Spectral and Spatial Parameter Resolution Requirements for Parametric, Filter-Bank-Based HRTF Processing*

JEROEN BREEBAART,[1] *AES Member,* (jeroen.breebaart@philips.com)
FABIAN NATER,[2] (fnater@vision.ee.ethz.ch)

AND ARMIN KOHLRAUSCH,[1,3] *AES Member* (armin.kohlrausch@philips.com)

[1] *Philips Research, Eindhoven, The Netherlands*
[2] *Information Technology and Electrical Engineering, Eidgenössische Technische Hochschule, Zurich, Switzerland*
[3] *Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands*

The audibility of HRTF information reduction was investigated using a parametric analysis and synthesis approach. Nonindividualized HRTFs were characterized by magnitude and interaural phase properties computed for warped critical bands. The minimum number of parameters was established as a function of the HRTF set under test, the sound-source position, whether overlapping or nonoverlapping parameter bands were used, and whether spectral characteristics were derived from the HRTF magnitude or power spectrum domain. A three-interval, forced-choice procedure was employed to determine the required spectral resolution of the parameters and the minimum requirements for interaural phase reconstruction. The results indicated that, for pink-noise stimuli, the estimation of magnitude and interaural phase spectra per critical band is a sufficient prerequisite for transparent HRTF parameterization. Furthermore the low-frequency HRTF phase characteristics can be efficiently described by a single interaural delay while disregarding the absolute phase response of the individual HRTFs. Also, high-frequency phase characteristics were found to be irrelevant for the HRTFs and the stimuli used for the test. Estimation of parameters in the spectral magnitude domain using overlapping parameter bands resulted in better quality compared to either power-domain parameter estimation or the use of nonoverlapping parameter bands. When HRTFs were reconstructed by interpolation of parameters in the spatial domain, a spatial measurement resolution of about 10° was shown to be sufficient for high-quality binaural processing. Further reductions in spatial resolution predominantly give rise to monaural cues, which are stronger for interpolation in the vertical direction than in the horizontal direction. The results obtained provide clear design criteria for parametric HRTF processing in filter-bank-based applications such as MPEG Surround.

0 INTRODUCTION

The conversion from analog to digital television is progressing steadily. In the last few years, analog television broadcast over the air has been replaced by digital transmission formats. Mobile television networks are also expanding, allowing people to watch television on their mobile phones. Furthermore television based on internet protocols is evolving rapidly. The main drivers for these trends include mobility, convenience, personalization, and the ongoing quest for improved audio/video quality at low bit rates to lower the cost of distribution.

With the trend of digital transmission to mobile devices, new opportunities exist to enhance the user's experience. The mobility aspect often requires headphones as audio reproduction devices. Furthermore television content (movie material in particular) is often accompanied by multichannel audio signals. The reproduction of such multichannel audio content through headphones is particularly challenging if the same spatial percept is desired as if the audio were reproduced on a multichannel loudspeaker setup. The generation of virtual sound sources through headphones has been subject to research for several decades, and dedicated

*Manuscript received 2009 March 9; revised 2009 December 21 and 2010 February 3.

research software environments have been developed (see [1],[2]). The most popular approach for binaural processing is to measure head-related impulse responses (HRIRs) or binaural room impulse responses (BRIRs) and to convolve each source signal with a pair of HRIRs or BRIRs that correspond to the desired sound source position [3]–[5]. To gain efficiency, the convolution is often performed in the frequency domain employing head-related transfer functions (HRTFs) or binaural room transfer functions (BRTFs).

Several challenges can be identified for binaural rendering applications which are related to individual differences in HRTFs [6]–[10], the incorporation of head movements [11], [12], and the necessity for an accurate room simulation for a convincing "out-of-head" percept [11], [13]. In particular for mobile applications, the complexity associated with the binaural algorithm needs to be reduced as the convolution of multichannel signals with HRTFs or BRTFs demands a high processing power, which can be costly and is often impractical given the constraints on the memory size and the battery lifetime. A promising approach to complexity reduction is based on parametric HRTFs. This method is based on the hypothesis that HRTF spectra contain more information than strictly required from a perceptual point of view (see [14], [18]), and the fact that the frequency-dependence of interaural time differences can in many cases be simplified [19], [15], [1]. Moreover it has been shown that the information of low-frequency interaural time differences is more important than that at high frequencies [20]–[22], which suggests that the high-frequency interaural timing information may be omitted in some cases.

In summary it has been shown that all perceptually relevant aspects can be captured using a limited set of HRTF parameters that are closely linked to predominant sound-source localization cues (interaural level and time differences for azimuth, and spectral cues for elevation; see [23]). Previous experiments indicated that if these parameters are extracted as a function of frequency with a resolution of approximately one set per critical band, subjects cannot discriminate the parameterized from the original HRTFs [18]. In the same study the authors also showed that the frequency-dependent phase characteristics of an HRTF pair can be replaced by a frequency-independent interaural time difference (ITD), while discarding the absolute phase characteristics of each individual HRTF.

The parameterization technique can be used for mobile broadcast applications, as it is currently integrated into recent multichannel audio compression schemes such as MPEG Surround [24]–[26], which significantly reduces the required bit rate and the system complexity compared to other conventional techniques [25], [27], [28], without compromising good localization performance [29]. Contrary to conventional methods for HRTF processing that typically increase the computational complexity of an application (such as by adding a binaural rendering stage in the form of fast convolution of full HRTFs or by employing FIR or IIR approximations), the incorporation of HRTF processing in the parameter domain typically reduces the computational complexity by about a factor of 4 or more [27]. This significant increase in efficiency is obtained by 1) the ability to perform HRTF processing directly in the parameter domain, and 2) the reduction of the number of synthesis filter banks from 5 or 6 to only 2.

More recently parametric HRTFs have also been incorporated in the audio compression standards that support interactive positioning and modification of individual sound objects (spatial audio object coding (SAOC); see [30]). In this standard users can freely reposition sound sources in a virtual environment using headphones or in a real environment based on loudspeaker playback. In addition it also allows users to specify their own HRTFs using a standardized HRTF parameter interface. Since the parameters and filter-bank structures of SAOC used are very similar to those of MPEG Surround, similar efficiency gains as for MPEG Surround are expected for this application.

Although the parameterization method, its performance, and its incorporation in standards have been described in the literature, several aspects of this scheme have not been covered yet. This study aims at providing more detailed insight in the relation between several design aspects of the parameterization scheme and the associated discriminability between original and parameterized HRTFs. Specifically it addresses the following topics:

- The effect of spectral parameter-band overlap on the detectability of changes induced by the HRTF parameterization scheme. Understanding the consequences of spectral overlap is important to consider when designing a proper filter bank for a binaural application employing the parameterization scheme.
- The effect of the parameter estimation domain, that is, whether level parameters are estimated in the spectral magnitude or the power domain may have consequences for the perceived quality of binaurally rendered signals.
- The perceptual consequences of changes in the number of parameters, expressed in a quantitative tradeoff between audibility and number of parameters. Such information is important for the design tradeoffs in terms of memory requirements and filter-bank complexity.
- The maximum required frequency for ITD synthesis. If ITD synthesis is only perceptually relevant in the low-frequency range, while for high frequencies only magnitude spectrum attributes are relevant, filter banks may be designed that operate in the complex-valued domain for low frequencies, and are real-valued for the remaining bandwidth, resulting in a substantial reduction in computational and memory requirements (the so-called low-power processing mode of MPEG Surround; see [28]).
- The effect of changes in the spatial resolution of the HRTF parameter database used. Especially for SAOC applications, sound sources can be positioned freely in a virtual environment. An indication of the minimum spatial resolution required for high-quality binaural rendering would be worthwhile when building a consumer-domain application that is typically subject to stringent cost minimization.

The paper is organized as follows. In Section 1 the method of HRTF analysis and synthesis using parameters is explained. In Section 2 the minimum number of parameters that produce no audible differences between the original and the parameterized HRTFs is investigated in listening tests addressing the research questions mentioned. Section 3 discusses the results, and conclusions are given in Section 4.

1 HRTF PARAMETERIZATION

1.1 HRTF Analysis

The goal of HRTF parameterization is to allow analysis and resynthesis of the perceptually relevant attributes of any HRTF pair. Given the evidence that fine-structure details in HRTF spectra within critical bands are in most cases irrelevant, a straightforward method is to describe parameters as average values for each critical band, an approach that was also used in [18]. This process results in the following parameters to be extracted for a limited set of parameter bands b:

- a level parameter vector for the left ear $\boldsymbol{\sigma}_\mathrm{l}$
- a level parameter vector for the right ear $\boldsymbol{\sigma}_\mathrm{r}$
- an interaural time or phase parameter vector $\boldsymbol{\phi}$.

The level parameters $\boldsymbol{\sigma}_\mathrm{l}$ and $\boldsymbol{\sigma}_\mathrm{r}$ define the spectral cues for elevation. The azimuthal localization cues (interaural level and time differences) are captured by the elementwise ratio of the level parameters and $\boldsymbol{\phi}$, respectively. A fourth parameter that could optionally be employed is the interaural coherence $\boldsymbol{\rho}$. This parameter describes potential spatial diffuseness properties introduced by HRTF convolution, and has also been shown to influence the perceived distance [31]. For the HRTF sets under test in the current paper, however, the coherence is very close to $+1$, and hence the effect of this parameter is not investigated here.

