Smoothing head-related transfer functions for a virtual artificial head

E. Rasumow[a], M. Blau[a], M. Hansen[a], S. Doclo[b], S. Van De Par[b], D. Püschel[c] and V. Mellert[b]

[a]Institut für Hörtechnik und Audiologie, Jade Hochschule, Ofener Straße 16/19, 26121 Oldenburg, Germany
[b]Carl von Ossietzky Universität Oldenburg, Ammerländer Heerstraße 114-118, 26121 Oldenburg, Germany
[c]Akustik Technologie Göttingen, Bunsenstraße 9c, 37073 Göttingen, Germany
eugen.rasumow@jade-hs.de

The frequency-dependent directivity pattern of a human head (head-related transfer functions, HRTFs) can be synthesized with a microphone array and digital filtering (Rasumow et al. 2011), which may be referred to as a virtual artificial head (VAH). However, in order to synthesize the peaky directivity patterns (especially at high frequencies) with a sufficiently high accuracy, a large number of microphones is required, resulting in an increased sensitivity to gain, phase and positioning errors. Therefore, it is beneficial to smooth the HRTFs in such a manner that they become (spatially) as smooth as possible while still yielding an unmodified individual perception. Since peaky directional characteristics of the HRTF are reflected in the frequency response, smoothing is conducted firstly by truncating the head related impulse responses (hrirs) in the time domain and secondly by smoothing the transfer functions in the frequency domain. The aim of this study is to investigate the limits of these smoothing operations in terms of the truncation length, the relative bandwidth and phase approximations, respectively. Using 3-AFC listening tests, the limiting parameters were determined as a function of the angle of incidence in the horizontal plane. The conducted experiments indicate that a truncation of hrirs from 512 to ≈250 samples and a phase linearization for frequencies $f \gtrsim 1700$ Hz yielded a discrimination ability approximately at 50% correct score.

1 Introduction

The usual way to incorporate (subjective) binaural spatial information into recordings is to use the so-called artificial heads, which are a replica of real human heads. A comprehensive summary of previous binaural recording techniques can be found in [10]. Alternatively, the desired frequency-dependent beam patterns of human head related transfer functions (HRTFs) can also be approximated using a set of spatially distributed microphones with appropriate filters (cf. [3]). Such a setup may be referred to as a virtual artificial head (VAH). The performance of the VAH depends (inter alia) on the microphone topology, the number of microphones and the applied cost function (cf. [3]) but also on the desired directivity pattern.

In general, a microphone array with a fixed number of microphones achieves a more accurate fit to the desired directivity if the latter is spatially smooth. One indirect way of obtaining spatially smooth HRTFs is to smooth them individually in the frequency domain (or equivalently, truncate the corresponding HRIRs). This is however complicated by the fact that at high frequencies (above ≈ 5 kHz), the phase of measured HRTFs often exhibits a too steep frequency response.

On the other hand, it is known that the exact phase spectrum of the HRTFs, at least at high frequencies, is not needed in order to recreate a perfect virtualization [1, 2]. One can therefore take advantage of this insensitivity to the exact phase spectrum at high frequencies by substituting the original phase by something better suited to obtain spatially smooth HRTFs, e.g. a linear phase.

Therefore, the following experiments explored a listeners ability to discriminate between (individual) reference-HRTFs and HRTFs containing reduced information. The experiments were supposed to answer the following questions:

1. What ist the appropriate length for the direction-dependent head related impulse responses (hrir) that leads to an unaltered perception compared to a chosen reference setting?

2. What is a reasonable crossover frequency at which the individual phase can be faded into a linear phase for higher frequencies?

2 Test preparation

2.1 HRTF measurement

At the beginning of each test session, individual HRTFs were measured in an anechoic room with a circular loudspeaker array (radius 1.1 m, custom-made loudspeaker boxes).

