



Audio Engineering Society Convention Paper

Presented at the 125th Convention
2008 October 2–5 San Francisco, CA, USA

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Obtaining Binaural Room Impulse Responses from B-Format Impulse Responses

Fritz Menzer¹, and Christof Faller¹

¹*Audiovisual Communications Laboratory, EPFL Lausanne, CH-1015 Lausanne, Switzerland*

Correspondence should be addressed to Fritz Menzer (fritz.menzer@epfl.ch)

ABSTRACT

Given a set of head related transfer functions (HRTFs) and a room impulse response measured with a Soundfield microphone, the proposed technique computes binaural room impulse responses (BRIRs) which are similar to binaural room impulse responses that would be measured if in place of the Soundfield microphone, the dummy head used for the HRTF set was directly recording the BRIRs. The proposed technique enables that from a set of HRTFs corresponding BRIRs for different rooms are obtained without a need for the dummy head or person to be present for measurement.

1. INTRODUCTION

Binaural room impulse responses (BRIRs) are important tools for high-quality 3D audio [1] rendering because they take into account both the properties of the listener (or dummy head) as well as the properties of the room in which the BRIR has been recorded and therefore allow to give the listener the impression of being in this 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 therefore lack any room-related properties.

In this paper we propose a method that allows to compute BRIRs from Soundfield B-Format impulse responses and HRTF sets. This means that recording the listener-specific properties (HRTFs) is now independent from recording the room-specific properties. In particular, this very much simplifies the task of providing individualized BRIRs for a big number of different acoustic environments for many different persons – something which may be relevant if high quality 3D audio for movies or video games should become popular.

Inspired by current models of reverberation [2], we consider the B-Format 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.

2. PROPOSED PROCESSING

2.1. B-Format room impulse responses

A B-Format room impulse response (B-Format RIR) is a room impulse response measured with a Soundfield microphone [3, 4]. Ideally, it corresponds to a measurement of a room impulse response with four coincident microphones. These four room impulse responses are denoted:

- $w(n)$: RIR measured with an omni microphone
- $x(n)$: RIR measured with a dipole microphone pointing forward
- $y(n)$: RIR measured with a dipole microphone pointing to the side
- $z(n)$: RIR measured with a dipole microphone pointing upwards

Note that usually B-Format is defined such that the dipoles have a gain which is $\sqrt{2}$ larger than the omni gain. Panel (a) in Figure 1 shows an excerpt of a B-Format BRIR.

2.2. RIR separation

Since the direct sound and the early reflections are processed in a different way than the diffuse reverberation, it is necessary to separate the Soundfield RIR into these two parts.

To summarize this procedure, the omni response $w(n)$ is separated into a coherent part $w_{\text{coh}}(n)$ and a diffuse part $w_{\text{dif}}(n)$. From $w_{\text{coh}}(n)$ the direct sound and a certain number of early reflections are extracted, while their angle of arrival is estimated from the original B-Format RIR. Given the early reflections and the original B-Format RIR, an approximate late B-Format RIR $w_{\text{late}}(n)$, $x_{\text{late}}(n)$, $y_{\text{late}}(n)$, $z_{\text{late}}(n)$ is calculated.

In order to obtain $w_{\text{coh}}(n)$ and $w_{\text{dif}}(n)$ from the B-Format RIR, the following assumption is made: $w_{\text{coh}}(n)$ is the part of $w(n)$ that can be predicted from $x(n)$, $y(n)$ and $z(n)$.