The HRTF analysis and synthesis processes can be implemented using a variety of time–frequency transforms. In the following examples HRTFs are assumed to be represented by discrete-sampled complex-valued frequency spectra $\boldsymbol{H}_\mathrm{l}$ and $\boldsymbol{H}_\mathrm{r}$ of length K, for the left and right ears, respectively, and the parameter bands are defined by a parameter-band definition matrix $\boldsymbol{M}$ of size (K, B), which determines the contribution of each spectrum component k to each parameter band b. This matrix typically describes a set of adjacent parameter bands that may be partially overlapping. More details on this matrix are provided in Section 1.2. Given a parameter-band matrix $\boldsymbol{M}$, the level parameters of the power spectrum of HRTF $\boldsymbol{H}_x$ can be estimated:

$$\boldsymbol{\sigma}^2{}_{xx} = \boldsymbol{M}^{+}\boldsymbol{H}_{xx} \tag{1}$$

with $\boldsymbol{H}_{xy}$ the elementwise cross product of $\boldsymbol{H}_x$ and the complex conjugate of $\boldsymbol{H}_y$,

$$H_{xy}(k) = H_x(k)H_y^*(k) \tag{2}$$

and $\boldsymbol{M}^+$ the pseudo inverse of $\boldsymbol{M}$,

$$\boldsymbol{M}^+ = \left(\boldsymbol{M}^\mathrm{T}\boldsymbol{M}\right)^{-1}\boldsymbol{M}^\mathrm{T}. \tag{3}$$

In this example the level parameters $\boldsymbol{\sigma}^2{}_{xx}$ are expressed in the power domain. In a similar way the level parameters can also be extracted from the magnitude spectrum of $\boldsymbol{H}_x$, resulting in magnitude-domain parameters $\boldsymbol{\sigma}_x$.

The interaural phase parameters $\boldsymbol{\phi}$ can be extracted according to

$$\boldsymbol{\phi} = \boldsymbol{M}^{+}\angle_\mathrm{u}(\boldsymbol{H}_\mathrm{lr}) \tag{4}$$

with $\angle_\mathrm{u}(\boldsymbol{H})$ the unwrapped phase angles of the elements of $(\boldsymbol{H})$.

1.2 Parameter-Band Matrix *M*

The parameter-band matrix $\boldsymbol{M}$ defines the (weighted) mapping between spectral components of an HRTF and the parameter bands. In the current frameworks, the parameter-band matrix is defined by two attributes. The first is the parameter γ, which defines the bandwidth of the parameter bands. The parameter bands are defined on a warped ERB (equivalent rectangular bandwidth, see [32]) scale, given by

$$f(b) = QL\left[\exp\left(\frac{b\gamma}{Q}\right) - 1\right] \tag{5}$$

with b the parameter-band number, γ the warp factor, Q the filter Q factor, L the minimum bandwidth, and $f(b)$ the center frequency of parameter band b. The bandwidth $w(f)$ at frequency f is given by

$$w(f) = \gamma\left(L + \frac{f}{Q}\right) \tag{6}$$

and hence the bandwidth as a function of the parameter-band number b is given by

$$w(b) = \gamma L \exp\left(\frac{b\gamma}{Q}\right). \tag{7}$$

According to [32], the bandwidth of the auditory filters can be described with $\gamma = 1$, $Q = 9.265$, and $L = 24.7$. In the experiments reported in this paper the effect of changing the warp factor γ will be investigated, which is equivalent to changing the HRTF parameter bandwidth. Larger values of γ will result in wider (and hence fewer) parameter bands. Said differently, if various parameterization methods are compared, larger values of γ indicate a more efficient parameterization. As an example, Table 1 provides the number of parameter bands B needed to describe the frequency range from 0 to 20 kHz as a function of the warp factor γ, assuming the highest filter is centered at around 18 kHz.

The second attribute that determines the parameter-band matrix M is the spectral shape of the parameter bands. Two different parameter-band shapes will be used in this study. The first comprises rectangular, spectrally nonoverlapping parameter bands. The parameter bands are defined as rectangularly shaped, adjacent bands with

Table 1. Number of parameter bands B as a function of warp factor γ.

γ	0.8	1.0	1.2	1.5	2.0	3.0	4.0
B	49	40	33	28	20	14	10

a bandwidth equal to $w(b)$ for parameter band b. These parameter bands reflect analysis and synthesis methods with (almost) infinitely high frequency separation. The second spectral shape is based on triangular, 50% overlapping parameter bands. This case simulates analysis and synthesis methods using spectrally (partially) overlapping bands. The two spectral parameter shapes are visualized in Fig. 1. The elements of $\boldsymbol{M}$ are shown for three parameter bands $b-1$, b, $b+1$ using nonoverlapping bands [Fig. 1(a)] and overlapping bands [Fig. 1(b)].

1.3 HRTF Synthesis

Reconstruction of the power spectrum $\boldsymbol{H}'_{xx}$ or interaural phase spectrum now follows from multiplication of the representative parameter vector with the matrix $\boldsymbol{M}$,

$$\boldsymbol{H}'_{xx} = \boldsymbol{M}\boldsymbol{\sigma}^2_{xx}. \tag{8}$$

The following process was used to reconstruct HRTF impulse responses from HRTF parameter sets:

- A desired impulse response length N was chosen given a sample rate f_s.
- A Dirac pulse was created at the temporal center of the N samples. Subsequently the set of N samples was transformed to the frequency domain using a discrete Fourier transform of length N.
- The reconstructed magnitude or power spectrum was superimposed on the magnitude or power spectrum of the transformed Dirac pulse for the left- and right-ear HRTFs independently.
- The reconstructed interaural phase spectrum was applied by distributing the interaural phase values symmetrically across the left- and right-ear responses (but with different signs). Hence only relative phase information was preserved; absolute phase information was discarded. To result in real-valued impulse responses, Hermitian symmetry was ensured during the spectrum reconstruction process.
- The resulting complex spectra were transformed to the time domain using an inverse discrete Fourier transform. Finally a symmetric Hanning window of length N was applied to smoothen the edges of the impulse response.

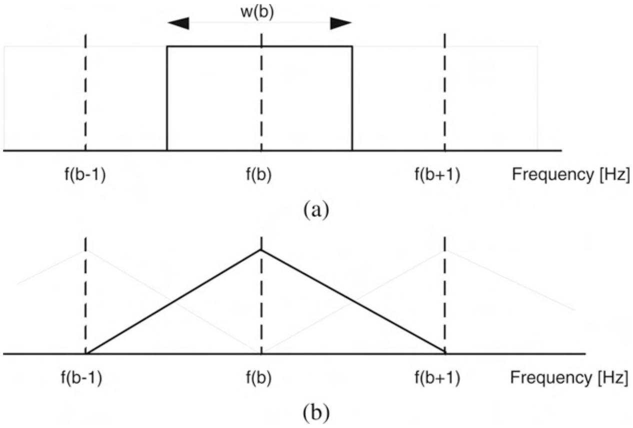


Fig. 1. Visualization of matrix $\boldsymbol{M}$. (a) Based on nonoverlapping parameter bands. (b) Based on overlapping parameter bands.

Exemplary magnitude spectra using a warp factor $\gamma = 2$ resulting in 20 spectral coefficients are shown in Fig. 2. Fig. 2(a), (b) represents overlapping parameter bands, Fig. 2(c), (d) nonoverlapping bands. Fig. 2(a), (c) was created using parameter extraction and reconstruction in the magnitude domain; Fig. 2(b), (d) uses parameterization in the power domain. The dashed lines represent the original (unmodified) spectra for the left ear of HRTF set 003 from the CIPIC database [33] for an elevation of 0° and an azimuth of $-55°$. The solid lines represent the parametric reconstructions. As can be observed from Fig. 2, both the nonoverlapping and the overlapping reconstructions are quite accurate for frequencies up to about 6 kHz, while for higher frequencies fine-structure details are lost. Furthermore for nonoverlapping bands one can clearly see the effect of the rectangular parameter bands that result in a stepwise constant spectrum reconstruction. This phenomenon does not occur for overlapping bands because of the intrinsic interpolation of the associated $\boldsymbol{M}$ to reconstruct the spectra.

The time-domain HRIR and its parameter-based reconstruction are shown in Fig. 3. The parameterization was performed for the same impulse response and parameter resolution as used in Fig. 2, using overlapping parameter bands in the magnitude domain. As can be observed by comparing Fig. 3(a) and 3(b), both impulse responses differ significantly. Also the impulse response of the reconstructed HRIR is more symmetric than the original one.