The HRTF measurements were carried out with one loudspeaker signal and two ear signals using the blocked ear method (cf. [9]) with custom made ear shells and Knowles FG-23329 miniature electret microphones. The transfer functions were estimated using white noise test signals and standard FFT-based techniques (H_1 estimate with 8192-point Hann window, 50% overlap, 52 averages). The measured transfer functions were divided by the free field transfer function obtained from a measurement (G.R.A.S. Microphone Type 40AF) at the position of the centre of the head without the acoustic influence of the subject.

In order to provide a rough overview of a potential direction-dependency of the tested variables and to limit the number of experiments to a manageable amount, three different source positions (out of the 24 measured) in the horizontal plane with azimuth angles 0° (front), 90° (left) and 225° (back right) were chosen.

2.2 Reference setting

In order to obtain perceptive limits for reducing information from individual impulse responses it is crucial to define a reference setting that represents the basis for comparison within the discrimination experiments. Ideally, this reference setting should contain all desired aspects/information regarding the individual HRTF, whereas all disturbances (e.g. measurement noise, reflections due to the measuring apparatus etc.) should be omitted to avoid potential misleading cues in the evaluation. The available measuring apparatus (especially the loudspeakers, microphones and anechoic room) enables a signal-to-noise ratio (SNR) of about 50-55 dB. On a logarithmic view of the hrir as a function of time (not shown), the decay reaches this noise floor after approximately 400 − 500 samples, corresponding to approximately 9−11 milliseconds. It is partly for this reason that the hrir-length l_{hrir} in the reference-setting was chosen as $l_{hrir} = 2^9 = 512$ samples for all subsequent experiments. As a first control, white noise stimuli filtered with individual HRTF-filters with $l_{hrir} = 512$ from 24 directions (equidistant 15°-spacing) in the horizontal plane were played dichotically to each subject. All of the

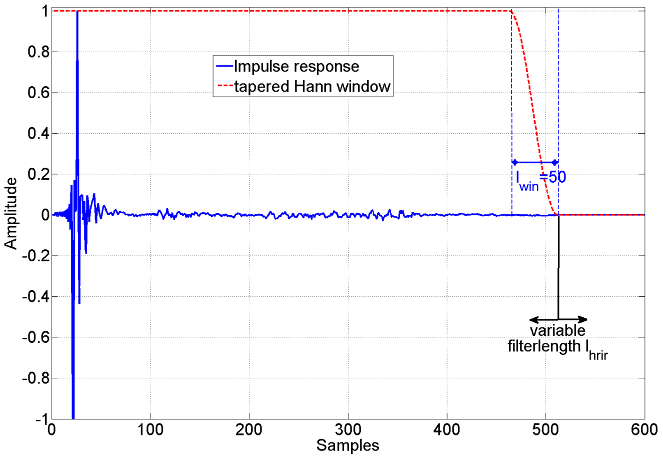

Figure 1: Example hrir (left ear, azimuth=90°) with the reference-length l_{hrir} = 512 samples and a tapered Hann-window with the fixed length l_{win} = 50 samples for the descending flank.

tested subjects could assign the corresponding various source positions and perceive the presented signals outside the head.

All of the used hrirs were flanked with a tapered Hann-window with a descending flank of l_{win} = 50 samples. An exemplary hrir (blue line) and its associated window (red dashed line) with the previously explained parameters are shown in figure 1.

