Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3-5 April 2014

THE IMPACT OF THE WHITE NOISE GAIN (WNG) OF A VIRTUAL ARTIFICIAL HEAD ON THE APPRAISAL OF BINAURAL SOUND REPRODUCTION

Eugen Rasumow, Matthias Blau, Martin Hansen, *

Institute of hearing technology and audiology
Jade University of Applied Sciences
Oldenburg, Germany
eugen.rasumow@jade-hs.de

Simon Doclo, Steven van de Par, Volker Mellert

Institute of Physics
Carl-von-Ossietzky University
Oldenburg, Germany

Dirk Püschel

Akustik Technologie Göttingen
Göttingen, Germany

ABSTRACT

As an individualized alternative to traditional artificial heads, individual head-related transfer functions (HRTFs) can be synthesized with a microphone array and digital filtering. This strategy is referred to as "virtual artificial head" (VAH). The VAH filter coefficients are calculated by incorporating regularization to account for small errors in the characteristics and/or the position of the microphones. A common way to increase robustness is to impose a so-called white noise gain (WNG) constraint. The higher the WNG, the more robust the HRTF synthesis will be. On the other hand, this comes at the cost of decreasing the synthesis accuracy for the given sample of the HRTF set in question. Thus, a compromise between robustness and accuracy must be found, which furthermore depends on the used setup (sensor noise, mechanical stability etc.). In this study, different WNG are evaluated perceptually by four expert listeners for two different microphone arrays. The aim of the study is to find microphone array-dependent WNG regions which result in appropriate perceptual performances. It turns out that the perceptually optimal WNG varies with the microphone array, depending on the sensor noise and mechanical stability but also on the individual HRTFs and preferences. These results may be used to optimize VAH regularization strategies with respect to microphone characteristics, in particular self noise and stability.

1. INTRODUCTION

In order to take into account spatial cues within a binaural reproduction, the use of so-called artificial heads, which are a replica of real human heads and pinnae, is common practice today. By this means the signals at the ears receive characteristic spatial information, which encompasses interaural time and level difference cues, but also spectral cues due to the shape of the pinna, for instance. Disadvantageously, artificial heads are inherently bound to non-individual (average) anthropometric geometries and are most often implemented as bulky devices. Alternatively, the individual frequency-dependent directivity patterns of a human head (HRTFs) can be synthesized with a microphone array and digital filtering (cf, [1], [2], [3], [4] and [5]), which will be referred to as a virtual artificial head (VAH). A VAH is more flexible than real artificial heads, since, e.g., the filters can be adjusted post-hoc to match any individual sets of HRTFs. In contrast to approaches in the spherical harmonics domain (i.e. applying spherical harmonics decomposition, optimization and re-synthesis, cf. [3] and [6]), the VAH re-synthesis in this study is optimized in the frequency domain for discrete directions in the horizontal plane only, assuming the intermediate directions to be inherently interpolated by the VAH. One advantage of this approach is that much fewer microphones are needed in comparison to e.g. spherical harmonics based approaches (cf. [7] and [8]). The individual filter coefficients can be calculated by optimizing various cost functions, where a least square cost function is known to yield appropriate perceptual results (cf. [5]) and is thus used in this study (cf. section 2). The robustness of the filter coefficients is usually assured by imposing a constraint on the so-called white noise gain (WNG), in order to consider small deviations of the microphone characteristics and/or positions (cf. [4]). By doing so, the robustness of the filter coefficients increases with higher WNG while the accuracy decreases at the same time for a given HRTF set and vice versa (cf. Figure 1). Thus, it seems reasonable to find a compromise in the regularization, where the perceptual appraisal of a HRTF re-synthesis using the VAH is assessed best as a function of the WNG. Two microphone arrays (cf. Figure 2) were applied in this study. These arrays enabled the use of measured steering vectors (as opposed to the application of analytical steering vectors in cf. [3], [4] or [6]) and to re-synthesize individual ear signals by individually recalculating pre-recorded signals.

2. REGULARIZED LEAST SQUARES COST FUNCTION

Consider the desired directivity pattern $\mathrm{D}(\omega, \Theta)$ as a function of frequency ω and discrete azimuthal angles Θ, as well as the $\mathrm{N}\times 1$ steering vector $\mathbf{d}(\omega, \Theta)$ which represent the frequency- and direction-dependent transfer functions between the source and the N microphones. Then the re-synthesized directivity pattern of the VAH $\mathrm{H}(\omega, \Theta)$ for one particular set of steering vectors $\mathbf{d}(\omega, \Theta)$

* Author to whom correspondence should be addressed. Electronic mail: eugen.rasumow@jade-hs.de

can be expressed as[1]

$$\mathrm{H}(\omega,\Theta) = \mathbf{w}^{\mathbf{H}}(\omega)\mathbf{d}(\omega,\Theta). \tag{1}$$

Here, the N×1 vector $\mathbf{w}(\omega)$ contains the complex-valued filter coefficients for each microphone per frequency ω and a given set of steering vectors $\mathbf{d}(\omega,\Theta)$.
In order to calculate the filter coefficients $\mathbf{w}(\omega)$ for the steering vectors $\mathbf{d}(\omega,\Theta)$, one may employ a narrowband least squares cost function $\mathrm{J_{LS}}$, being the sum over P directions of the squared absolute differences between $\mathrm{H}(\omega,\Theta)$ and $\mathrm{D}(\omega,\Theta)$ that is to be minimized, i.e.

$$\mathrm{J_{LS}}(\mathbf{w}(\omega)) = \sum_{\Theta=1}^{\mathrm{P}} \left|\mathbf{w}^{\mathbf{H}}(\omega)\mathbf{d}(\omega,\Theta) - \mathrm{D}(\omega,\Theta)\right|^2 . \tag{2}$$