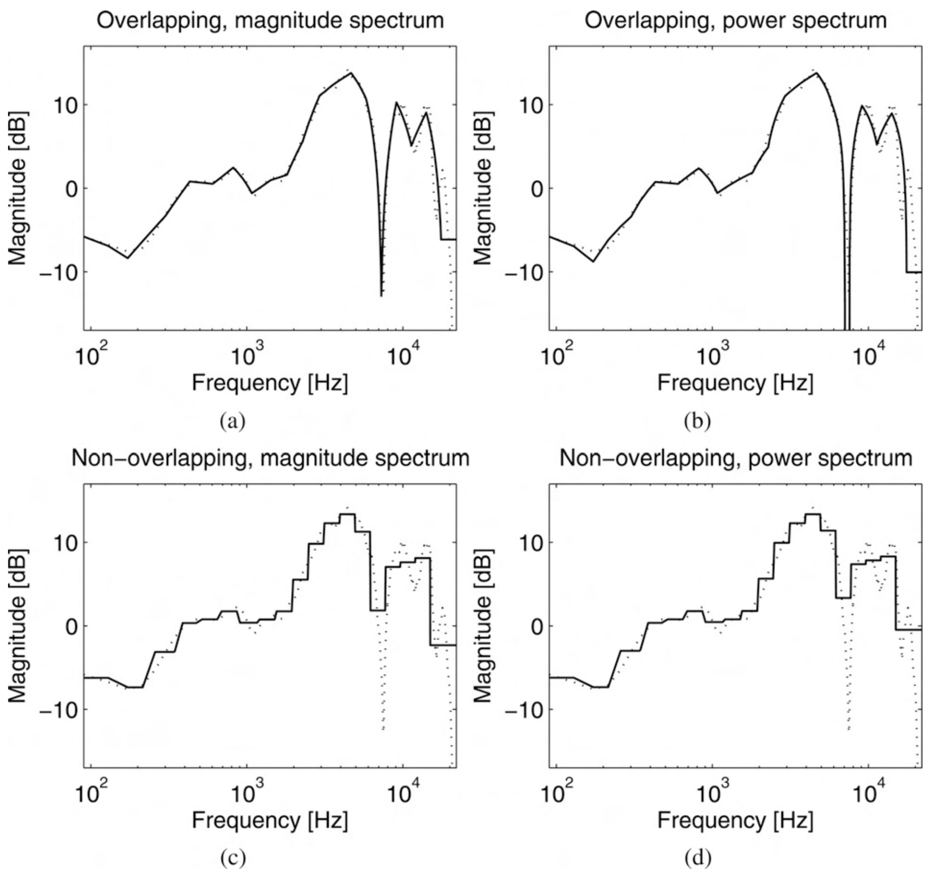
2 EXPERIMENTS

2.1 Generic Procedure and Stimuli

To evaluate the quality of the HRTF parameterization method, several listening tests were carried out to estimate the minimum number of spectral parameters required for inaudible differences between original and reconstructed HRTFs. For this purpose short bursts of pink noise (of 300-ms duration with 20-ms onset–offset ramps and 200-ms between-interval silences) were used as test signals. A three-interval, forced-choice procedure was employed to determine threshold values for variables under test. Subjects were presented with three intervals of filtered noise in random order, of which one was processed by parameterized HRTFs whereas the remaining two intervals were convolved with unmodified HRTFs. Subjects had to indicate the "odd one out," and feedback was provided after each trial. For those experiments that varied an experimental variable adaptively to establish a threshold value, this variable was modified according to a two-down, one-up procedure [34] to estimate the 70.7% correct score. The step size of the experimental variable was also modified adaptively. For experiments where γ was the experimental variable, an initial value of $\gamma = 4$ and a step size of 1 were used. Initially the step size of the experimental variable was modified adaptively. After two reversals of the experimental variable track, the step size was halved until a total of six reversals were obtained. Subsequently another eight reversals with a fixed step size were employed of which the median was used as threshold value. Each combination of experimental variable and subject was repeated three times.

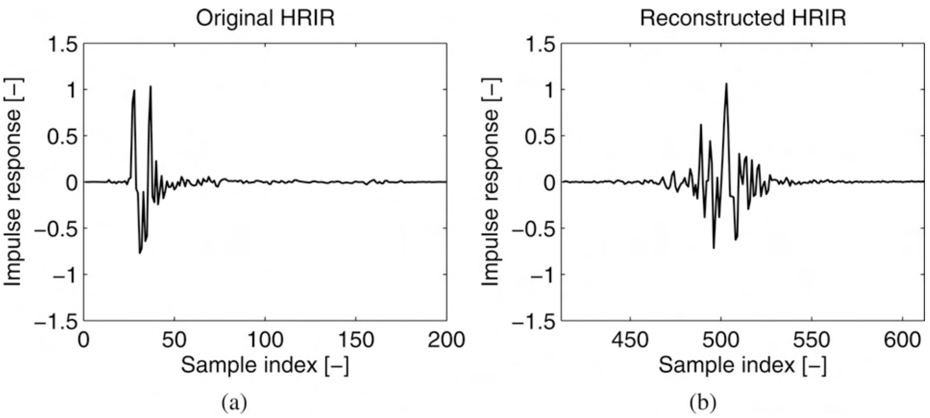
For each trial a new pink-noise sample was generated to prevent dependencies on specific noise tokens. The filtered signals were generated by a PC using an RME DIGI 96/8 PAD soundcard, amplified (Tucker-Davis PA5 and RP2.1), and presented through headphones (Beyerdynamic DT 990) at a level of about 75 dB SPL. No compensation for headphone characteristics was used. Subjects were seated in a soundproof listening booth. Five subjects participated in each test with ages between 25 and 35 years, all of whom had substantial experience in psychoacoustic listening tests.

HRTF sets 003, 018, 137, and 147 were used from the CIPIC database [33]. This database contains (anechoic) HRTF pairs for a wide range of positions and persons. To limit the number of experiments to a reasonable amount, a total of nine sound-source positions was selected for listening tests, which are specified in Table 2 in terms of azimuth and elevation. The positions are specified for both interaural polar coordinates as used in the CIPIC database, and vertical–polar coordinates. Positions 1 and 2 are exactly in front and above the listener, whereas the remaining positions are scattered around the listener. The CIPIC HRTFs have a length of



(a) (b) (c) (d)

Fig. 2. Original (···) and parametric(—) HRTF reconstructions for left ear of HRTF set 003 of the CIPIC HRTF database at 0° elevation and $-55°$ azimuth and 20 spectral coefficients. (a), (b) Overlapping parameter bands. (c), (d) Nonoverlapping bands. (a), (c) and (b), (d) Parameterization employed in magnitude and power spectrum domains, respectively.



(a) (b)

Fig. 3. Original and reconstructed HRIR for left ear of HRTF set 003 of the CIPIC HRTF database at 0° elevation and $-55°$ azimuth and 20 spectral coefficients. Parameterization was performed using overlapping parameter bands in the magnitude domain and $N = 1024$. (a) Original HRIR. (b) Reconstructed HRIR. In (b) only a 200-sample subsection of the 1024-sample HRIR is shown for clarity.

200 samples which were zero-padded to a length of 1024 samples. Parameter extraction and HRTF reconstruction were subsequently employed using $N = 2K = 1024$.

2.2 Experiment I: Effect of Analysis Filter Shapes and Parameterization Domain

In a first test the effects of overlapping and nonoverlapping parameter bands, as well as potential differences between power and magnitude domain parameters were investigated. In these experiments the minimum value for γ was acquired that resulted in just-inaudible differences between parametric and original HRTFs. The CIPIC HRTF set 003 was employed in this test using the following three configurations:

1) nO-P—Nonoverlapping parameter bands applied in the power spectrum domain
2) O-P—Overlapping parameter bands in the power spectrum domain (for this configuration only sound-source positions 1, 6, and 9 were measured)
3) O-M—Overlapping parameter bands in the magnitude spectrum domain.

The threshold values as a function of position, averaged across subjects and repetitions, are shown in Fig. 4. Error bars denote the standard errors of the mean. Different symbols represent different configurations: downward triangles represent nonoverlapping bands in the power domain (nO-P); upward triangles represent overlapping bands in the power domain (O-P), and circles represent overlapping bands in the magnitude domain (O-M). As can be observed in Fig. 4, threshold values are in the range of between 1.0 and 3.5. Large differences are observed across the various sound-source positions. Sound-source position 2 is associated with the highest threshold values, whereas positions 1, 3, and 6 have the lowest thresholds. Furthermore the overlapping parameter bands used in the magnitude domain consistently give higher thresholds (that is, lower numbers of parameter bands) than the two other configurations, which tend to perform very similarly (except for position 1). For the best performing configuration (O-M) thresholds vary between 1.35 for position 3 and 3.5 for position 2.

A four-way analysis of variance was carried out with the sound-source position, parameterization method, subject, and trial number as independent variables (including second-order interactions), and the threshold as dependent variable. Only the nO-P and O-M configurations were included to ensure a balanced data set. The results are shown in Table 3. As can be observed from the F ratios

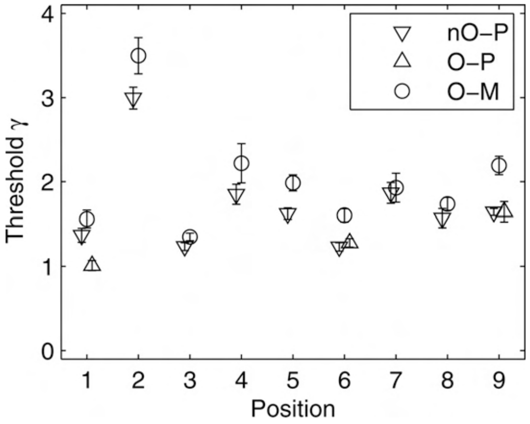


Fig. 4. Threshold γ as a function of sound- source position and three different parameter extraction and reconstruction methods. Data represent mean across five subjects and three repetitions. Error bars — standard errors of the mean.

Table 3. ANOVA table for results of experiment I.*

Source	SS	df	MS	F	p
Subject	7.629	4	1.9073	17.89	0
Trial	0.518	2	0.2591	2.43	0.0909
Position	80.198	8	10.0248	94.02	0
Method	6.033	1	6.0331	56.58	0
Subject * trial	1.138	8	0.1422	1.33	0.2292
Subject * position	12.686	32	0.3965	3.72	0
Subject * method	7.819	4	1.9548	18.33	0
Trial * position	2.827	16	0.1767	1.66	0.0585
Trial * method	0.424	2	0.2120	1.99	0.1398
Position * method	1.797	8	0.2246	2.11	0.0372
Error	19.62	184	0.1066		
Total	140.69	269			

*SS—sum of squares; df—degrees of freedom; MS—mean square; F ratio; p—maximum probability for observed F ratio under the null hypothesis.