3 Listening tests - material and methods

3.1 Generic Procedure and Stimuli

In order to cover a wide frequency range and include temporal cues, the test signal contained short bursts of white noise with a spectral content of 350 Hz ≤ f ≤ 18000 Hz. Each signal was constructed out of three segments, each with a noise burst of 0.15 seconds with 0.01 seconds raised-cosine onset- and offset ramps, followed by silence of 0.15 seconds, resulting in a total signal length of 0.75 secnds. A three-alternative forced-choice paradigm was employed to determine threshold values for the tested variables. Three intervals of the filtered signals were presented to the subjects in ranomized order each separated by 0.3 seconds silence. One out of the three signals was filtered with the modified HRTF and two with the reference HRTF. The subjects were instructed to indicate the odd of the three intervals. Feedback for the correct answer was presented after each trial. The variable of the test signal was modified according to the weighted up-down method as to estimate the 80%-, as well as the 40%-correct score. These particular values of the psychometric function were targeted by adaptively modifying the step size according to [8]. The initial variable was a filter length of l_{hrir} = 256 for the truncating experiments (see section 3.2) and a crossover frequency of f_g = 3000 Hz for the phase experiments (see section 3.3). The variable was varied adaptively in the familiarisation phase until a minimal step of 3 samples (truncating experiments) and 10 Hz (phase experiments) was reached. As soon as this minimal step was reached, the measurement phase started and the variable was not reduced any more. The threshold of the examined variable was defined as the median of 6 following reversals whithin the measurement phase.

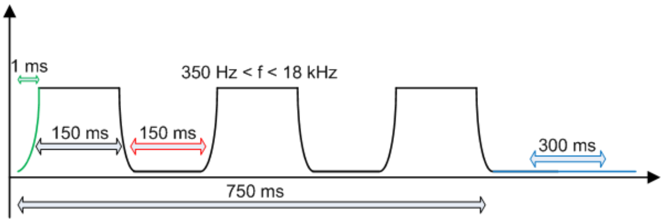

Figure 2: Sketch of the envelope of the test signal as a function of time. Each noise burst (350Hz ≤ f ≤ 18000Hz) lasted 0.15 seconds and was windowed with 0.01 seconds onset-offset ramps. The pauses inbetween the noise bursts lasted 0.15 seconds as well.

Four male and experienced subjects participated in these experiments in which each combination of experimental variable, desired correct score and subject was repeated at least three times.

All experiments were designed using *psylab*[1] (cf. [7]), which is a set of MATLAB–scripts for the design of various psychoacoustical detection- and discrimination-experiments. The signals were generated digitally and presented via a RME ADI-8 DA-converter and AKG K 240 Studio headphones (individually equalized) at an overall sound pressure level of 78 dB SPL to the subject, seated in an anechoic room. The system was calibrated using the Artificial Ear Type 43AA (G.R.A.S.) with the preamplifier Type 27A (G.R.A.S.).

3.2 Truncation experiments

In these experiments, the length of the impulse response (l_{hrir}) was varied adaptively, whereas in each run, the target correct score (either 40% or 80%) and the source direction (one of 0°,90° and 225°) were kept constant. It should be noted that the length of the fading window was set to l_{win} = 50 samples for all various l_{hrir} (see figure 1). In the case that a particular length of the impulse response l_{hrir} was shorter than the window-length l_{win}, the whole impulse response was windowed.

3.3 Phase experiments

While the parameters remained unchanged, the variable for the phase experiments was the crossover frequency f_g. Here, f_g specifies the frequency above which the original phase is replaced by a linear phase. The linear phase was estimated for each combination of parameters and both ears, by evaluating the delay of the maximum of the envelope of the hrirs. This resulting linear phase was subtracted from the original phase to obtain a temporary phase (temp-phase, solid brown line in figure 3). In order to avoid abrupt changes, a descending Hann-flank with the length of 5 samples (corresponding to a frequency range of ≈ 430 Hz) was applied to the temp-phase (dashed green line in figure 3). Thus, the test-phase (solid red line in figure 3) which is the sum of the windowed- and the linear phase does not contain any individual phase -information but only information regarding the overall delay for frequencies $f \geq f_g$.