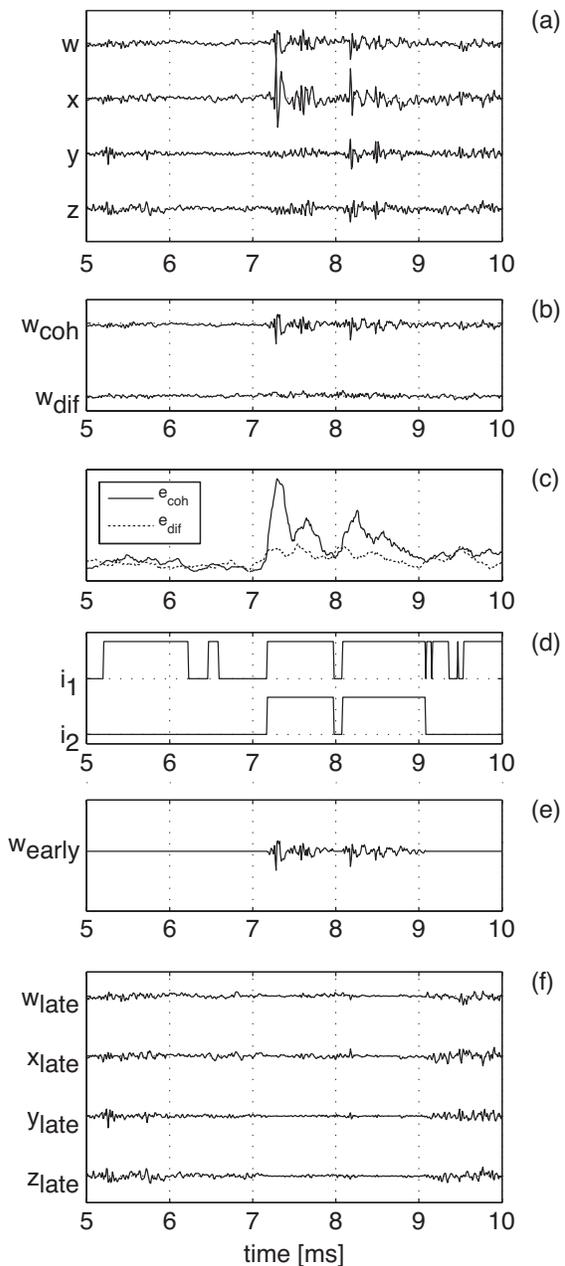


Fig. 1: Soundfield RIR separation. (a) original B-Format RIR; (b) coherent and diffuse omni signals; (c) their envelopes; (d) reflection indicator functions; (e) selected reflections; (f) late B-Format RIR. The excerpt shows the signals 5ms to 10ms after the direct sound and Panels (a), (b), (e) and (f) have the same scale.

$$\begin{bmatrix} x(-M) & \cdots & x(M) & y(-M) & \cdots & y(M) & z(-M) & \cdots & z(M) \\ x(1-M) & \cdots & x(1+M) & y(1-M) & \cdots & y(1+M) & z(1-M) & \cdots & z(1+M) \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ x(N-M) & \cdots & x(N+M) & y(N-M) & \cdots & y(N+M) & z(N-M) & \cdots & z(N+M) \end{bmatrix} \begin{bmatrix} c_{x,-M} \\ \vdots \\ c_{x,M} \\ c_{y,-M} \\ \vdots \\ c_{y,M} \\ c_{z,-M} \\ \vdots \\ c_{z,M} \end{bmatrix} = \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(N) \end{bmatrix} \quad (1)$$

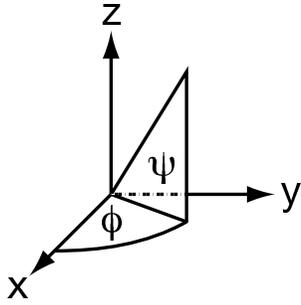


Fig. 2: Cartesian coordinates corresponding to the directions of the dipole microphones recording $x(n)$, $y(n)$ and $z(n)$ and polar coordinates used to specify the directions of early reflections.

In the ideal coherent case, i.e. for a single source in free field, we would have

$$w(n) = c_x x(n) + c_y y(n) + c_z z(n),$$

where c_x , c_y and c_z would be constants that depend only on azimuth ϕ and the elevation ψ in the coordinate system (see Figure 2) defined by the X, Y and Z directions of the Soundfield microphone, i.e.

$$\begin{aligned} c_x &= \cos(\psi) \cos(\phi) \\ c_y &= \cos(\psi) \sin(\phi) \\ c_z &= \sin(\psi). \end{aligned}$$

However, the Soundfield microphone is not perfect and real sound sources are not ideal point sources. Therefore it is better to model $w_{\text{coh}}(n)$ as follows:

$$w_{\text{coh}}(n) = \sum_{i=-M}^M c_{x,i} x(n+i) + c_{y,i} y(n+i) + c_{z,i} z(n+i).$$

For an excerpt of the B-Format RIR of length $N+1$, this can be written in matrix form as in Equation (1) which has the structure $X \cdot C = W$. Letting \hat{X} be the Moore-Penrose pseudo-inverse of X , one obtains the filter $C = [C_x C_y C_z]^T = \hat{X} \cdot W$ (T denotes transpose). This filter is optimal in the least squares sense [5].