Table 2. Azimuth and elevation for nine sound-source positions under test expressed in two coordinate systems.*

Position number	1	2	3	4	5	6	7	8	9
CS1									
Azimuth (deg)	0	0	−10	−55	−20	10	45	65	25
Elevation (deg) (1)	0	90	−22.5	11.25	135	191.25	157.5	−33.75	33.75
CS2									
Azimuth (deg)	0	0	−10	−55	160	−170	−135	65	25
Elevation (deg) (CS2)	0	90	−22.5	11.25	45	−11.25	22.5	−33.75	33.75

*CS1 — interaural polar coordinates as used in CIPIC database for which azimuth is limited to ±90° (see [33], ftn.3]); CS2 — vertical–polar coordinates using an azimuth range of ±180°.

obtained for the main factors subject, position, and method, these factors were found to have a significant effect on the measured thresholds. Furthermore all second-order interactions between these three factors were also found to be significant. On the other hand the trial number and interactions involving the trial number were not found to be of significance if a 5% confidence level is used. One can thus conclude that the O-M method performs significantly better than the nO-P method, that thresholds depend on the subject and sound-source position, and that subjects tend to respond differently to changes in the HRTF parameterization method and the sound-source position. Given the significance of the various interactions, intersubject differences cannot be explained by a difference in sensitivity only.

To investigate the effect of γ on the strength of the cues that subjects could use, psychometric curves were reconstructed from the adaptive level tracks. For each sound-source position, subject, parameterization method, and trial the subject's answer and the associated value of γ were logged. From these data the proportion of correct answers could be constructed as a function of γ. Low values of γ are associated with a relatively high number of parameters and hence weak cues that subjects could use. With increasing γ the proportion of correct answers is expected to increase. The observed γ values were grouped in bins centered around values of 0.6, 0.8, 1.0, 1.25, 1.5, 2.0, 2.5, 3.0, and 4.0. For each bin the proportion of correct answers was computed across all sound-source positions and subjects. The results are visualized in Fig. 5 for the nO-P and O-M methods by the downward triangles and circles, respectively. The dashed horizontal line indicates the percentage of correct responses based on chance (33.33%). The results indicate that for $\gamma = 0.6$ both methods have a score that is very close to the result obtained by random guessing. For $\gamma = 0.8$ the O-M method still performs quite well with proportion correct of 31.4%. The nO-P method, on the other hand, has scores that are significantly above chance (56.9% for $\gamma = 0.8$). For higher values of γ both curves increase monotonically, and the nO-P scores remain higher than the O-M scores.

2.3 Experiment II: Effect of HRTF Set

In this experiment γ thresholds were measured for sound-source positions 1 and 9 of CIPIC HRTF sets 003, 018, 137, and 147. Position 1 was selected because it showed to be quite critical in the previous experiment; position 9 gave rise to large differences between parameterization methods. The best performing parameterization method of the previous experiment was used (based on overlapping bands applied on the magnitude spectra). The results of the listening test are visualized in Fig. 6 as average threshold across subjects and repetitions. Different symbols denote different HRTF sets; error bars indicate the standard errors of the mean. As can be observed from Fig. 6, the thresholds for γ are between 1.0 and 2.6. The HRTF set 003 (crosses) resulted in relatively high threshold values for γ compared to the results for 137 (upward triangles).

An analysis of variance was carried out on the results shown in Fig. 6 with the factors HRTF set, subject, position, and trial number as independent variables (including second-order interactions), and the γ threshold value as dependent variable. Qualitatively the results are the same as those of the previous experiment: all factors and interactions were found to be significant at a 5% confidence level, except those involving the trial number. The trial number itself resulted in $p = 0.0774$. The interactions with subject, position, and HRTF set resulted in $p = 0.4288$, $p = 0.8003$, and $p = 0.9938$, respectively. All other p values were found to be zero, except for the interaction between subject and HRTF set, which yielded $p = 0.0436$, and the interaction between position and subject with $p = 0.0013$. This result indicates that subjects did not only respond differently to changes in the sound-source position, but that their responses also depended on the HRTF pair under test.

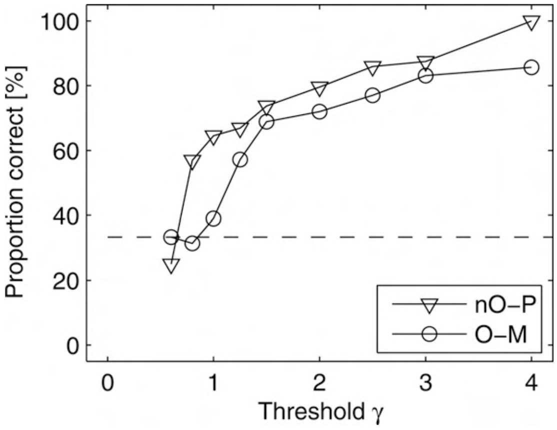


Fig. 5. Psychometric functions for nO-P method (downward triangles) and O-M method (circles) reconstructed from adaptive level tracks and pooled across sound-source positions and subjects. – – – expected proportion correct at chance level.

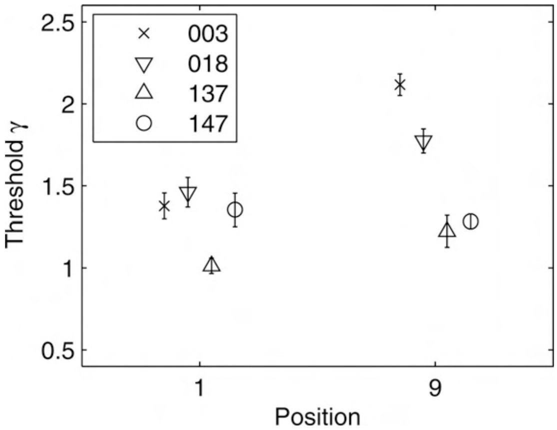


Fig. 6. Threshold γ for sound-source positions 1 and 9 using different HRTF sets (see legend). Data represent mean across five subjects and three repetitions. Error bars indicate standard errors of the mean.

2.4 Experiment III: Interaural Phase Parameterization

In this experiment further reductions in the HRTF parameter set are investigated by partial removal and simplification of the interaural phase data. In a first test the frequency range for which interaural phase parameters were synthesized was adaptively decreased. The interaural phase parameter of the last (highest) band was repeated for all remaining bands to circumvent strong phase discontinuities in the reconstructed HRTF phase spectrum. Subjects had to indicate the pink-noise stimulus constructed with the parameterized HRTFs out of a trial, including two references convolved with original HRTFs. The nine positions given in Table 2 of CIPIC HRTF set 003 were used. Parameterization was employed according to the O-M method with a fixed value of $\gamma = 1$. For the initial trial during an adaptive run, all interaural phase parameters were set to zero. Depending on the answer of the subjects, the number of phase parameters was adaptively varied according to a one-down, two-up rule. Hence after two subsequent correct answers the maximum frequency for phase synthesis was increased by one parameter band; after each incorrect answer the frequency range was decreased by one parameter band. The run continued until eight reversals were observed. The median value of these eight reversals was used as threshold value. A run was aborted if subjects provided four erroneous answers in a row for stimuli that did not include any interaural phase synthesis, assuming that no significant cue was available for that subject. Three repetitions were carried out for each sound-source position.

The pink-noise stimulus employed in this experiment was amplitude modulated with a 50-Hz raised sinusoid. The sine-wave modulation served to enhance temporal envelope cues at high frequencies [35]–[39].

The threshold frequencies as a function of the sound-source position are provided in Fig. 7. Thresholds are shown as average across repetitions and subjects; error bars denote the standard errors of the mean. Depending on the sound-source position, thresholds vary between 221 Hz for position 1 and 1215 Hz for position 4. Not surprisingly positions 1 and 2, which are positioned on the midsagittal plane, have very low thresholds, which can be attributed to the very small interaural time differences present in the original HRTF pairs. The most lateral positions (4, 7, and 8) resulted in the highest thresholds.

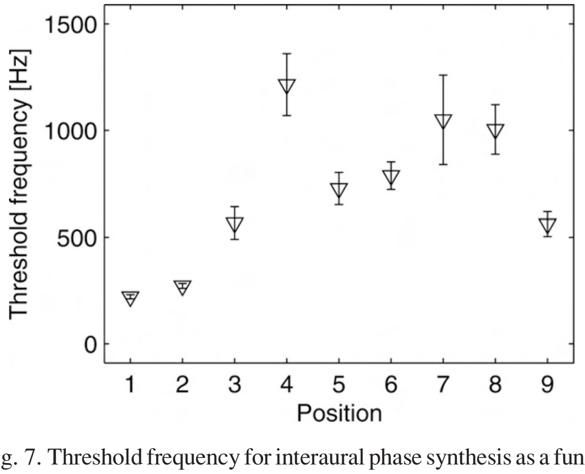


Fig. 7. Threshold frequency for interaural phase synthesis as a function of sound-source position averaged across five subjects and three repetitions. Error bars indicate standard errors of the mean.