[1] *Psylab* is available at http://www.hoertechnik-audiologie.de/psylab/

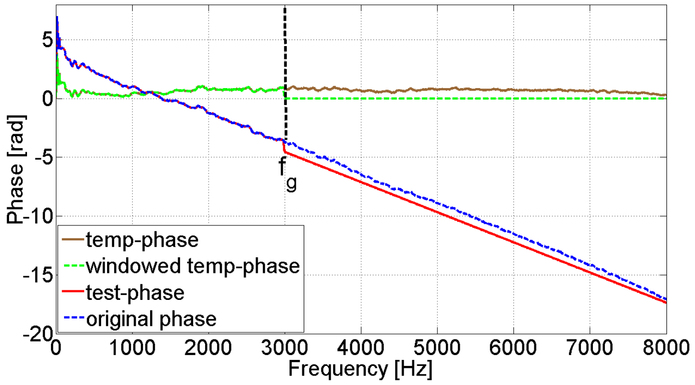

Figure 3: Exemple of the (original) unwraped HRTF-phase, the temp-phase (original-phase - linear-phase), the windowed temp-phase and the test-phase with f_g = 3 kHz (windowed temp-phase + linear-phase).

3.3.1 Preliminary investigation

A preliminary investigation was conducted to examine the general sensitivity of listeners when the individual phase is faded into a broadband linear- or minimum-phase. Although these results will not be discussed here in detail, it is interesting to mention the observed tendencies. Broadband-changes of the actual measured phase into linear, as well as minimum-phase seem to alter the perception noticeably. These conclusions can also be drawn from the results presented in [1] (see figure 8 and 12).

By evaluating a reasonable crossover frequency f_g, not only the phase of HRTFs could be simplified appropriately, but this also would constitute a significant preliminary stage for a perceptively appropriate smoothing of complex-valued HRTFs (cf. for instance [4] and [6]).

4 Listening tests - results

Equation (1) describes the model psychometric function $\Psi(x)$ which is controlled by the two parameters x_m and m with the following meaning: x_m is the midpoint of the base function $\Psi_0(x)$, i.e. $\Psi_0(x_m) = \frac{1}{2}$, and m is the slope of $\Psi_0(x)$ at x_m, i.e. $\Psi_0{}'(x_m) = m$. The value $\frac{1}{3} = p_{\text{guess}}$ results from the 3-AFC method.

$$\Psi_0(x) = \frac{1}{1 + e^{-4m(x-x_m)}}, \qquad \Psi(x) = \frac{1}{3} + \frac{2}{3} \cdot \Psi_0(x) \qquad (1)$$

The experiments yielded the 40% and 80% correct points, x_{40} and x_{80}, on the psychometric function for each condition. The two parameters x_m and m could therefore be expressed in a closed solution as a function of x_{40} and x_{80}.

$$x_m = \left(-\frac{ln\,(3/7)}{ln(9)} \cdot x_{40} + x_{80}\right) \cdot \left(\frac{1}{1 - \dfrac{ln\,(3/7)}{ln(9)}}\right) \qquad (2)$$

$$m = \frac{ln(9)}{4 \cdot (-x_{40} + x_m)} \qquad (3)$$

Subsequently, also the 50% threshold x_{50}, defined by $\Psi(x_{50}) = 50\%$, could then be computed. This was deemed particularly interesting since it corresponded to the JND of the investigated variable.

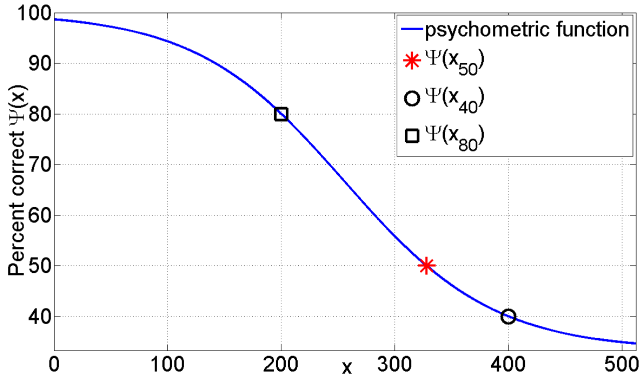

Figure 4: Exemplary psychometric function $\Psi(x)$ estimated from two supposed arguments x_{80} = 200 and x_{40} = 400. According to equation 1 and 2, this yields an argument at threshold ($\Psi(x_{50})$ = 0.5) of $x_{50} \approx 328$.