In this study, filters were optimized to represent individual HRTFs measured in the horizontal plane with an equidistant angular spacing of $\Delta\Theta = 15°$, resulting in P = 24 directions. A straightforward minimization of Eq. 2, however, may result in non robust filter coefficients $\mathbf{w}(\omega)$, where already small errors of the microphone positions and/or characteristics may cause huge errors of the re-synthesized directivity patterns (cf. [4] and [9]) and which may lead to a not desirable amplification of spatially uncorrelated noise at the microphones. More robust filter coefficients can be obtained by imposing a constraint on the derived filter coefficients. To this end, we propose a modified definition of the white noise gain ($\mathrm{WNG_m}$), given as

$$\mathrm{WNG_m}(\omega) = 10\cdot\log_{10}\left(\frac{\mathbf{w}^{\mathbf{H}}(\omega)\mathbf{Q_m}(\omega)\ \mathbf{w}(\omega)}{\mathbf{w}^{\mathbf{H}}(\omega)\mathbb{I}_{\mathrm{N}}\ \mathbf{w}(\omega)}\right), \text{ with}$$

$$\mathbf{Q_m}(\omega) = \frac{1}{\mathrm{P}}\sum_{\Theta=1}^{\mathrm{P}} \mathbf{d}(\omega,\Theta)\mathbf{d}^{\mathbf{H}}(\omega,\Theta) \tag{3}$$

and $\mathbb{I}_{\mathrm{N}}$ being the N×N-dimensional unity matrix. By doing so, $\mathrm{WNG_m}(\omega)$ relates the mean array gain in the measured acoustic field (determined by $\mathbf{Q_m}(\omega)$ and $\mathbf{w}(\omega)$) to the inner product of the filter coefficients, i.e. to the array gain for spatially uncorrelated noise at the microphones (cf. [10]). Usually, regarding beamforming applications the WNG is given for a certain direction (discrete steering direction Θ_0) only (cf. [11],[12] and [5]), whereas the $\mathrm{WNG_m}$ in Eq. 3 may be referred to as the mean WNG over all considered directions Θ. This modification of the WNG was applied since a direction-dependent constraint (as is realized in the classical WNG) would consequently yield a direction-dependent regularization, which is not desirable for a VAH re-synthesis. Hence, the mean $\mathrm{WNG_m}$ incorporating all associated directions is introduced in this study (Eq. 3). Positive $\mathrm{WNG_m}$ represent an attenuation of spatially uncorrelated noise, whereas negative $\mathrm{WNG_m}$ represent an amplification ([11]) relative to the mean array gain in the measured acoustic field. We suggest to apply the constraint $\mathrm{WNG_m}(\omega) \geq \beta$ for regularization, where the gain β (in dB) has to be chosen manually according to the expected error of the steering vectors (cf. [4]). The combination of the least squares cost function from Eq. 2 with the constraint incorporating Eq. 3 results in the cost function

$$\mathrm{J_{LS}}_{\rho}(\mathbf{w}(\omega)) = \sum_{\Theta=1}^{\mathrm{P}} \left|\mathbf{w}^{\mathbf{H}}(\omega)\mathbf{d}(\omega,\Theta) - \mathrm{D}(\omega,\Theta)\right|^2 + \mu\left(\left(\mathbf{w}^{\mathbf{H}}(\omega)\mathbb{I}_{\mathrm{N}}\mathbf{w}(\omega)\right) - \frac{1}{\beta_{\mathrm{pow}}}\left(\mathbf{w}^{\mathbf{H}}(\omega)\mathbf{Q_m}(\omega)\ \mathbf{w}(\omega)\right)\right), \tag{4}$$

where μ represents the Lagrange multiplier and $\beta_{\mathrm{pow}} = 10^{\frac{\beta}{10}}$. The closed form solution of $\mathrm{J_{LS}}_{\rho}(\mathbf{w}(\omega))$, yielding the regularized filter coefficients $\mathbf{w}(\omega)$, is given by

$$\mathbf{w}(\omega) = \left(\mathbf{Q}(\omega) + \mu\left(\mathbb{I}_{\mathrm{N}} - \frac{1}{\beta_{\mathrm{pow}}}\cdot\mathbf{Q_m}(\omega)\right)\right)^{-1}\cdot\mathbf{a}(\omega)\ , \quad \text{with} \tag{5}$$

$$\mathbf{Q}(\omega) = \sum_{\Theta=1}^{\mathrm{P}} \mathbf{d}(\omega,\Theta)\mathbf{d}^{\mathbf{H}}(\omega,\Theta) \ \text{ and} \tag{6}$$

$$\mathbf{a}(\omega) = \sum_{\Theta=1}^{\mathrm{P}} \mathbf{d}(\omega,\Theta)\ \mathrm{D}^*(\omega,\Theta)\ . \tag{7}$$

While the least squares solution of the cost function in Eq. 2 is quite well known in literature (cf. [9], [5]), the regularization term in Eq. 5 differs from usual regularization strategies, as for instance known from diagonal loading (cf. [13]), Tikhonov-regularization or similar regularization approaches (cf. [14]). The main difference lies in the dependence of the regularization on the applied steering vectors ($\mathbf{Q_m}(\omega)$) and the desired $\mathrm{WNG_m}$ β. However, the presented regularization approaches the diagonal loading or Tikhonov-regularization for very large β_{pow} (i.e., for the most stringent regularization possible).
The optimal μ to satisfy the desired WNG-constraint was chosen iteratively. Analogous to the procedure in [5], μ was increased in steps of $\Delta\mu = \frac{1}{100}$ for each ω until $\mathrm{WNG_m}(\omega,\mu) \geq \beta$ or $\mu_{\mathrm{max}} = 100$ were reached (if existent at all, this only occurred at very high frequencies).