In a second test the interaural phase information was reduced by fitting a single interaural time delay for each sound-source position and reconstruction of the interaural phase characteristic from this constant delay. The interaural time delay for an HRTF pair was computed by minimizing the squared error between the synthesized and actual interaural phase curve in the frequency range of 0 to 1500 Hz. During synthesis the estimated interaural time delay was employed to the full frequency range and applied in the frequency domain. The magnitude spectra of the HRTFs were parameterized using overlapping parameter bands operating in the magnitude domain with a fixed value of $\gamma = 1$.

The CIPIC HRTF set 003 was employed using the same nine sound-source positions as in the previous experiments. Four subjects were presented with three intervals of pink noise. Two intervals were filtered with the unmodified HRTFs; one interval was processed with parameterized HRTFs using a single interaural time difference value. Feedback was provided after each trial. Subjects had to evaluate 30 trials for each condition. Conditions were repeated three times. The percentage of correct responses across all 90 trials was used as performance indicator.

A control test was also performed, which used the same experimental procedure as described, except for the fact that full interaural phase parameterization was used (that is, one interaural phase parameter for each parameter band). This condition allows to verify whether results in the experiment can be attributed to parameterization per se, or result from the constant interaural time difference property.

The results averaged across subjects are shown in Fig. 8. The percentage of correct responses is shown as a function of the sound-source position. Error bars denote

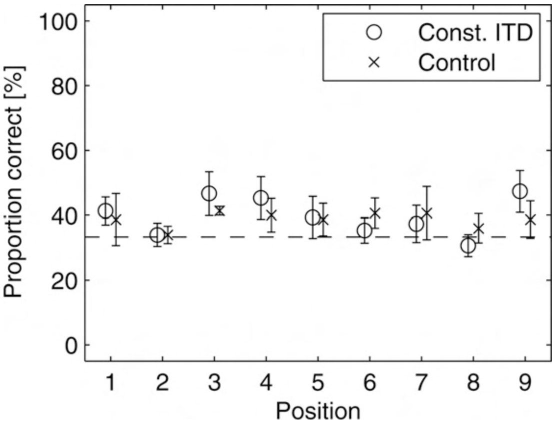


Fig. 8. Percentage correct responses for HRTFs parameterized using a constant interaural time difference as a function of the sound-source position. Error bars indicate standard errors of the mean across subjects. – – – proportion of correct responses based on chance; ○ — constant interaural time delay configuration; × — control condition employing an individual interaural phase parameter for each parameter band.

the standard errors of the mean. The dashed line represents the percentage correct based on random responses (33.3%). The circles represent the constant interaural time difference (Const. ITD) configuration; the control condition employing one individual interaural phase parameter for each parameter band is denoted by the crosses.

As can be observed from Fig. 8, the percentage correct responses for the constant interaural time difference condition varies between 30.7% for position 8 and 47.3% correct for position 9. The pooled data differ significantly from the chance level of 33% (single-sided $t(44) = 3.48$, $p = 0.0006$), indicating performance better than chance. For the control condition the percentage of correct responses varies between 34 and 41.3% for positions 2 and 3, respectively. The pooled data for the control condition were also observed to be significantly better than chance (single-sided $t(44) = 3.23$, $p = 0.0012$). However, when both conditions were compared, no significant differences were observed (single-sided $t(44) = 0.49$, $p = 0.3137$). Even if one compensates for the fact that multiple-comparison tests are used by Bonferroni adjustments or using the Holm–Bonferroni method, the conclusions remain the same if confidence levels of 5% are employed. Despite the statistical significance of both conditions with respect to chance levels, it should be noted that the average percentages correct for both conditions amount to only 39.6 and 38.7%, respectively, indicating only very weak cues that subjects could use.

2.5 Experiment IV: Effect of Spatial Interpolation

Most HRTF databases comprise impulse responses for a limited set of measurement positions, typically with an angular distance of about 5°, which seems the minimum resolution from a spatial sampling point of view [40], [41]. When sound-source positions in between measurement positions are required (for example, when dynamic positioning of sound sources is required), HRTF estimates for positions in between measurement positions are required to circumvent switching artifacts. Various schemes have been proposed for interpolation (see [42], [43], [5], [1], [44]), which in many cases combine multiple HRTF representations to compute a weighted average. The weights are typically inversely proportional to the distance between desired and measured positions.

In this experiment the effect of interpolation was investigated. The goal was to further reduce the amount of data required to represent an HRTF set by reducing the spatial sampling frequency of HRTF parameters. The CIPIC database employs a horizontal spatial sampling period of 5° and a vertical period of 5.625°. In the current experiment the spatial sampling frequency was reduced by a factor 2, 3, or 4, and the eliminated positions were reconstructed by means of linear interpolation of the HRTF parameters of the surrounding positions, similar to what was proposed in [1]. The interpolated HRTF parameters were reconstructed either from their neighboring positions in a vertical direction or from neighboring positions in a horizontal direction. The experiments were performed for positions 1 and 9 as outlined in Table 2 and for HRTF set 003. All combinations of sound-source position, interpolation direction, and spatial sampling reduction factor are summarized in Table 4. For a reduction in measurement positions by a factor of 2 or 4, the interpolated position was situated in the spatial center of the remaining positions. For a spatial resolution reduction by a factor of 3, both of the two eliminated positions were tested separately.

Table 4. Overview of all configurations used in parameter interpolation experiment.*

Configuration	Position	Azi	Ele	Dir	Azil	Ele1	Azi2	Ele2
1H2	1	0.00	0.00	Hor	−5.000	0.000	5.000	0.000
1H3a	1	0.00	0.00	Hor	−5.000	0.000	10.000	0.000
1H3b	1	0.00	0.00	Hor	−10.000	0.000	5.000	0.000
1H4	1	0.00	0.00	Hor	−10.000	0.000	10.000	0.000
1V2	1	0.00	0.00	Ver	0.000	−5.625	0.000	5.625
1V3a	1	0.00	0.00	Ver	0.000	−5.625	0.000	11.250
1V3b	1	0.00	0.00	Ver	0.000	−11.250	0.000	5.625
1V4	1	0.00	0.00	Ver	0.000	−11.250	0.000	11.250
9H2	9	25.00	33.75	Hor	20.000	33.750	30.000	33.750
9H3a	9	25.00	33.75	Hor	20.000	33.750	35.000	33.750
9H3b	9	25.00	33.75	Hor	15.000	33.750	30.000	33.750
9H4	9	25.00	33.75	Hor	15.000	33.750	35.000	33.750
9V2	9	25.00	33.75	Ver	25.000	28.125	25.000	39.375
9V3a	9	25.00	33.75	Ver	25.000	28.125	25.000	45.000
9V3b	9	25.00	33.75	Ver	25.000	22.500	25.000	39.375
9V4	9	25.00	33.75	Ver	25.000	22.500	25.000	45.000

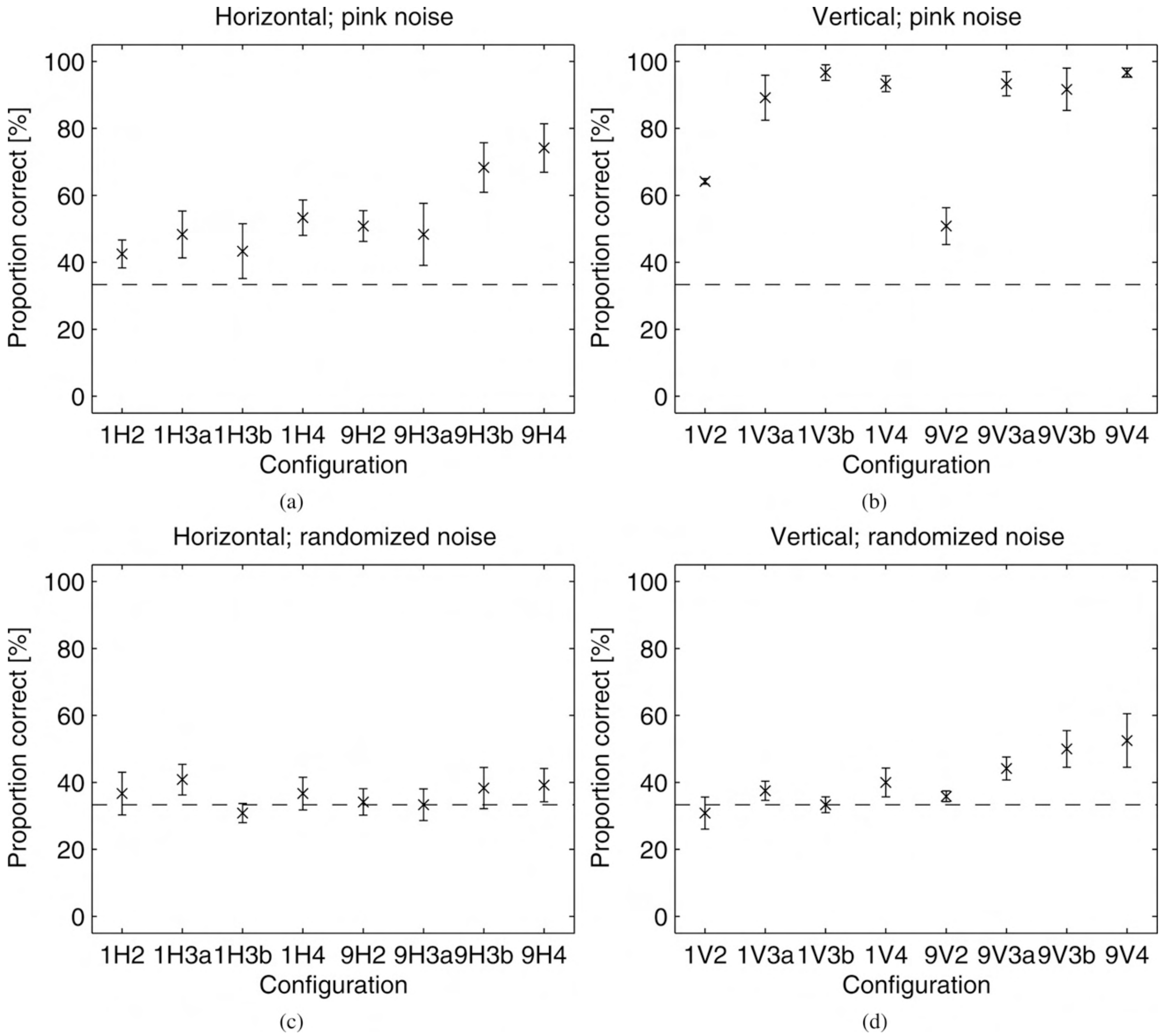
*Azi, Ele—target sound source position; Dir—interpolation direction; Azi1, Ele1, Azi2, Ele2—neighboring positions for interpolation.