Since for a real room impulse response, we have reflections from many directions, the prediction of w from x , y and z cannot hold globally, but only locally. Therefore the B-Format RIR is split into windows (128 samples, 50% overlap) and the coefficient matrix C is calculated separately for each frame. For the recordings we have considered, an 11-tap prediction filter per channel (i.e. $M = 5$) is a reasonable choice.

The predicted (coherent) w_{coh} is calculated by applying frame-by-frame the filter C_x to $x(n)$, C_y to $y(n)$ and C_z to $z(n)$. During the following overlap-add operation, a Hann window is applied to avoid discontinuities.

Because the separation is not perfect, the time of arrival t_1 of the first reflection is estimated and we let $w_{\text{coh}}(0 \dots t_1) = w(0 \dots t_1)$.

An approximation of the diffuse room impulse response, w_{dif} is calculated as

$$w_{\text{dif}}(n) = w(n) - w_{\text{coh}}(n).$$

For a numerical example of $w_{\text{coh}}(n)$ and $w_{\text{dif}}(n)$ see Figure 1, Panel (b).

Comparing the temporal envelopes e_{coh} of w_{coh} and e_{dif} of w_{dif} (see Figure 1, Panel (c)), the reflections are separated, by considering an indicator function

$$i_1(n) = \begin{cases} 1 & e_{\text{coh}}(n) > e_{\text{dif}}(n) \\ 0 & \text{otherwise.} \end{cases}$$

Each time interval where $i_1(n) = 1$ is considered to be a single reflection. Since the first interval often contains not only the direct sound, but also one or more early reflections, there is an option to manually split the first interval.

Only the intervals that contain most energy are chosen to be part of the early BRIR. An early RIR is calculated as

$$w_{\text{early}}(n) = w_{\text{coh}}(n) \cdot i_2(n),$$

where $i_2(n)$ is 1 only for the intervals that are chosen to be part of the early BRIR. See Figure 1, Panels (d) and (e).

Since for the modeling of the late BRIR a B-Format RIR is needed, it is approximated in the following way: first the envelope $e_{\text{early}}(n)$ of $w_{\text{early}}(n)$ is calculated, as well as the envelope $e_w(n)$ of $w(n)$. The late B-Format RIR is calculated as follows:

$$w_{\text{late}}(n) = \frac{e_w(n) - e_{\text{early}}(n)}{e_w(n)} w(n)$$

$$x_{\text{late}}(n) = \frac{e_w(n) - e_{\text{early}}(n)}{e_w(n)} x(n)$$

$$y_{\text{late}}(n) = \frac{e_w(n) - e_{\text{early}}(n)}{e_w(n)} y(n)$$

$$z_{\text{late}}(n) = \frac{e_w(n) - e_{\text{early}}(n)}{e_w(n)} z(n).$$

An example of the late B-Format RIR is shown in Panel (f) of Figure 1.

2.3. Modeling the early BRIR

For each reflection and for the direct sound, the direction of arrival is determined by calculating

$$p_x = \sum_{n \in I_r} x(n)w(n)$$

$$p_y = \sum_{n \in I_r} y(n)w(n)$$

$$p_z = \sum_{n \in I_r} z(n)w(n)$$

on the time interval I_r that corresponds to the reflection. This method of direction of arrival estimate is related to the method used in [6]. It is only an approximation because $w(n)$, $x(n)$, $y(n)$ and $z(n)$ contain also diffuse sound, but since only the reflections containing high energy are considered, the coherent part contains considerably more energy than the diffuse part. Furthermore, the inner product will be less affected by the non-coherent diffuse sound than by the coherent reflection.

The angles are calculated as

$$\phi = \arg(p_x + ip_y)$$

$$\psi = \arg\left(\sqrt{p_x^2 + p_y^2} + ip_z\right).$$

Knowing the coherent part of the impulse response for each reflection as well as the angle of arrival, it is easy to calculate the early BRIR. It is sufficient to apply to each reflection the HRTF that corresponds to its angle of arrival.

2.4. 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 computed late BRIR are the same as the power spectra of the true BRIR.
- The normalized cross-correlation coefficient between the left and right computed BRIRs is the same as the normalized cross-correlation coefficient between the true left and right late BRIRs at each frequency.
- At each frequency the temporal envelope of the computed BRIR is the same as for the true BRIR.

Note that the first two items result in that the important perceptual spatial cues interaural level difference and coherence [7] are the same for the synthesized and true BRIRs at each frequency.