2.1. Influence of the WNG-constraint on the VAH re-syntheses

The accuracy of the VAH re-syntheses depends on the desired HRTFs, the number of microphones, the topology of the microphone array, the cost function and also the applied Lagrangian

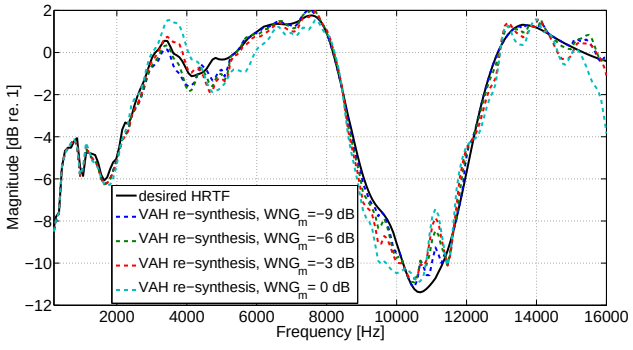

Figure 1: *Magnitude of the desired HRTF ($\Theta = 90°$) for the left ear of subject* $\mathrm{S_1}$ *(black line) and VAH re-syntheses with various* $\mathrm{WNG_m}$ *(dashed lines) for* $\mathrm{array_2}$ *as a function of frequency.*

[1] In the following $x^{\mathbf{H}}$ denotes the Hermitian transpose of x and x^* denotes the complex conjugate of x.

multiplier μ (cf. Eq. 5). In general, the desired $\mathrm{WNG_m}$ is approached by gradually increasing μ. This in turn will cause increasing deviations of the re-syntheses from the desired HRTF. The magnitude of the resulting μ is primarily determined by the desired $\mathrm{WNG_m}$ β. Thus, the regularization yielding a desired $\mathrm{WNG_m}$ unavoidably causes distortions of the VAH re-syntheses which may vary individually with the desired HRTFs and steering vectors. This aspect is exemplarily depicted in Figure 1. On the other hand, higher $\mathrm{WNG_m}$ are associated with more robustness regarding small changes of the microphone characteristics and/or with a lower amplification of spatially uncorrelated noise at the microphones.

3. MICROPHONE ARRAYS USED

The main goal of this study is to investigate the perceptually optimal $\mathrm{WNG_m}$ for different subjects, using different microphone arrays. For this reason, the perceptual evaluation was made with recordings using two open planar microphone arrays incorporating different kinds of microphones and support structures but the same number of microphones and an identical topology which was chosen according to [4]. The advantage of using open planar arrays over rigid spheres or the like is the opportunity to realize various two-dimensional inter-microphone distances. By this means, a mathematically motivated microphone topology according to [4] was chosen, which is assumed to yield appropriate results regarding the accuracy and robustness of the re-syntheses.
The first microphone array (array_1, left panel in Figure 2) consisted of 24 Sennheiser KE 4-211-2 microphones. The individual microphones were mounted on a wooden plate using a solid wire construction. Together with analog preamplifiers the sensor noise of each single microphone signal was approximately 35 dB(A). No absorbent material was used for the support structure of array_1.

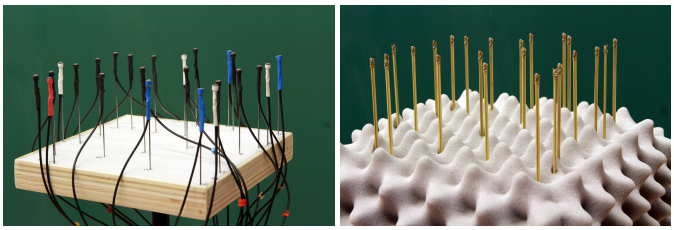

Figure 2: *Two used microphone arrays with 24 KE-4 microphones (*array_1*, left) and 24 sensors composed of 48 MEMS microphones (*array_2*, right) with the same planar microphone topology according to [4].*

For the second array (array_2), micro-electromechanical system (MEMS) microphones (Analog Devices ADMP 504 Ultralow Noise Microphone) were used in an custom-made electrical circuit. Here, each sensor is composed of two MEMS microphones. A composed sensor yielded a sensor noise of approximately 27 dB(A), which is quite low for this kind of microphones. The directivity of such a composed sensor can be assumed to be negligible for frequencies of interest (i.e. $f \lesssim 16$ kHz). For array_2, 24 of these sensors (consisting of 48 MEMS microphones) were mounted on a printed circuit board (cf. right panel in Figure 2) with the same topology as for array_1. In order to reduce effects of standing waves between the sensors and the board, array_2 is covered with absorbent material.

4. EXPERIMENTAL PROCEDURE

4.1. Material

Prior to the experiment, individual HRTFs and headphone (AKG K-240 Studio) transfer functions (HPTFs) were measured for four subjects using the blocked ear method according to [15]. For measuring the HPTFs, subjects were instructed to reposition the headphone ten times to various realistic carrying positions which successively yielded ten different individual HPTFs. The individual HPTF resulting in the smallest dynamic range of its magnitude for frequencies 300 Hz $\leq f \leq$ 16000 Hz was inverted in the frequency domain and transformed into the time domain. The HRTFs as well as the inverse HPTFs were implemented as finite impulse response (FIR) filters with a filter length of 256 taps, corresponding to ≈ 5.8 ms at a sampling frequency of $f_s = 44100$ Hz. This filter length was chosen to incorporate all aspects associated with an appropriate binaural reproduction (cf. [16]). The individual HRTFs as well as the steering vectors $\mathbf{d}(\omega, \Theta)$ for the two microphone arrays were measured in the horizontal plane with an angular spacing of 15°. All HRTFs were smoothed in the frequency and spatial domain prior to the VAH re-syntheses according to the perceptual limits derived in [17]. Moreover, the associated impulse responses of all measured steering vectors $\mathbf{d}(\omega, \Theta)$ were also truncated to a filter length of 256 taps in order to achieve smoother transfer functions.