Two types of stimuli were used in this test. The first stimulus comprises a pink-noise burst with the same duration, level, and onset/offset ramps as in experiment I. A second stimulus consisted of the same pink noise, but included an additional level roving per critical band of −5, 0, or +5 dB. The level rove was applied independently in each critical band. This stimulus was denoted "randomized noise." It served to exclude monaural timbre cues from the experimental data and forced subjects to use spatial cues only. Experiments were performed using the O-M parameter band configuration and a fixed value of $\gamma = 1$. In a three-interval forced-choice procedure, four subjects had to indicate which interval was created using interpolated HRTF parameters. The reference intervals were generated using convolution with the original HRTFs for the interpolation position.

The results expressed as proportion of correct responses are shown in Fig. 9. Fig. 9(a), (c) represent interpolation in the horizontal direction; Fig. 9(b), (d) show data obtained for vertical interpolation. Data in the Fig. 9(a), (b) represent the pink-noise stimulus; these in Fig. 9(c), (d) correspond to the randomized noise stimulus. All error bars denote standard errors of the mean across all subjects and repetitions. The horizontal dashed line is the expected percentage of correct responses based on chance.

The data obtained for horizontal interpolation and a pink-noise stimulus resulted in proportions correct of between 42.5 and 72.2%. There is a general tendency of increasing subject performance with a decrease in spatial HRTF resolution: the 1H4 and 9H4 conditions score higher than all other configurations, and 1H2 and 9H2 are among the lowest scores. A similar trend can be observed for vertical interpolation [Fig. 9(b)]. However, with the exception of the 9V2 configuration, all scores are substantially higher than for horizontal interpolation with maximum values of up to 96.7% correct.

The randomized noise stimuli [Fig. 9(c), (d)] give rise to a different response pattern. Most configurations have



(a)

(b)

(c)

(d)

Fig. 9. Percentage correct responses for HRTFs reconstructed from interpolated parameters. (a), (c) Interpolation in horizontal direction. (b), (d) Interpolation in vertical direction. (a), (b) Data obtained for pink-noise stimuli. (c), (d) Randomized noise stimuli. Error bars indicate standard errors of the mean.

scores that are equal to or just above chance level. The highest scores are obtained for vertical interpolation of position 9, with a maximum score of 52.5% for the 9V4 configuration.

3 DISCUSSION

The results of the various experiments confirm the hypothesis that HRTF magnitude and phase spectra contain fine-structure details which seem perceptually irrelevant in the case of broadband stimuli, in line with results reported in other publications (see [15]–[18]). For the parameterization method presented, all measured thresholds for γ amounted to at least 1.0, indicating that none of the configurations required more than one set of parameters per critical band, assuming a threshold defined by 70.7% correct responses. This finding supports the notion that sound-source localization is mediated by the level and interaural phase cues that are acquired at a critical-band level. Moreover, subjects were allowed to use any cue to detect the target interval, and feedback was provided after each trial. One can thus conclude that for $\gamma = 1$ cues such as spatial or timbral changes induced by HRTF parameterization were not strong enough for detection.

The results on interaural phase information reduction confirm earlier reports (see [19], [15]) that the frequency-dependent interaural delay can be replaced by a constant, frequency-independent delay. Furthermore for the HRTFs employed in the current test, subjects were not able to use high-frequency (above 1500 Hz) mismatches in interaural phase characteristics as cues to detect parameterized HRTFs, indicating that high-frequency phase information is perceptually irrelevant for the stimuli used in this test, in line with earlier observations [20]–[22]. Interestingly the parameterization method used ignores all absolute phase information of the HRTF spectra, a property that could not be exploited by the subjects. Hence the symmetric distribution of interaural phase information in combination with pink-noise stimuli seems a sufficient prerequisite for transparent HRTF parameterization.

The exact number of parameters required for transparent HRTF representations was found to be dependent on various variables, including the HRTF set, the sound-source position, the subject under test, and the method of parameter extraction. Furthermore interactions between these factors were found to be significant, indicating that different subjects responded differently to changes in HRTF sets, sound-source positions, and HRTF parameterization methods. Given the complexity of the responses, it is difficult to pinpoint the underlying mechanisms for the observed effects exactly. A factor that could be of importance is the difference between the employed HRTFs and the personal HRTF characteristics of the subject under test. The sound-source localization cues available in the test could be different from those in normal listening conditions, potentially resulting in an increased difficulty to detect the target interval. Another effect that could be of importance is the detection strategy employed by the listener. Subjects could use any cue to detect the target interval, including spatial as well as timbral attributes. They could be focusing on different aspects depending on the sound-source position, the HRTF set, and the parameterization method.

When comparing different HRTF parameterization methods it was observed that overlapping parameter bands operating in the HRTF magnitude domain result in the most efficient parameterization, that is, requiring the lowest number of parameters. The mean value of γ across all nine sound-source positions amounted to 2.0, indicating that on average one parameter set for two critical bands allowed subjects to just reach 70.7% correct responses.

The results obtained are especially of interest in relation to the recent audio coding standards MPEG Surround and Spatial Audio Object Coding. These standards provide HRTF parameter interfaces that individuals or manufacturers can use to specify their proprietary HRTFs. The parameters used are qualitatively identical to those employed in this study, but differ slightly in terms of the exact form of the parameter band definition matrix. More specifically, these standards prescribe 28 parameter bands that are approximately equally spaced on an ERB scale. In the current context this would correspond to an average γ value of 1.43. Based on the results obtained, this value seems sufficient for high-quality HRTF parameterization in which subjects have only weak cues (if any) to detect any mismatches between original and parameterized HRTF pairs. Also the fact that these standards use only interaural phase reconstruction up to about 1700 Hz seems in line with the results obtained in the experiments involving limitations in the interaural phase reconstruction.

The results for parameter-based interpolation of HRTFs indicate that a reduction of the spatial sampling period from 5 to 10° in the horizontal direction results in only a small increase in the percentage of correct responses. In this experiment the references consisted of original HRTFs, and hence the subjects' responses comprise the combined effect of parameterization and spatial interpolation. The results of Fig. 8 indicate that the parameterization only gives rise to percentages of correct responses of 38.7% for positions 1 and 9. These values increase to 42.5 and 50.8%, respectively, when both parameterization and interpolation are employed. For vertical interpolation these values amounted to 64.2 and 50.8% for positions 1 and 9, respectively. These values are all below the 70.7% correct that was used to determine the thresholds in the experiments employing adaptive procedures. One can thus conclude that for the proposed parameterization scheme, interpolation of HRTF parameters for static sound sources is feasible for HRTF measurement positions that are 10° apart. However, further reductions in the number of HRTF measurement positions lead to substantial increases in correct responses, and hence interpolation across larger angular distances is not recommended for near-transparent binaural rendering.

Interestingly, vertical interpolation seems more critical than horizontal interpolation. In particular for spatial sampling frequency reductions of factors of 3 or 4, vertical interpolation gives rise to substantially more correct responses than horizontal interpolation. A likely cause of this observation is the position dependence of the HRTF spectral characteristics. Changes in elevation are typically associated with shifts of peaks and troughs across frequency. Azimuth, on the other hand, is coded by relative level and time differences between the left and right-ear HRTF. Interpolation of HRTF parameters effectively crossfades the HRTF spectra, and this process will have a greater distorting effect on spectral elevation cues than on binaural azimuth cues.

The parameter-interpolation results obtained with the randomized stimulus resulted in significantly lower performance scores than those for pink noise. This suggests that subjects were predominantly using monaural timbral cues to detect the effect of interpolation, and could not rely on interaural differences or elevation cues. This result was to be expected for position 1 since interaural differences should be close to 0 for a frontal position for both original and interpolated HRTF parameters. The fact that position 9 also resulted in relatively low performance scores for the randomized stimulus indicates that the interaural cues are well represented by parameter interpolation, and that subjects performed the task using monaural cues. It should be noted, however, that it is unclear whether these conclusions also hold for dynamically varying sound-source positions in which the interpolation process (that is, the parameter values and their relative weights) changes over time. Other studies have shown similar HRTF spatial resolution requirements for dynamic stimuli [5], [45].