4.1 Truncation experiments

The results from the truncation experiments are shown in figure 5. The mean values and standard deviation of the 80% scores and 40% scores, $l_{hrir,80}$ (green symbols) and $l_{hrir,40}$ (blue symbols), and the interpolated 50% threshold ($l_{hrir,50}$, red symbols) is indicated by individual symbols for each subject. The inter-subject mean value of $l_{hrir,50}$ is shown with a red dashed line. The results are separated into three panels according to the source direction (0°, 90°, and 225°).

It is evident for the left and the middle panel that the standard deviation for the 80% threshold is smaller than for the 40% threshold. This phenomenon also is apparent for all conditions whithin the phase experiments (cf. figure 6). One possible explanation may be the fact that the acoustical cue to be distinguished is more present for the shorter l_{hrir} at 80% score (cf. figure 4). A 40% score on the other hand corresponds to rather long l_{hrir} and consequently less audible cues which may introduce a greater variance in the individual judgement process. Furthermore, also a large standard deviation of the 80% score is observed for the source direction 225° for three of the subjects. Although each subject received a training phase of at least one complete run, three of four subjects might have altered their judgement successively (not shown). This may suggest that some subjects experienced a learning effect or potentially evaluated different acoustical cues in successive runs. Due to the small population and a deviating focus, this aspect will not be investigated further in the present study.

The inter-subject mean $l_{hrir,50}$ at threshold for the source direction 0° was 184 samples, meaning that for the mean performance the impulse response from 0° could be truncated to this length, in order to yield a 50% score for discrimination relative to the reference setting (cf. 2.2). For the most sensitive subject, this threshold was at $l_{hrir,50}$ = 235 samples. For the 90° source direction the mean $l_{hrir,50}$ at threshold increased to 224 samples and to 259 samples for the most sensitive subject. The longest $l_{hrir,50}$ could be observed for source direction 225°. The mean $l_{hrir,50}$ at threshold for this condition was 250 samples, with the most sensitive subject yielding 320 samples.

To summarise, the $l_{hrir,50}$ at threshold seems to depend on the subject, the source direction and in some cases on a potential learning effect (source direction of 225°). For source directions of 0° and 90°, a truncation of the hrirs from