4.2. Test stimulus

As to cover a wide frequency range and simultaneously to include temporal cues, the test stimulus for perceptual evaluation consisted of 3 short bursts of pink noise filtered with an eighth order bandpass with the cutoff frequencies of $f_{\mathrm{low}} = 300$ Hz and $f_{\mathrm{hi}} = 16000$ Hz. The lower bandwidth limitation of the test stimulus f_{low} was chosen due to the limits of the loudspeakers used. However, since the influence of varying the $\mathrm{WNG_m}$ is primarily evident for frequencies $f \geq 3$ kHz (cf. Figure 1) it seems reasonable to assume that this limitation does not have a significant influence on the perceptual evaluations. Each noise burst lasted $\frac{1}{3}$ s with 0.01 s onset-offset ramps followed by silence of $\frac{1}{6}$ s. This test stimulus was intended to facilitate the evaluation of spectral deviations, temporal dispersion but also the influence of the sensor noise. The presented stimuli were calibrated with a G.R.A.S. type 43AA artificial ear to have 70 dB SPL for the frontal direction $\Theta = 0^\circ$.

4.3. Methods

A listening test was carried out with four experienced listeners (two of them are authors of this article). The subjects were instructed to rate four different aspects (localization, sensor noise, overall performance and spectral coloration, cf. section 4.3.1) of a test presentation with respect to the reference presentation (binaural reproduction with original individual HRTFs and HPTFs). The quality of the reference setting (representing desirable re-syntheses) has a major effect on the evaluations. Thus it needed to be assured that the individual binaural reproductions incorporated all essential individual spatial characteristics. For this reason, the individual binaural reproductions used in the reference setting were played to the subjects before the experimental procedure in a preliminary listening test. All subjects were able to perceive the presented stimuli outside the head and correctly assigned the corresponding directions in the horizontal plane.

Prior to the listening tests, the steering vectors were measured and the test stimuli were recorded using the two microphone arrays (cf. section 3) in an anechoic chamber. Furthermore, the individual VAH filters were optimized to re-synthesize the individual HRTFs in the horizontal plane with an angular spacing of $\Delta\Theta = 15°$. In the test condition, the sum of the filtered stimuli (representing the re-synthesized ear signals, cf. Eq.1) was also filtered with the inverse HPTF filters (same procedure as in the reference setting) and played to the subject via headphones. In both conditions, the stimuli were played back in an infinite loop with the possibility to switch between the reference- and test condition or to stop the playback. To limit the number of experiments to a manageable amount, three directions in the horizontal plane were chosen for evaluation with azimuth angles $\Theta = 0°$ (front), $\Theta = 90°$ (left) and $\Theta = 225°$ (back right) and the $\mathrm{WNG_m}$ was one of $\mathrm{WNG_m}(\omega)$ = -9 dB, -6 dB, -3 dB or 0 dB for all ω. These preselected $\mathrm{WNG_m}$ were assumed to roughly cover the area with the best suited $\mathrm{WNG_m}$ based on previous preliminary tests.

The three tested azimuthal directions Θ, the two microphone arrays as well as the four $\mathrm{WNG_m}$ were varied in randomized order within one experimental run with three random presentations (retest) for each condition. The true identities of the signals in the reference and test setting were hidden to the subjects. In sum, 216 conditions (presented signal pairs) were evaluated by each subject, whereas one of the tested parameters (impact of various calibration strategies) was eliminated from the analysis in this article in hindsight. Hence, 3 directions $\times$ 2 arrays $\times$ 3 presentations $\times$ 4 $\mathrm{WNG_m}$ = 72 individual evaluations (of a total of originally 216 individually gathered evaluations) will be analyzed and discussed in section 5 and 6. Within each condition, subjects were able to switch between the reference and the test setting arbitrarily. The entire experiment was performed applying an English category scale, ranging between *bad*, *poor*, *fair*, *good* and *excellent* with four intermediate undeclared steps (cf. [5]). Each session lasted approximately 120-180 minutes, where subjects were able to subdivide the session arbitrarily and to do as many breaks as they wanted. Prior to the evaluation each subject had time for familiarization with the various reference and test conditions.

4.3.1. Assessed aspects

The subjects were instructed to evaluate the quality of the test setting with respect to reference setting for four chosen aspects which are assumed to be significant for appropriate VAH re-syntheses:

- localization: The evaluation of localization incorporated the perceived angle of incidence (azimuth and elevation) and the perceived distance in combination.
- sensor noise: Subjects were instructed to evaluate the perceived sensor noise which was primarily apparent in the temporal pauses of the test stimulus.
- overall performance: The evaluation of the perceived overall performance incorporated all feasible aspects depending on the taste and preferences of the individual subject.
- spectral coloration: Subjects were instructed to evaluate the perceived spectral coloration without evaluating the potential deviations of localization or other cues.

5. RESULTS AND DISCUSSION - PERCEPTUAL EVALUATION

The mean and the standard deviations (over three randomized presentations) of all individual evaluations are depicted in Figure 3 as functions of the $\mathrm{WNG_m}$ on the x-axis with the assessed aspects separated in rows, the directions Θ separated in columns and the color indicating the subjects. The average performance (means and standard deviations over subject) is depicted in Figure 4, with the color indicating the assessed aspects (see legend).
In general, the perceptual evaluations and their variation within repeated trials in Figure 3 (standard deviation depicted as error bars) seem to depend on the direction of incidence Θ and the used microphone array, but as well on the subject. This is an effect of individual preferences with individual internal scales and was to be expected according to analogous studies (cf. [5]). In order to analyze potential preferences regarding the $\mathrm{WNG_m}$ for the application of a VAH, primarily the relative tendencies of intra- and inter-individual perceptual evaluations depending on the $\mathrm{WNG_m}$ are focused on.