Although this study provides important insights into perceptual irrelevancies in HRTF pairs, some questions have not been addressed. For example, only stimuli consisting of pink noise and amplitude-modulated pink noise were employed. Tonal stimuli or stimuli involving strong transients may be more susceptible to phase and magnitude modifications than the current signals, possibly resulting in lower γ thresholds. Furthermore high-pass stimuli may also be more critical to determine the (in) sensitivity to high-frequency interaural phase information.

A second limitation involves the use of generic HRTFs. A mismatch between the HRTFs used and the HRTF characteristics of the individual subjects may result in a modified sensitivity to changes in spectral or timbral cues. For localization accuracy, on the other hand, the use of generic HRTFs hardly seems to degrade the performance compared to individualized HRTFs [11] except for positions in the median plane [9]. Other studies indicate that personalized HRTFs do not always provide the most natural or realistic simulation of a virtual auditory space [46]. A cross-check of the current experiments involving personalized HRTFs would therefore be a valuable exercise. Another interesting topic that is not covered in this study is the incorporation of head tracking. The incorporation of head movements has been shown to resolve sound-source localization errors for both personalized and generic HRTFs, essentially resolving the majority of localization accuracy differences between generic and personalized HRTFs [47], [48], [12], [49].

A third limitation of this study is the small number of subjects used in relation to intersubject differences in the audibility of the parameterization process. Although our statistical analysis indicates both significant main effects as well as interactions related to subject variability, the number of tests and subjects is too small to identify underlying causes or mechanisms for such dependencies. Future, more extensive work of subject dependencies is needed to better understand the implications of this result, and to investigate potential links between the audibility of parameterization and localization accuracy or naturalness of the virtual sound sources.

A fourth limitation relates to the compensation of headphone characteristics, which was not used in the current study. It has been suggested that spectral changes induced by headphones can be of the same order of magnitude as intersubject differences in HRTFs [50]. Hence the absence of headphone equalization may have had an impact on the results obtained. It is, however, difficult to predict what this impact would be qualitatively. Notches in the headphone frequency characteristic may mask certain spectral discrepancies induced by the HRTF parameterization method. However, a peak in the headphone characteristic my have the opposite effect. Moreover it has been indicated that correct equalization of headphone characteristics is difficult in practice because of its dependence on headphone positioning [51], [52], resulting in potentially considerable spectral discrepancies, even in the case of careful headphone compensation [53], [50]. Also there is evidence that for generic HRTFs and a limited set of transducers, headphone equalization does not improve sound-source localization [54] while it does seem to have some benefits for individualized HRTFs [52].

On the positive side, the current experiments were performed using single sound-source positions. Results by Breebaart and Kohlrausch [17] have indicated that multiple simultaneous sound sources result in a reduced sensitivity to HRTF information reduction. Although their stimuli and experimental procedure differ significantly from this study, their results indicated that two simultaneous sound sources can be subjected to a higher degree of HRTF fine-structure removal than single sound sources.

4 CONCLUSIONS

From the experimental results we can derive the following conclusions. First the results indicate that sound-source localization involving pink-noise stimuli is mediated by level and interaural phase cues that are evaluated on averaged values per critical band. As a result HRTFs can be accurately described using a limited set of magnitude parameters and a single interaural delay for each HRTF pair. Furthermore it was observed that an absence of both high-frequency interaural phase information as well as absolute HRTF phase characteristics does not result in audible changes.

Depending on the HRTF set used, the subject under test, the sound-source position, and the details of the HRTF parameterization scheme, a further reduction in the number of parameters can in some cases be employed. However, given the complexity of subjects' responses, it seems difficult to predict what the minimum number of parameters should be, except for the fact that one parameter set per critical band is a safe lower boundary.

The parameter-interpolation experiment indicated that the spatial resolution of HRTF measurement positions can be reduced from 5 to about 10° without introducing strong perceptual differences. If the spatial resolution is decreased further, the results indicate that subjects use predominantly monaural cues to detect differences between a parameter-interpolated HRTF pair and original HRTFs, and that spatial attributes remain intact up to about 20° angular spacing of HRTF measurement positions. The results also indicated that decreasing the vertical resolution seems more critical than decreasing the horizontal resolution.

The results so far have been obtained with anechoic HRTFs only. It is not evident how the parametric approach can be extended to echoic responses. The work by Merimaa [55] could be a good starting point for developing perceptual parameterization schemes for binaural room impulse responses.

5 ACKNOWLEDGMENT

The authors would like to thank the subjects who participated in the listening tests for their valuable time, and their colleagues Mun Hum Park and Janto Skowronek for their helpful suggestions to improve the manuscript. The authors would also like to thank the anonymous reviewers for their judicious comments to improve the manuscript.

6 REFERENCES

[1] E. M. Wenzel, J. D. Miller, and J. S. Abel, "Sound Lab: A Real-Time, Software-Based System for the Study of Spatial Hearing," presented at the 108th AES Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 48, p. 360 (2000 Apr.), preprint 5140.

[2] J. D. Miller and E. M. Wenzel, "Recent Developments in SLAB: A Software-Based System for Interactive Spatial Sound Synthesis," in *Proc. Int. Conf. on Auditory Display* (*ICAD 2002*) (Kyoto, Japan, 2002 July).

[3] F. L. Wightman and D. J. Kistler, "Headphone Simulation of Free-Field Listening. I. Stimulus Synthesis," *J. Acoust. Soc. Am.*, vol. 85, pp. 858–867 (1989).

[4] F. L. Wightman and D. J. Kistler, "Headphone Simulation of Free-Field Listening. II: Psychophysical Validation," *J. Acoust. Soc. Am.*, vol. 85, pp. 868–878 (1989).

[5] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of Three-Dimensional-Sound Reproduction Using a Virtual Auditory Display," *J. Acoust. Soc. Am.*, vol. 107, pp. 528–537 (2000).

[6] D. R. Begault, "Challenges to the Successful Implementation of 3-D Sound," *J. Audio Eng. Soc.*, vol. 39, pp. 864–870 (1991 Nov.).

[7] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Nonindividualized Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 94, pp. 111–123 (1993).

[8] F. L. Wightman and D. J. Kistler, "Individual Differences in Human Sound Localization Behavior," *J. Acoust. Soc. Am.*, vol. 99, p. 2470 (1996).

[9] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural Technique: Do We Need Individual Recordings?," *J. Audio Eng. Soc.*, vol. 44, pp. 451–469 (1996 June).

[10] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Evaluation of Artificial Heads in Listening Tests," *J. Audio Eng. Soc.*, vol. 47, pp. 83–100 (1999 Mar.).

[11] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," *J. Audio Eng. Soc.*, vol. 49, pp. 904–916 (2001 Oct.).

[12] P. J. Minnaar, S. K. Olesen, F. Christensen, and H. Møller, "The Importance of Head Movements for Binaural Room Synthesis," In *Proc. ICAD* (Espoo, Finland, 2001 July), pp. 21–25.

[13] B. G. Shinn-Cunningham, "The Perceptual Consequences of Creating a Realistic, Reverberant 3-D Audio Display," in *Proc. Int. Cong. on Acoustics* (Kyoto, Japan, 2004 Apr.).

[14] D. J. Kistler and F. L. Wightman "A Model of Head-Related Transfer Functions Based on Principal Components Analysis and Minimum-Phase Reconstruction," *J. Acoust. Soc. Am.*, vol. 91, pp. 1637–1647 (1992).

[15] A. Kulkarni and H. S. Colburn, "Role of Spectral Detail in Sound-Source Localization," *Nature*, vol. 396, pp. 747–749 (1998).

[16] J. Huopaniemi and N. Zacharov, "Objective and Subjective Evaluation of Head-Related Transfer Function Filter Design," *J. Audio. Eng. Soc.*, vol. 47, pp. 218–239 (1999 Apr.).

[17] J. Breebaart and A. Kohlrausch, "The Perceptual (Ir)relevance of HRTF Magnitude and Phase Spectra," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstract)*, vol. 49, p. 546 (2001 June), convention paper 5406.

[18] J. Breebaart and C. Faller, *Spatial Audio Processing: MPEG Surround and Other Applications* (Wiley, Chichester, UK, 2007).

[19] W. M. Hartmann and A. Wittenberg, "On the Externalization of Sound Images," *J. Acoust. Soc. Am.*, vol. 99, pp. 3678–3688 (1996).

[20] R. G. Klumpp and H. R. Eady, "Some Measurements of Interaural Time Difference Thresholds," *J. Acoust. Soc. Am.*, vol. 28, pp. 859–860 (1956).

[21] W. A. Yost, "Tone-in-Tone Masking for Three Binaural Listening Conditions," *J. Acoust. Soc. Am.*, vol. 52, pp. 1234–1237 (1972).

[22] F. L. Wightman and D. J. Kistler, "The Dominant Role of Low-Frequency Interaural Time Differences in Sound Localization," *J. Acoust. Soc. Am.*, vol. 91, pp. 1648–1661 (1992).

[23] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1997).

[24] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjörling, E. Schuijers, J. Hilpert, and

F. Myburg, "The Reference Model Architecture of MPEG Spatial Audio Coding," presented at the 118th Convention, of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 53, pp. 693, 694 (2005 July/Aug.), convention paper 6447.

[25] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Röden, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG Surround — The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," vol. 56, pp. 932–955 (2008 Nov.).

[26] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. van de Par, "Background, Concept, and Architecture for the Recent MPEG Surround Standard on Multichannel Audio Compression," *J. Audio Eng. Soc.*, vol. 55, pp. 331–351 (2007 May).