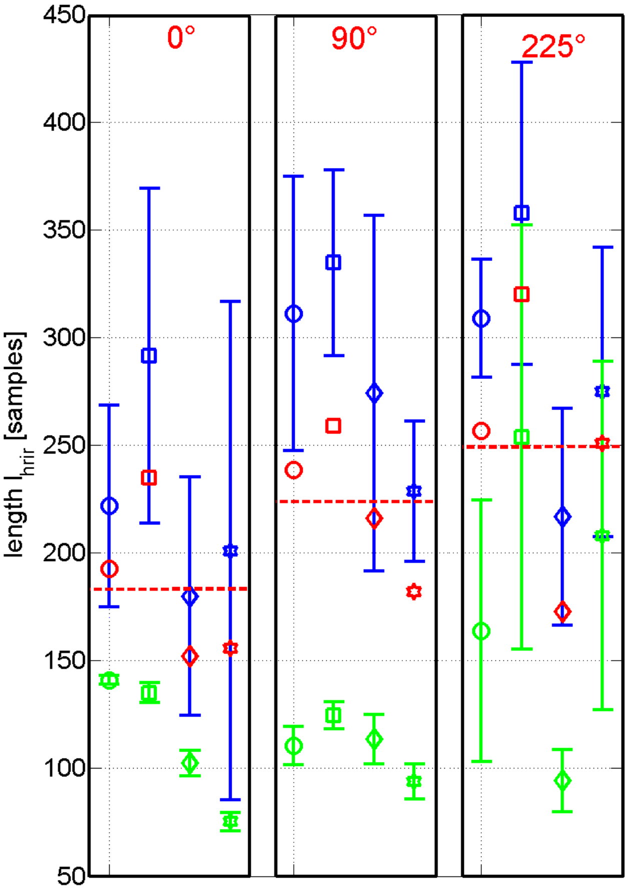

Figure 5: Means and standard deviations of thresholds in the truncation experiment for 80% score (green) and 40% score (blue) for four subject (○, □, ◇ and ✡). Red symbols indicate the calculated 50% scores and the dashed lines indicate their mean across subjects.

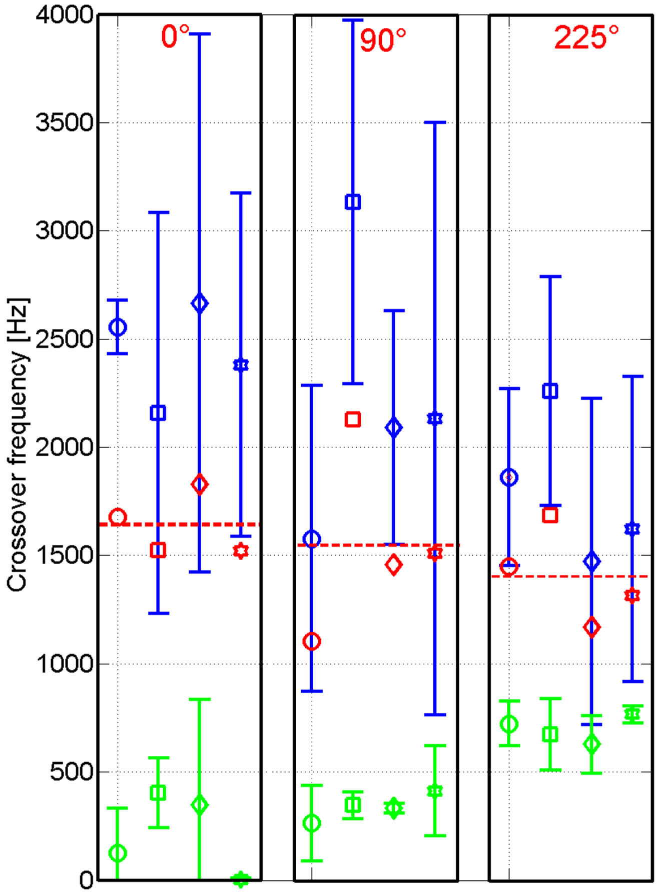

Figure 6: Means and standard deviations of thresholds in the phase experiment for 80% score (green) and 40% score (blue) for four subject (○, □, ◇ and ✡). Red symbols indicate the calculated 50% scores and the dashed lines indicate their mean across subjects.

l_{hrir} = 512 samples to $l_{hrir} \approx 250$ samples (corresponding to ≈ 5.7 ms) seems to retain all essential information, since the last part of the hrirs does not seem to alter the discrimination noticeably. For the source direction of 225°, it appears that $l_{hrir,50}$ increases. Especially with subjects who repeated the experiments several times (more than 3 runs plus training phase) the $l_{hrir,50}$ at threshold seems to increase steadily. However, one must bear in mind that the discrimination in the conducted experiments is based on all feasible aspects of the hrirs (as opposed to e.g. localisation only). Since the learning effect may also be due to a reflection or noise in the HRTF (generally HRTFs from 225° exhibit the worst sound-to-noise ratio (SNR) of the tested source directions), this effect requires further investigation with a larger population and finer directional grid.

