a MOTU 896HD firewire sound interface at 96 kHz. To measure the impulse responses, a logarithmic sweep signal of 2.5-s length, covering the frequency range between 20 Hz and 48 kHz was used.

The B-format room impulses were measured using a Soundfield ST350 microphone and the BRIRs were measured with a KEMAR artificial head with torso. The artificial head was put on a remote-controlled turntable in order to measure BRIRs precisely every 5° in azimuth.

TABLE II
LOUDSPEAKER POSITIONS RELATIVE TO THE MICROPHONE POSITION

Distance	Elevation
1.2 m	60°
1.5 m	30°
1.5 m	15°
1.5 m	0°
1.5 m	−15°
1.5 m	−30°
1.2 m	−60°

The setup was designed such that the first 3 ms of the BRIRs could be used as HRTFs (i.e., no reflections arrive at the microphone position in the first 3 ms after the direct sound). Therefore the measurements yielded at the same time a BRIR set and an HRTF set for 7 elevation angles and 72 azimuth angles.

ACKNOWLEDGMENT

The authors would like to thank everybody from EPFL's Electromagnetics and Acoustics Laboratory (LEMA) for their help and advice for the room impulse response measurements. The authors would also like to thank all the (unpaid) subjects of the listening test for spending their time for this project.

REFERENCES

[1] J. Huopaniemi, "Virtual acoustics and 3D Sound in multimedia signal processing," Ph.D. dissertation, Lab. of Acoust. Audio Signal Process., Helsinki Univ. of Technol., Espoo, Finland, 1999.
[2] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *J. Aud. Eng. Soc.*, vol. 53, no. 12, 2005.
[3] V. Pulkki and J. Merimaa, "Spatial impulse response rendering II: Reproduction of diffuse sound and listening tests," *J. Aud. Eng. Soc.*, vol. 54, no. 1, 2006.
[4] J. Merimaa, "Analysis, synthesis, and perception of spatial sound—binaural localization modeling and multichannel loudspeaker reproduction," Ph.D. dissertation, Helsinki Univ. of Technology, Espoo, Finland, 2006.
[5] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456–466, Jun. 1997.
[6] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, Revised ed. Cambridge, MA: The MIT Press, 1997.
[7] M. A. Gerzon, "Periphony: Width-height sound reproduction," *J. Aud. Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.
[8] K. Farrar, "Soundfield microphone," *Wireless World*, pp. 48–50, Oct. 1979.
[9] W. G. Gardner, "Reverberation algorithms," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brenburg, Eds. Norwell, MA: Kluwer, 1998, ch. 2.
[10] F. Menzer and C. Faller, "Obtaining binaural room impulse responses from B-format impulse responses," in *Preprint 125th Conv. Aud. Eng. Soc.*, Oct. 2008.
[11] J. -M. Jot, "An analysis/synthesis approach to real-time artificial reverberation," in *Proc. ICASSP-92*, 1992, vol. 2, pp. 221–224.
[12] F. Menzer and C. Faller, "Investigations on modeling BRIR tails with filtered and coherence-matched noise," in *Preprint 127th Conv. Aud. Eng. Soc.*, Oct. 2009.
[13] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, Oct. 2001.
[14] C. Faller and M. Kolundzija, "Design and limitations of non-coincidence correction filters for soundfield microphones," in *Preprint 126th Conv. Aud. Eng. Soc.*, May 2009.
[15] E. A. G. Shaw, "Transformation of sound pressure level from the free field to the eardrum in the horizontal plane," *J. Acoust. Soc. Amer.*, vol. 56, no. 6, pp. 1848–1861, 1974.
[16] J. C. Middlebrooks, J. C. Makous, and D. M. Green, "Directional sensitivity of sound-pressure levels in the human ear canal," *J. Acoust. Soc. Amer.*, vol. 86, no. 1, pp. 89–108, 1989.
[17] "Methods for subjective assessment of small impairments in audio systems including multichannel surround systems," ITU, 1997 [Online]. Available: http://www.itu.org

Fritz Menzer received the M.S. (Ing.) degree in communication systems engineering and the Ph.D. degree for his thesis "Binaural audio signal processing using interaural coherence matching" from EPFL, Lausanne, Switzerland, in 2004 and 2010, respectively.

His main research interests are the perception of binaural cues, binaural and multichannel reverberation, spatial audio signal processing, sound synthesis, and music applications.

Christof Faller received the M.S. (Ing.) degree in electrical engineering from ETH Zurich, Zurich, Switzerland, in 2000, and the Ph.D. degree from EPFL Lausanne, Switzerland, in 2004, for his work on parametric multichannel audio coding.

From 2000 to 2004, he worked in the Speech and Acoustics Research Department at Bell Labs Lucent and Agere Systems, where he worked on audio coding for satellite radio, MP3 Surround, and MPEG Surround. He is currently Managing Director at Illusonic, a company he founded in 2006, and part-time Research Associate at the Swiss Federal Institute of Technology (EPFL), Lausanne.

Hervé Lissek was born in Strasbourg, France, in 1974. He graduated in fundamental physics from Université Paris XI, Orsay, France, in 1998, and received the Ph.D. degree from Université du Maine, Le Mans, France, in July 2002, with a specialty acoustics.

From 2003 to 2005, he was a Research Assistant at Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, with a specialization in the fields of electroacoustics and active noise control. Since 2006, he has been heading the Acoustic Group of the Laboratoire d'Electromagnétisme et d'Acoustique at EPFL, working on numerous applicative fields of electroacoustics and audio engineering.