Table 1: p-values (rounded to 3 digits) according to the Friedman test regarding localization, overall performance, sensor noise and coloration for the three tested directions separately. p-values indicating significantly different evaluations when varying the $\mathrm{WNG_m}$ ($p \leq \frac{0.05}{24} = 0.0021$) are depicted as bold numbers.

localization	array_1	array_2	overall	array_1	array_2
$\Theta = 0°$	0.164	0.445	$\Theta = 0°$	0.341	0.081
$\Theta = 90°$	0.004	0.006	$\Theta = 90°$	**0.000**	0.129
$\Theta = 225°$	0.147	0.933	$\Theta = 225°$	0.109	0.188
sensor noise	array_1	array_2	**coloration**	array_1	array_2
$\Theta = 0°$	0.004	0.049	$\Theta = 0°$	0.035	0.578
$\Theta = 90°$	**0.000**	0.340	$\Theta = 90°$	**0.000**	0.827
$\Theta = 225°$	**0.000**	0.079	$\Theta = 225°$	0.015	0.319

Although means and standard deviations were used for illustrating the evaluations in Figs. 3 and 4 (for increased clarity), a non parametric statistical test was applied. The Friedman test was applied to analyze whether the evaluations for at least one of the tested $\mathrm{WNG_m}$ (for a fixed direction, array and assessed aspect) was considerably different than the evaluations for the other $\mathrm{WNG_m}$. A sufficiently small p-value indicated an effect of the $\mathrm{WNG_m}$ on the evaluations. The p-values for the assessed aspects (separate boxes), the applied arrays (columns) and directions (rows) are given in Table 1. The p-values for conditions indicating a significant effect of the $\mathrm{WNG_m}$ on the perceptual evaluations (considering the Bonferroni correction for 24 repeated tests, a p-value of $p \leq \frac{0.05}{24}$ is assumed to indicate a significant effect of the $\mathrm{WNG_m}$) are depicted as bold numbers. However, due to the rather small number of subjects and the presumably low test power, the p-values in Table 1 may primarily be used to highlight tendencies of all evaluations for fixed conditions without postulating any statistical (in)significances for the effect of the $\mathrm{WNG_m}$.
In sum, it emerges that the tested $\mathrm{WNG_m}$ mainly seem to have an effect on the evaluations for array_1 with regard to sensor noise and coloration. The evaluations regarding localization seem primarily to be affected by the $\mathrm{WNG_m}$ for $\Theta = 90°$ and both arrays. The evaluations regarding the overall performance seem to be affected by the $\mathrm{WNG_m}$ mainly for array_1 and $\Theta = 90°$.