In the following we are computing the left and right true BRIR power spectra and cross-correlation coefficient as a function of frequency between the left

and right BRIR. Then, it is shown how to compute BRIRs by linear B-Format decoding from the B-Format room impulse responses such that the power spectra and normalized cross-correlation coefficient are the same as in the true BRIRs. The decay of the BRIR will be the same as the decay of the B-Format room impulse response, since linear B-Format decoding does not change the decay.

2.5. Computation of the true BRIR parameters

The left and right power spectra of the true BRIRs are obtained by averaging the HRTF power spectra for all directions (it is assumed that diffuse sound arrives from all directions with the same average power and that diffuse sound from each direction is orthogonal to diffuse sound from any other direction):

$$\begin{aligned} P_L(\omega) &= \frac{\sum_{i=1}^I |L_i(\omega)|^2}{I} |W_{\text{late}}(\omega)|^2 \\ P_R(\omega) &= \frac{\sum_{i=1}^I |R_i(\omega)|^2}{I} |W_{\text{late}}(\omega)|^2, \end{aligned} \quad (2)$$

where $|\cdot|$ is the magnitude of a complex number and $L_i(\omega)$ and $R_i(\omega)$ are left and right transfer functions in a given HRTF set covering I equally spaced angles in the horizontal plane and $W_{\text{late}}(\omega)$ is the spectrum of $w_{\text{late}}(n)$. The normalized cross-correlation coefficient between the true left and right late BRIRs as a function of frequency is

$$\Phi(\omega) = \frac{\text{Re}\{\sum_{i=1}^I L_i(\omega)R_i^*(\omega)\}}{\sqrt{\sum_{i=1}^I |L_i(\omega)|^2 \sum_{i=1}^I |R_i(\omega)|^2}}, \quad (3)$$

where $\text{Re}\{\cdot\}$ is the real part of a complex number.

2.6. Computation of the modeled BRIR

From the B-Format late room impulse response signals, denoted $W_{\text{late}}(\omega)$, $X_{\text{late}}(\omega)$, $Y_{\text{late}}(\omega)$, $Z_{\text{late}}(\omega)$, the left and right channels of the late BRIR, $B_{L,\text{late}}$ and $B_{R,\text{late}}$ are computed. The directional response of the left channel is pointing towards the left and the directional response of the right channel is pointing towards the right:

$$\begin{aligned} B_{L,\text{late}}(\omega) &= H_L(\omega) \left(v(\omega)W_{\text{late}}(\omega) + \frac{1-v(\omega)}{\sqrt{2}}Y_{\text{late}}(\omega) \right) \\ B_{R,\text{late}}(\omega) &= H_R(\omega) \left(v(\omega)W_{\text{late}}(\omega) - \frac{1-v(\omega)}{\sqrt{2}}Y_{\text{late}}(\omega) \right), \end{aligned} \quad (4)$$

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.

First the constant $v(\omega)$ is determined. The normalized directional responses of the two signals (4) are

$$\begin{aligned} D_L(\omega, \phi) &= H_L(\omega) (v(\omega) + (1-v(\omega)) \cos \phi) \\ D_R(\omega, \phi) &= H_R(\omega) (v(\omega) - (1-v(\omega)) \cos \phi). \end{aligned} \quad (5)$$

Figure 3 shows a few example directional responses for different B-Format decoding constants v .

From these the normalized cross-correlation coefficient for the computed BRIRs (4) can be determined, assuming diffuse sound:

$$\Phi(\omega) = \frac{\int_{-\pi}^{\pi} D_L(\omega, \phi)D_R(\omega, \phi)d\phi}{\sqrt{\int_{-\pi}^{\pi} D_L^2(\omega, \phi)d\phi \int_{-\pi}^{\pi} D_R^2(\omega, \phi)d\phi}}. \quad (6)$$

By substituting (5) into (6) it can be shown that

$$\Phi(\omega) = \frac{v^2(\omega) + 2v(\omega) - 1}{3v^2(\omega) - 2v(\omega) + 1}. \quad (7)$$

Figure 4 shows the normalized cross-correlation coefficient $\Phi(\omega)$ as a function of the B-Format decoding constant $v(\omega)$. Equation (7) is equivalent to the quadratic equation

$$(3\Phi(\omega) - 1)v^2(\omega) - 2(\Phi(\omega) + 1)v(\omega) + \Phi(\omega) + 1 = 0. \quad (8)$$

The solution of (8) which fulfills $v(\omega) \in [0, 1]$ is

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

Figure 5 shows the the B-Format decoding constant $v(\omega)$ as a function of the normalized cross-correlation coefficient $\Phi(\omega)$. Note that Figure 4 describes the same function as the upper part ($\Phi > 0$) of the curve in Figure 5.