[27] J. Breebaart, L. Villemoes, and K. Kjörling, "Binaural Rendering in MPEG Surround," *EURASIP J. Appl. Signal Process.*, paper 732895 (2008).

[28] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG Surround — The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," *J. Audio Eng. Soc.*, vol. 56, pp. 932–955 (2008 Nov.).

[29] J. Breebaart, J. Herre, L. Villemoes, C. Jin, K. Kjörling, and J. Plogsties, "Multichannel Goes Mobile: MPEG Surround Binaural Rendering," in *Proc. AES 29th Int. Conf.* "Audio for Mobile and Handheld Devices" (Seoul, Korea, 2006 Sept. 2–4).

[30] J. Engdegard, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial Audio Object Coding (SAOC): The Upcoming MPEG Standard on Parametric Object Based Audio Coding," presented at the 124th Convention of the Audio Engineering Society, (Abstracts) www.aes.org/events/124/124thWrapUp.pdf, (2008 May), convention paper 7377.

[31] A. W. Bronkhorst, "Effect of Stimulus Properties on Auditory Distance Perception in Rooms," in *Physiological and Psychophysical Bases of Auditory Function*, D. J. Breebaart, A. J. M. Houtsma, A. Kohlrausch, V. Prijs, and R. Schoonhoven, Eds. (Shaker Publ., Maastricht, The Netherlands, 2001), pp. 184–191.

[32] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data," *Hear. Res.*, vol. 47, pp. 103–138 (1990).

[33] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF Database," in *Proc. 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics* (New Paltz, NY, 2001 Oct.), pp. 99–102.

[34] H. Levitt, "Transformed up–down Methods in Psychoacoustics," *J. Acoust. Soc. Am.*, vol. 49, pp. 467–477 (1971).

[35] N. Y.-S. Kiang, "Stimulus Representation in the Discharge Pattern of Auditory Neurons," in *The Nervous System: Human Communications and Its Disorders*, D. B. Tower, ed., vol. 3, pp. 81–96 (Raven Press, New York, 1975).

[36] D. H. Johnson, "The Relationship between Spike Rate and Synchrony in Responses of Auditory-Nerve Fibers to Single Tones," *J. Acoust. Soc. Am.*, vol. 68, pp. 1115–1122 (1980).

[37] L. R. Bernstein and C. Trahiotis, "Detection of Interaural Delay in High-Frequency Noise," *J. Acoust. Soc. Am.*, vol. 71, pp. 147–152 (1982).

[38] L. R. Bernstein and C. Trahiotis, "The Normalized Correlation: Accounting for Binaural Detection across Center Frequency," *J. Acoust. Soc. Am.*, vol. 100, pp. 3774–3787 (1996).

[39] S. van de Par and A. Kohlrausch, "A New Approach to Comparing Binaural Masking Level Differences at Low and High Frequencies," *J. Acoust. Soc. Am.*, vol. 101, pp. 1671–1680 (1997).

[40] T. Ajdler, L. Sbaiz, and M. Vetterli, "The Plenacoustic Function on the Circle with Application to HRTF Interpolation," in *Proc. ICASSP* (IEEE, 2005), pp. 273–276.

[41] T. Ajdler, C. Taller, L. Sbaiz, and M. Vetterli, "Interpolation of Head Related Transfer Functions Considering Acoustics," presented at the 118th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 53, p. 662 (2005 July/Aug.), convention paper 6327.

[42] E. M. Wenzel and S. H. Foster, "Perceptual Consequences of Interpolating Head-Related Transfer Functions during Spatial Synthesis," in *Proc. 1993 Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY, 1993).

[43] J. Chen, B. D. Van Veen, and K. E. Hecox, "A Spatial Feature Extraction and Regularization Model for the Head-Related Transfer Function," *J. Acoust. Soc. Am.*, vol. 97, pp. 439–452 (1995).

[44] F. P. Freeland, L. W. P. Biscainho, and P. S. R. Diniz, "Efficient HRTF Interpolation in 3D Moving Sound," in *Proc. AES 22nd Int. conf.* "Virtual, Synthetic, and Entertainment Audio" (Espoo, Finland, 2002 June 15–17).

[45] A. Lindau, H. J. Maempel, and S. Weinzierl, "Minimum BRIR Grid Resolution for Dynamic Binaural Synthesis," in *Proc. ASA and EAA Joint Conf. on Acoustics* (Paris, France, 2008), pp. 3853–3858.

[46] J. Usher and W. L. Martens, "Perceived Naturalness of Speech Sounds Presented Using Personalized versus Nonpersonalized HRTFs," in *Proc. 13th Int. Conf. on Auditory Display* (Montreal, Canada, 2007 June 26–29), pp. 10–16.

[47] F. L. Wightman and D. J. Kistler, "Resolution of Front–Back Ambiguity in Spatial Hearing by Listener and Source Movement," *J. Acoust. Soc. Am.*, vol. 105, pp. 2841–2853 (1999).

[48] U. Horbach, A. Karamustafaoglu, R. Pellegrini, P. Mackensen, and G. Theile, "Design and Applications of a Data-Based Auralization System for Surround Sound," presented at the 106th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 47, p. 528 (1999 June), preprint 4976.

[49] P. Mackensen, "Head Movements, an Additional Cue in Localization," Ph.D. thesis, Technische Universität Berlin, Berlin, Germany (2004).

[50] Z. Schärer and A. Lindau, "Evaluation of Equalization Methods for Binaural Signals," presented at the 126th Convention of the Audio Engineering Society,

(Abstracts) www.aes.org/events/126/126thWrapUp.pdf, (2009 May), convention paper 7721.

[51] A. Kulkarni and H. S. Colburn, "Variability in the Characterization of the Headphone Transfer Function," *J. Acoust. Soc. Am.*, vol. 107, pp. 1071–1074 (2000).

[52] F. Wightman and D. Kistler, "Measurement and Validation of Human HRTFs for Use in Hearing Research," *Acta Acustica/Acustica*, vol. 91, pp. 429–439 (2005).

[53] A. Lindau, T. Hohn, and S. Weinzierl, "Binaural Resynthesis for Comparative Studies of Acoustical Environments," presented at the 122nd Convention of the Audio Engineering Society, (Abstracts) www.aes.org/events/122/122ndWrapUp.pdf, (2007 May), convention paper 7032.

[54] D. Schonstein, L. Ferré, and B. F. G. Katz, "Comparison of Headphones and Equalization for Virtual Auditory Source Localization," in *Proc. ASA and EAA Joint Conf. on Acoustics* (Paris, France, 2008), pp. 4617–4622.

[55] J. Merimaa, "Analysis, Synthesis, and Perception of Spatial Sound—Binaural Localization Modeling and Multichannel Loudspeaker Reproduction," Ph.D. thesis, Helsinki University of Technology, Espoo, Finland (2006).

THE AUTHORS

J. Breebaart

F. Nater

A. Kohlrausch

Jeroen Breebaart received an M.Sc. degree in biomedical engineering and a Ph.D. degree in auditory psychophysics from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 1997 and 2001, respectively.

From 2001 to 2007 he was with the Digital Signal Processing Group at Philips Research, conducting research in the areas of spatial hearing, parametric audio coding, automatic audio content analysis, and audio effects processing. Since 2007 he has been the leader of the biometrics cluster of the Information and System Security Group at Philips Research, expanding his research scope toward secure and convenient identification.

Dr. Breebaart is a member of the Audio Engineering Society and IEEE. He contributed to the development of audio coding algorithms as recently standardized in MPEG and 3GPP, such as HE-AAC, MPEG Surround, and the upcoming standard on spatial audio object coding. He also actively participates in the ISO/IEC IT security techniques standardization committee. He published more than 50 papers at international conferences and journals and coauthored the book, *Spatial Audio Processing: MPEG Surround and Other Applications* (Wiley, 2007).

●

Fabian Nater was born in Switzerland in 1981. He received an M.Sc. degree in electrical engineering from the Swiss Federal Institute of Technology in Lausanne in 2006. He did his graduation project at Philips Research in Eindhoven, The Netherlands, where he investigated perceptual HRTF coding.

After two industrial affiliations he returned to academia and is currently with the Computer Vision Lab at the Eidgenössische Technische Hochschule Zurich, where he is a research assistant and Ph.D. student investigating human motion analysis and abnormal event detection in video.

●

Armin Kohlrausch studied physics at the University of Göttingen, Germany, specializing in acoustics. He received a Master's degree in 1980 and a Ph.D. degree in 1984, both on perceptual aspects of sound.

From 1985 until 1990 he worked at the Drittes Physikalisches Institut, University of Göttingen, being responsible for research and teaching in the fields of psychoacoustics and room acoustics. In 1991 he joined Philips Research Laboratories in Eindhoven, The Netherlands, and worked in the speech and hearing group of the Institute for Perception Research (IPO), a joint venture between Philips and the Eindhoven University of Technology (TU/e). Since 1998 he has been combining his work at Philips Research Laboratories with a professor position for multisensory perception at TU/e. In 2004 he was appointed research fellow of Philips Research.

Dr. Kohlrausch is a member of a great number of scientific societies, both in Europe and the United States. He is a fellow of the Acoustical Society of America and has served as associate editor for the *Journal of the Acoustical Society of America* and the *Journal of the Association for Research in Otolaryngology*. His main scientific interest is the experimental study and modeling of auditory and multisensory perception in humans and the transfer of this knowledge to industrial applications.