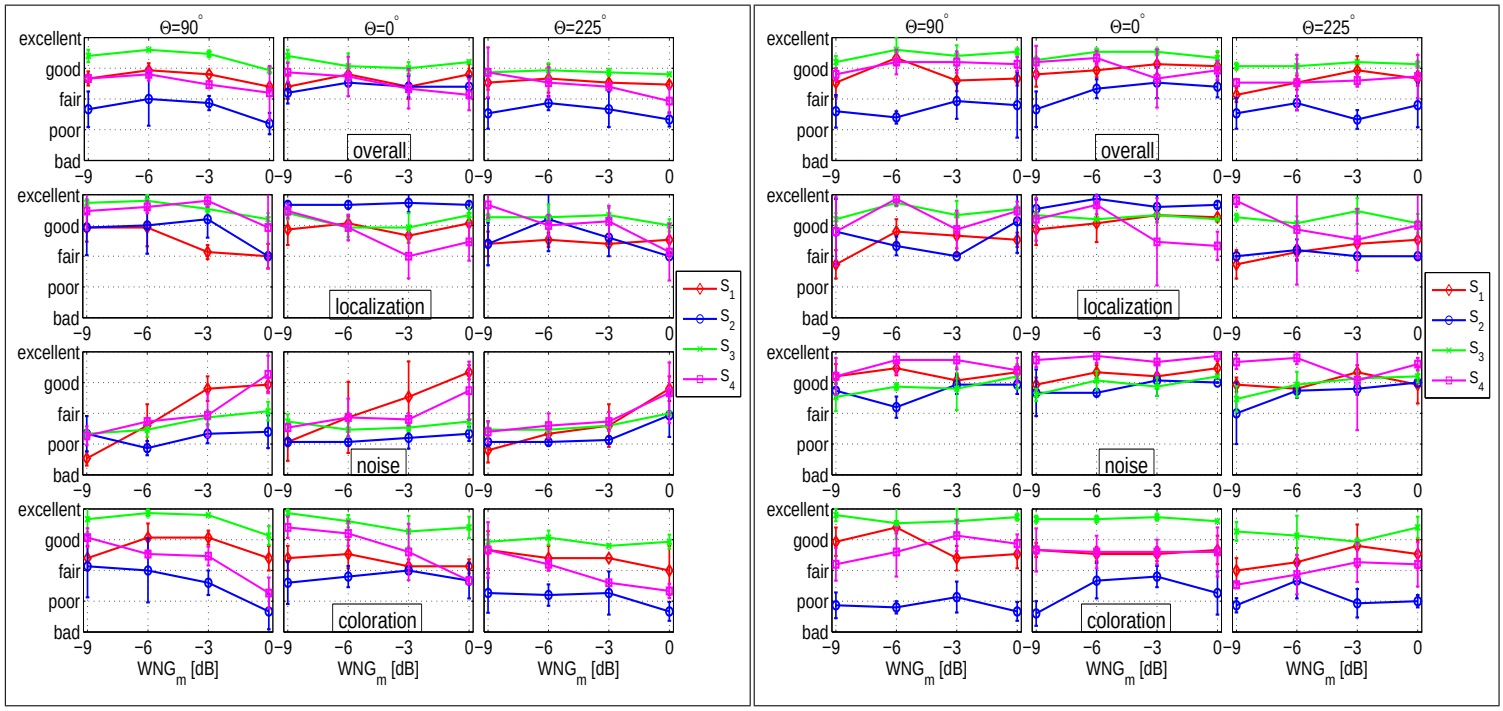

Figure 3: *Perceptual evaluations for* array_1 *(left block) and* array_2 *(right block). The aspects of evaluation are aligned in separate rows (first row: overall performance, second row: localization, third row: sensor noise and fourth row: spectral coloration) and the direction of arrival* Θ *is aligned in three columns (*$\Theta = 90°$ *in the left column,* $\Theta = 0°$ *in the middle column and* $\Theta = 225°$ *in the right column). The individual evaluations (mean and standard deviation over three randomized presentations) are depicted as a function of the* $\mathrm{WNG_m}$ *in dB. The colors and markers indicate the four subjects (*$\mathrm{S_1}$, $\mathrm{S_2}$, $\mathrm{S_3}$ *and* $\mathrm{S_4}$*).*

5.1. Localization

In general, all subjects concordantly reported the localization in the horizontal plane to be re-synthesized well by the VAH. However, the aspect localization was also used to evaluate the perceived distance of the sound source (cf. section 4.3.1). The perception of distance may vary noticeably when interaural level differences from lateral directions are not re-synthesized accurately. This may be a possible explanation for the better evaluations for $\Theta = 0°$, which is especially evident for subject $\mathrm{S_1}$ and $\mathrm{S_2}$ (cf. Figure 3).

For subject $\mathrm{S_3}$, the evaluations with regard to localization vary hardly with the tested $\mathrm{WNG_m}$ nor with the array. The p-values from Table 1 indicate the most notable effect of the $\mathrm{WNG_m}$ on the evaluations with regard to localization for $\Theta = 90°$ with both arrays. This aspect is also apparent in the averaged evaluations (cf. Figure 4) for array_1, where the evaluations decrease for higher $\mathrm{WNG_m}$. However, there does not seem to be such an unambiguous tendency for the evaluations with array_2 and $\Theta = 90°$. Moreover, the averaged evaluations seem also to decrease slightly with increasing $\mathrm{WNG_m}$ for $\Theta = 225°$ and array_1. This slight effect is concordantly associated with a relatively higher p-value from the Friedman test (p=0.147), as well indicating a less notable effect of the tested $\mathrm{WNG_m}$.

In sum, the evaluations of localization seems to decrease with higher $\mathrm{WNG_m}$ using array_1 and are approximately constant or do not vary in a clearly interpretable way for array_2.

5.2. Sensor noise

The evaluations with regard to the perceived sensor noise for array_1 are considerably different from the evaluations for array_2. Especially for lower $\mathrm{WNG_m}$ ($\mathrm{WNG_m} \leq -3$ dB), the sensor noise for array_1 is evaluated worse compared to the evaluations for array_2. The evaluations improve with increasing $\mathrm{WNG_m}$, especially for subjects $\mathrm{S_1}$ and $\mathrm{S_4}$ where the evaluations for $\mathrm{WNG_m}$=0 dB and array_1 are approximately in the range of the evaluations for array_2. The evaluations for array_2 vary much less with the $\mathrm{WNG_m}$, resulting for subjects $\mathrm{S_1}$ and $\mathrm{S_4}$ in variations of approximately the amount of their standard deviations (over randomized presentations). This effect is also represented by the associated p-values, with relatively small p-values ($p \leq 0.004$) for all directions Θ and array_1 and rather high p-values ($p \geq 0.049$) for all directions Θ and array_2. On the other hand, there also seems to be a slight trend towards better evaluations for higher $\mathrm{WNG_m}$ with array_2, with the worst evaluations for the lowest $\mathrm{WNG_m}$ of -9 dB (in the averaged evaluations in Figure 4 as well as for subject $\mathrm{S_2}$ and $\mathrm{S_3}$ and $\Theta = 225°$ in Figure 3). This indicates that sensor noise is not negligible for all subjects even with array_2. However, the averaged evaluations in Fig. 4 as well as the associated p-values in Table 1 indicate that the gathered evaluations vary much less with the tested $\mathrm{WNG_m}$ when using array_2 compared to array_1.

In sum, the perceptually optimal $\mathrm{WNG_m}$ with regard to sensor noise seems to vary with the used microphone array and its inherent sensor noise. The evaluations of the sensor noise (if detectable) seem generally to enhance with higher $\mathrm{WNG_m}$, which was to be expected.

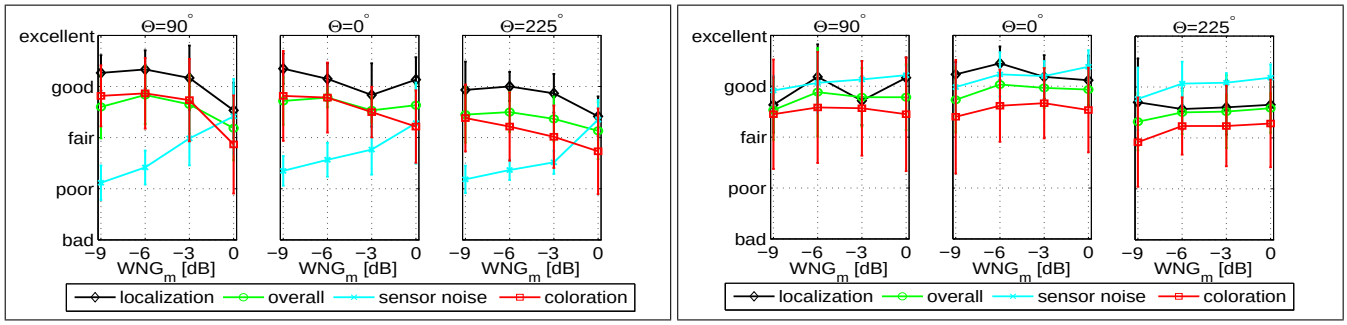

Figure 4: Perceptual evaluations averaged over all subjects for the array_1 (left block) and array_2 (right block) are depicted as the mean and the standard deviation for the four aspects to be evaluated (localization, overall performance, sensor noise and coloration).