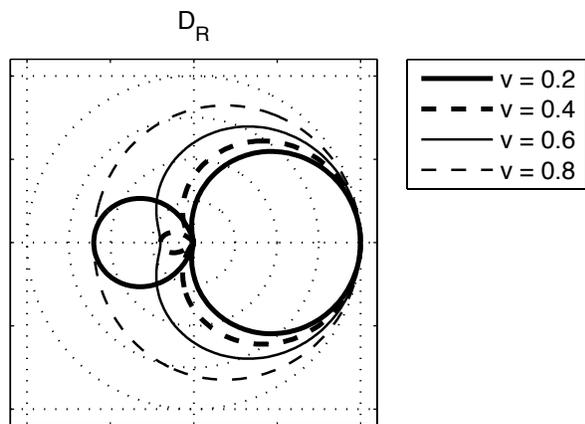
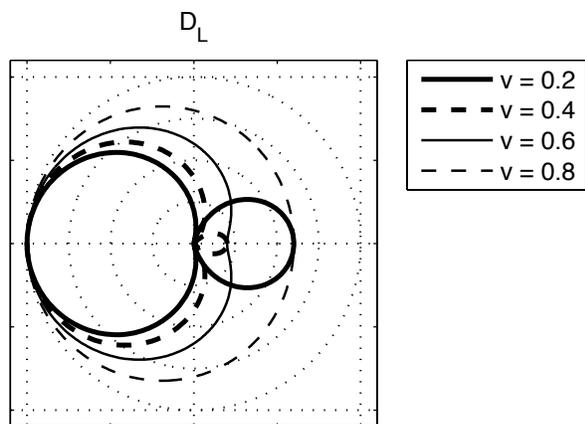


Fig. 3: Directional responses D_L and D_R for various B-Format decoding constants v .

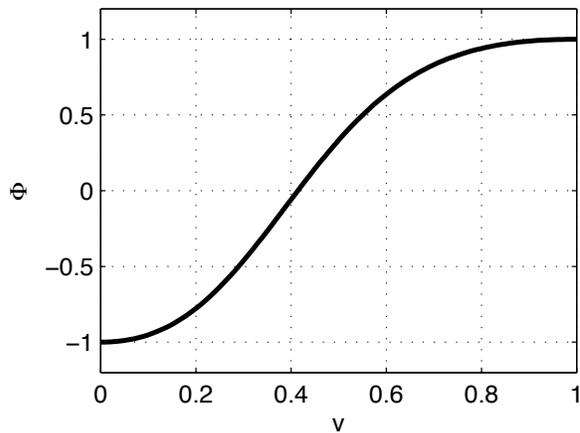


Fig. 4: The normalized cross-correlation coefficient Φ as a function of the B-Format decoding constant v assuming diffuse sound.

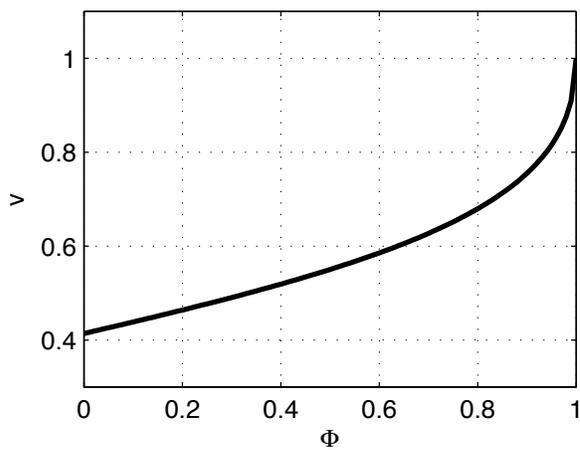


Fig. 5: B-Format decoding constant v as a function of the normalized cross-correlation coefficient Φ .