4.2 Phase experiments

The results from the phase experiments are illustrated in figure 6 in the same arrangement as in figure 5. Here the standard deviation of the 80% scores ($f_{g,80}$) again is smaller compared to the standard deviation of the 40% scores ($f_{g,40}$), except for subject ◇ and source direction 0°. As before, this phenomenon can be explained by the less audible cues at higher crossover frequencies f_g, which rather correspond to crossover frequencies at 40% score. Moreover, it can be noted that the mean crossover frequencies $f_{g,40}$ and $f_{g,80}$ are closer together for source directions of 90° and 225°, corresponding to a steaper psychometric function compared to the source direction of 0°. Whereas the mean crossover frequency at threshold $f_{g,50}$ seems to be slightly lower for source directions of 90° and 225° respectively, the individual f_g seem to show a different, partly reverse trend (cf. subject □).

For the source direction of 0° the mean crossover frequency at threshold is $f_{g,50}$ = 1638 Hz with $f_{g,50}$ = 1828 Hz for subject ◇, being the most sensitive. The mean crossover frequency at threshold deminishes to $f_{g,50}$ = 1549 Hz and $f_{g,50}$ = 1404 Hz with the most sensitive crossover ferquencies of $f_{g,50}$ = 2127 Hz and $f_{g,50}$ = 1686 Hz (subject □) for source directions of 90° and 225°, respectively.

In sum, the crossover frequency at threshold $f_{g,50}$ does not seem to alter characteristically with the tested source directions. With regard to the mean performance of the tested subjects and conditions, the phase responce above f_g = 1638 Hz can be replaced by a linear phase causing no characteristic alteration of the perception. In order to account for the most sensitive case of the tested conditions, the measured phase response should be linearized not below f_g = 2127 Hz. These results are in good agreement with former investigations, where [1] for instance supposed a $f_g \approx 2000$ Hz, and with experiments on phase locking, approximately assumed at $f_g \approx 1500$ Hz.

5 Impact of information reduction on beamformer performance

The results from the truncation experiments reveal the opportunity to truncate the hrirs and thus to yield a smoother, but psychoacoustically equivalent frequency response. Regarding the VAH, this raises the question whether the smoother frequency responses enable a more precise fit of the beamformer to the desired HRTFs. In order to compare the approximations of the VAH with respect to l_{hrir}, the mean absolute fitting error $\xi(l_{hrir})$ is introduced as

$$\xi(l_{hrir}) = \frac{1}{N \cdot M} \sum_{\phi=1}^{N} \sum_{\omega=1}^{M} \left| 20 \cdot \log_{10} \left(\left| \frac{D(\phi, \omega, l_{hrir})}{H(\phi, \omega, l_{hrir})} \right| \right) \right|, \quad (4)$$

with the resulting frequency response of the VAH H and the desired HRTF D, both in the frequency domain. The frequency range for analysis was set to $0\text{ Hz} \leq \frac{\omega}{2\pi} \leq 16000\text{ Hz}$ and the source direction varied in 15°-steps from 0° to 345°. The length of the impulse responses was either $l_{hrir} = 512$ samples or $l_{hrir} = 300$ samples, assuming that the finding from section 4.1 approximately holds for all directions in the horizontal plane. These hrirs were applied to the VAH with the same settings as in [3] (Golomb-topology, 24 microphones, no artificial positioning error).

It appears from the simulations that the advantage of truncating hrirs (i.e. $\Delta\xi = \xi(512) - \xi(300)$) depends on the regularisation of the VAH. Regularisation was applied by imposing a constraint ($\beta = \underline{w}^H \underline{w}$) on the calculation of the individual filterweights (see equation 6 and 7 in [3]). Hardly no advantage is achieved if no regularisation is applied, whereas the improvement of the VAH still stays small for reasonable white noise gains (β). The mean fitting error $\xi(512) = 0.83$ dB decreases to $\xi(300) = 0.81$ dB with $\beta = 0$ dB or from $\xi(512) = 0.47$ dB to $\xi(300) = 0.46$ dB with $\beta = 5$ dB, for instance. This analysis implies that truncating the impulse responses in reasonable dimensions causes only very small analytical improvements.