5.3. Overall performance

The largest variations of the evaluations with regard to overall performance can be observed across different subjects, while the evaluations remain rather constant over different $\mathrm{WNG_m}$, especially for subject $\mathrm{S_3}$ with both microphone arrays. However, there seems to be a slight trend to worse evaluations for higher $\mathrm{WNG_m}$ using array_1 (cf. $\Theta = 90°$ and $\Theta = 225°$) as well as for the lowest $\mathrm{WNG_m}$ of -9 dB (presumably due to the more disturbing sensor noise). This trend is also apparent from the averaged performance using array_1 in Figure 4, with the Friedman test indicating the largest effect of the $\mathrm{WNG_m}$ for $\Theta = 90°$.

The evaluations vary less clearly with the $\mathrm{WNG_m}$ for array_2. There, the best evaluations were mostly observed at higher $\mathrm{WNG_m}$ (cf. $\mathrm{S_1}$, $\Theta = 225°$ and $\mathrm{S_2}$, $\Theta = 0°$) and worsened slightly for the lowest $\mathrm{WNG_m}$ (cf. Figure 4). In general, the evaluations with regard to overall performance seem to be correlated to the evaluations with regard to spectral coloration (cf. section 5.4), again emphasizing the relevance of spectral coloration for the evaluation of a binaural re-synthesis with respect to a reference condition. Furthermore, comparing the averaged evaluations of the overall performance for both microphone arrays (cf. Figure 4) at higher $\mathrm{WNG_m}$, the evaluations seem better for array_2 compared to array_1. This aspect is assumed to be a consequence of the lower inherent sensor noise of array_2: Typically, the Lagrangian multiplier μ is lower for lower desired $\mathrm{WNG_m}$. To achieve a desired

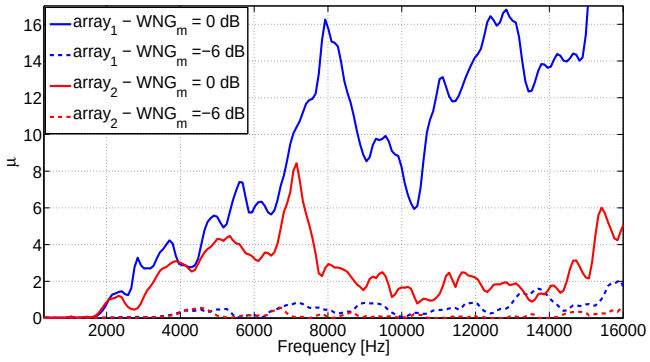

Figure 5: *Exemplary course of the Lagrangian multiplier μ (cf. Eq. 5) for* array_1 *and* array_2 *(blue and red lines, respectively) and* $\mathrm{WNG_m}$ *of 0 dB and -6 dB (solid and dashed lines, respectively) as a function of frequency of the left-ear re-synthesis for* $\mathrm{S_1}$.

$\mathrm{WNG_m}$, the required μ is usually lower (empirical observation) for array_2 compared to array_1, cf. Figure 5. Although not shown here, this tendency has also been observed for the other subjects and $\mathrm{WNG_m}$. A possible explanation could be that μ needs to be enlarged more in order to counteract the higher inherent sensor noise of array_1 (resulting in larger random errors on the measured steering vectors) in comparison to array_2. Considering that the accuracy of a re-synthesis decreases with larger μ, the higher inherent sensor noise of array_1 may therefore be a reasonable explanation for a worse accuracy of the re-syntheses and subsequently for the worse evaluations at $\mathrm{WNG_m} \gtrsim -3$ dB.

In sum, the evaluations with regard to overall performance seem best for $\mathrm{WNG_m}$=-6 dB and $\mathrm{WNG_m}$=-3 dB when using array_1 and for $\mathrm{WNG_m} \geq$=-6 dB when using array_2.

5.4. Spectral coloration

The evaluations with regard to spectral coloration seem to differ considerably for the four subjects. This phenomenon may be partly explained by the fact that the perception and evaluation of spectral coloration is influenced by the perceived localization and the interaction with the perceived sensor noise. This may introduce a certain degree of interpretation to assess this aspect. Furthermore, subjects have individual internal scales and assess individually. This is primarily evident when comparing the evaluations of subject $\mathrm{S_2}$ and $\mathrm{S_3}$, for instance. The evaluations of subject $\mathrm{S_3}$ vary roughly between good and excellent while the evaluations of subject $\mathrm{S_2}$ vary roughly between fair and poor, representing the most negative evaluations of this study.

In general, slightly better evaluations are evident for the frontal direction $\Theta = 0°$ compared with the lateral directions. The averaged evaluations in Figure 4 as well as the p-values in Table 1 indicate that the evaluations for array_1 vary considerably across the tested $\mathrm{WNG_m}$ for all tested directions Θ with decreasing averaged evaluations for higher $\mathrm{WNG_m}$ in Figure 4. This tendency does, however, not hold for array_2, with its p-values being relatively high ($p \geq 0.319$) for all directions. This array-dependent difference of evaluations may be explained by the differently sized Lagrangian multipliers μ for the two applied arrays (cf. Figure 5 and the discussion in section 5.3).

In sum, the evaluations of the perceived spectral coloration seem to vary with subjects and also with the used microphone arrays. Higher $\mathrm{WNG_m}$ seem to distort the perception of spectral coloration for array_1. On the other hand, the evaluations with regard to spectral coloration do not seem to vary considerably with the tested $\mathrm{WNG_m}$ when using array_2.