In addition to determining $v(\omega)$ in (4), the filters $H_L(\omega)$ and $H_R(\omega)$ need to be determined. From the condition that the power spectra of (4) need to be equal to the desired power spectra (2), it follows that

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

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

3. EXPERIMENTS

The proposed technique was implemented in Matlab, using B-Format RIR recordings made in two different rooms at EPFL and using HRTFs from the CIPIC database [8] for rendering the early reflections and for estimating the cross-correlation and power spectra needed for the late BRIR rendering.

The first room in which we recorded B-Format RIRs is a storage room with concrete walls and a characteristic “slap back echo”. This echo was modeled mainly by the late RIR algorithm because the individual reflections have low energy. Informal listening lead us to the conclusion that this is not a problem. Since this room contained several tables and cupboards, it was difficult to interpret the precision of the early reflection direction estimation.



Fig. 6: Setup used for recording B-Format RIRs in the conference room.

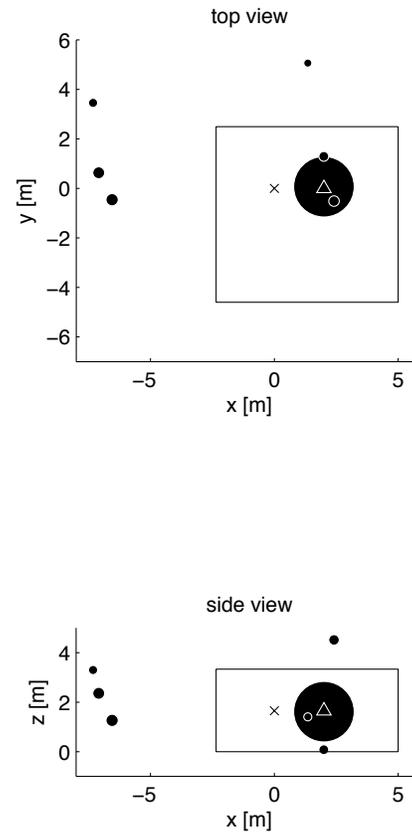


Fig. 7: Positions of the direct sound and 6 reflections extracted from a B-Format RIR of the conference room. The position of the listener is marked by \times and the position of the source by \triangle . Each dot corresponds to a reflection detected by our algorithm and the area of the dot is proportional to the logarithm of the energy contained in the reflection. The biggest dot represents the direct sound. The rectangles show the boundaries of the room. Reflections from two walls (left and top in the top view) and from the ceiling (top in the side view) can be identified.

Therefore we recorded several positions in a conference room that was completely emptied for the occasion (see Figure 6). In this case we could associate the directions and delays of the reflections to several positions that would be predicted by the image source model [9]. See Figure 7 for the image source positions extracted by our algorithm.

So far the proposed algorithm has only been evaluated by informally listening to synthesized BRIRs using HRTFs of the CIPIC database and the described B-Format RIRs. Further evaluations are planned using a listener's individual HRTFs and comparing the resulting BRIRs to reference BRIRs which are measured in the same room and position as the B-Format RIRs.

4. CONCLUSIONS

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 Soundfield microphone. The synthesis of the BRIRs is done differently for early reflections and diffuse sound. The early reflections are extracted from B-Format RIR and their direction of arrival is estimated. Each reflection is then filtered with the HRTFs corresponding to its direction of arrival to generate the corresponding reflection in the BRIRs. 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.

5. REFERENCES

- [1] J. Huopaniemi, *Virtual Acoustics and 3D Sound in Multimedia Signal Processing*, Ph.D. thesis, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland, 1999, Rep. 53.
- [2] W. G. Gardner, "Reverberation algorithms," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds., chapter 2. Kluwer Academic Publishing, Norwell, MA, USA, 1998.
- [3] M. A. Gerzon, "Periphony: Width-Height Sound Reproduction," *J. Aud. Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.
- [4] K. Farrar, "Soundfield microphone," *Wireless World*, pp. 48–50, Oct. 1979.
- [5] R. Penrose, "On best approximate solutions of linear matrix equations," *Proceedings of the Cambridge Philosophical Society*, vol. 52, pp. 17–19, 1956.
- [6] J. Merimaa and V. Pulkki, "Spatial impulse response rendering i: Analysis and synthesis," *J. Aud. Eng. Soc.*, vol. 53, no. 12, 2005.
- [7] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, The MIT Press, Cambridge, Massachusetts, USA, revised edition, 1997.
- [8] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF Database," in *Proc. Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, Mohonk Mountain House, New Palz, NY, Oct. 2001, IEEE.
- [9] J. B. Allen and D. A. Berkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.