This effect is even more prominent when varying the phase according to the observed results from paragraph 4.2 (with a conservatively chosen $f_g = 2500$ Hz), resulting in $\Delta\xi \approx 0$. Although the linearization of the phase shows no usable effect on the performance of the VAH, it is an essential step for smoothing complex HRTFs since this phase-manipulation may enable a reaonable smoothing of complex-valued spectral responses (see [4]).

6 Summary and further steps

The present article deals with the ability to distinguish between truncated hrirs and linearized hrir-phase responses. It was found that on average the tested subjects could distinguish between hrirs with $l_{hrir} = 512$ and $l_{hrir} \approx 250$ at a 50% correct score. For some hrirs (source direction of 225°) there appears to be a learning effect in recognizing truncated hrirs which needs further investigations. Furthermore it was found that on average the phase response above $f_g = 1638$ Hz can be linearized yielding a discrimination at 50% correct score.

It should be noted that all the reviewed aspects only depict tendencies based on experiments for three source directions. Further investigations with more subjects and source directions are needed in order to provide general statements. It could be shown that neither the reduction of information in the time-domain of hrirs, nor the simplification of the phase response for higher frequencies of HRTFs revealed a considerable enhancement regarding the application to a VAH. For this reason further steps may be for instance a psychoacoustically reasonable smoothing using constant relative bandwidths. Moreover, a major advantage regarding the VAH may be a reasonable spatial smoothing of the desired beam patterns using the decomposition in spherical/cylindrical harmonics with a variable order.

Acknowledgments

This project was partially funded by Bundesministerium für Bildung und Forschung under grant no. 17080X10.

References

[1] Kulkarni, A., Isabelle, S.K., Colburn, H.S., Sensitivity of human subjects to head-related transfer-function phase spectra, J. Acoust. Soc. Am. 105 (5), May 1999

[2] Breebaart, J., Nater, F., Kohlrausch, A., 2010, Spectral and Spatial Parameter Resolution Requirements for Parametric, Filter-Bank-Based HRTF Processing, J. Audio Eng. Soc., Vol. 58, No. 3, 2010 March

[3] Rasumow, E., Blau, M., Doclo, S., Püschel, D., van de Par, S., Hansen, M. and Mellert, V., 2011, Robustness of virtual artifcial head topologies with respect to microphone positioning, Forum Acousticum 2011. Aalborg, Denmark

[4] Rasumow, E., Blau, M., Doclo, S., Hansen, M., Mellert, V., Püschel, D. and van de Par, S.; 2012, Subjektiver Höreindruck bei synthetischer Erzeugung von kopfbezogenen Übertragungsfunktionen, DAGA-2012, Darmstadt

[5] Mellert, V. and Tohtuyeva, N., 1997. Multimicrophone arrangements as suitable for dummyhead recording technique. In Proc. 137th ASA Meeting. S. 3117

[6] Hatziantoniou, D. P. and Mourjopoulos, J. N., 2000, Generalized Fractional-Octave Smoothing of Audio and Acoustic Responses, J. Audio Eng. Soc., Vol. 48, No.4, April 2000

[7] Hansen, M., 2006, Lehre und Ausbildung in Psychoakustik mit psylab: Freie Software für psychoakustische Experimente, Fortschritte der Akustik – DAGA '06, Braunschweig

[8] Christian Kaernbach, 1991, Simple adaptive testing with the weighted up-down method, Perception & Psychophysics 1991, 49 W. 227-229

[9] Hammershoi, D., Moller, H., 1996, Sound transmission to and within the human ear canal, J. Acoust. Soc. Am. 100 (1), July 1996

[10] Stephan Paul, Binaural Recording Technology: A Historical Review and Possible Future Developments, Acta Acustica united with Acustica, Vol. 95 (2009) 767 - 788