6. CONCLUSIONS AND FURTHER WORK

In this work the effect of regularization on the appraisal of binaural reproduction was investigated. Firstly, we introduced an alternative definition of a WNG-criterion, which is better suited to re-synthesize HRTFs using microphone arrays.
Secondly, the evaluation of the perceived sensor noise (if noticeable) seems to improve considerably with increasing $\mathrm{WNG_m}$, whereas the explicit presence of sensor noise (primarily at lower $\mathrm{WNG_m}$ with array_1) does not consistently seem to deteriorate the overall performance. This latter observation may be due to the chosen test paradigm - it is conceivable that noise is more disturbing in other scenarios, e.g. when listening to music recordings. Furthermore, the higher sensor noise of array_1 seems also to have caused worse evaluations with regard to localization, coloration and overall performance for $\mathrm{WNG_m} \gtrsim -3$ dB. This phenomenon may be explained by the empirically higher Lagrangian multipliers μ that were required for array_1 to comply with a fixed $\mathrm{WNG_m}$ (cf. section 5.3).
The best compromise with regard to all assessed aspects and the associated robustness can be found at $\mathrm{WNG_m}$ of -6 dB and -3 dB for array_1 and at the highest of the tested $\mathrm{WNG_m}$ of 0 dB for array_2.
In general, the obtained evaluations confirm the validity of re-synthesizing HRTFs using microphone arrays in conjunction with individually suited $\mathrm{WNG_m}$. There is still room for improvement for the calculation and regularization of the filter coefficients, especially with regard to spectral coloration. Thus, one next step may be to elaborate a more appropriate and frequency-dependent regularization method.

7. ACKNOWLEDGMENTS

This project was partially funded by Bundesministerium für Bildung und Forschung under grant no. 17080X10, by Akustik Technologie Göttingen and by the Cluster of Excellence 1077 "Hearing4All", funded by the German Research Foundation (DFG).

8. REFERENCES

[1] V. Mellert and N. Tohtuyeva, "Multimicrophone arrangement as a substitute for dummy-head recording technique," in *In Proc. 137th ASA Meeting*, 1997, p. 3117.

[2] Y. Kahana, P.A. Nelson, O. Kirkeby, and H. Hamada, "A multiple microphone recording technique for the generation of virtual acoustic images," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1503–1516, 1999.

[3] J. Atkins, "Robust beamforming and steering of arbitrary-beam patterns using spherical arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 16-19 2011, pp. 237–240.

[4] E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. van de Par, V. Mellert, and D. Püschel, "Robustness of virtual artificial head topologies with respect to microphone positioning errors," in *Proc. Forum Acusticum, Aalborg*, Aalborg, 2011, pp. 2251–2256.

[5] E. Rasumow, M. Blau, S. Doclo, M. Hansen, S. Van de Par, D. Püschel, and V. Mellert, "Least squares versus non-linear cost functions for a vitual artificial head," in *Proceedings of Meetings on Acoustics*. 2013, vol. 19, pp. –, ASA.

[6] D. N. Zotkin, R. Duraiswami, and N.A Gumerov, "Regularized hrtf fitting using spherical harmonics," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 18-21 2009, pp. 257–260.

[7] Cesar D. Salvador Castaneda, Shuichi Sakamoto, Jorge A. Trevino Lopez, Junfeng Li, Yonghong Yan, and Yoiti Suzuki, "Accuracy of head-related transfer functions synthesized with spherical microphone arrays," in *Proceedings of Meetings on Acoustics*. 2013, vol. 19, pp. –, ASA.

[8] Shuichi Sakamoto, Satoshi Hongo, Takuma Okamoto, Yukio Iwaya, and Yoit Suzuki, "Improvement of accuracy of three-dimensional sound space synthesized by real-time "senzi", a sound space information acquisition system using spherical array with numerous microphones," in *Proceedings of Meetings on Acoustics*. 2013, vol. 19, pp. –, ASA.

[9] S. Doclo and M. Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, vol. 51, no. 10, pp. 2511–2526, October 2003.

[10] K.U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, Michael Brandstein and Darren Ward, Eds., Digital Signal Processing, pp. 39–60. Springer Berlin Heidelberg, Berlin, Heidelberg, New York, May 2001.

[11] J. Bitzer and K.U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, Michael Brandstein and Darren Ward, Eds., Digital Signal Processing, pp. 19–37. Springer Berlin Heidelberg, Berlin, Heidelberg, New York, May 2001.

[12] E. Mabande, A. Schad, and W. Kellermann, "Design of robust superdirective beamformers as a convex optimization problem," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 77–80.

[13] Jian Li, Petre Stoica, and Zhisong Wang, "On robust capon beamforming and diagonal loading," *Signal Processing, IEEE Transactions on*, vol. 51, no. 7, pp. 1702–1715, July 2003.

[14] Ole Kirkeby and Philip A. Nelson, "Digital filter design for inversion problems in sound reproduction," *J. Audio Eng. Soc*, vol. 47, no. 7/8, pp. 583–595, 1999.

[15] D. Hammershøi and H. Møller, "Sound transmission to and within the human ear canal.," *Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 408–427, 1996.

[16] E. Rasumow, M. Blau, M. Hansen, S. Doclo, S. van de Par, D. Püschel, and V. Mellert, "Smoothing head-related transfer functions for a virtual artificial head," in *Acoustics 2012*, Nantes, France, April 2012, pp. 1019–1024.

[17] E. Rasumow, M. Blau, M. Hansen, S. van de Par, S. Doclo, V. Mellert, and D. Püschel, "Smoothing individual head-related transfer functions in the frequency and spatial domains," *Journal of the Acoustical Society of America*, 2014, accepted for publication.