High spatial resolution binaural sound reproduction using a virtual artificial head

Mina Fallahi[1], Martin Hansen[1] , Simon Doclo[2], Steven van de Par[2], Volker Mellert[2]
Dirk Püschel[3] , Matthias Blau[1]
[1] *Jade Hochschule Oldenburg, Institut für Hörtechnik und Audiologie, Germany*
[2] *Carl von Ossietzky Universität Oldenburg and Cluster of Excellence Hearing4All, Germany*
[3] *Akustik Technologie Göttingen, Germany*
Email: mina.fallahi@jade-hs.de

Introduction

The objective of binaural sound reproduction is to preserve the spatial properties of the sound field that are present in the binaural sound acquisition. An established method to capture these spatial properties is the use of so-called artificial heads. However, one of the main disadvantages of traditional artificial heads is their non-individual anthropometric geometrics, often leading to perceptible distortions of the spatial image.

As an alternative, a microphone array can be used as a filter-and-sum beamformer to synthesize individual head related transfer functions (HRTFs). The major advantage of such a recording system, referred to as Virtual Artificial Head (VAH), is that the same recording can be adapted post hoc to different individual HRTFs, both statically as well as dynamically (i.e. using head tracking), by modifying the array directivity pattern with individually calculated filter coefficients. A VAH based on a regularized least-squares cost function was developed in [1] for a planar microphone array with 24 microphones. This VAH was evaluated as successful compared to traditional artificial heads for synthesizing individual HRTFs at discrete angles of incidence in the horizontal plane [2].

After reviewing the method proposed in [1], the current study introduces a new constrained optimization method, aiming at controlling the synthesis error for a large number of azimuthal angles of incidence. The performance of the new method in comparison to the existing method is discussed based on simulation results.

Review of the regularized least-squares beamformer with a WNG constraint [1]

The directivity pattern $H(f,\theta_k)$ of a filter-and-sum beamformer at direction θ_k and frequency f is defined as

$$H(f,\theta_k) = \boldsymbol{w}^H(f)\boldsymbol{d}(f,\theta_k), \quad (1)$$

where $\boldsymbol{d}(f,\theta_k)$ denotes the steering vector describing the free-field acoustic transfer function between the sound source at direction θ_k and the N microphones of the array and $\boldsymbol{w}(f) = [w_1(f), w_2(f), \dots, w_N(f)]^T$ contains the complex-valued filter coefficients. Aiming at synthesizing a desired directivity pattern $D(f,\theta_k)$, e.g., an individual HRTF at either the left or the right ear, these filter coefficients can be calculated by minimizing a narrowband least-squares cost function J_{LS}, defined as the sum of the squared absolute differences between the synthesized and the desired directivity pattern over P directions, i.e.

$$J_{LS}(\boldsymbol{w}(f)) = \sum_{k=1}^{P} |H(f,\theta_k) - D(f,\theta_k)|^2 \quad (2)$$

In order to increase the robustness against microphone noise and small changes in the microphone positions or characteristics, some regularization constraints are usually added. A commonly used robustness measure is the White Noise Gain (WNG), which is defined as the ratio between the output power of the beamformer at direction θ_k and the output power for spatially uncorrelated noise. In contrast to many beamformers such as superdirective beamformers, for which the WNG is constrained only for the steering direction, it has been shown in [1] that for synthesizing HRTFs it is necessary to constrain the mean WNG over all considered directions, i.e.

$$WNG_m(\boldsymbol{w}(f)) = \frac{1}{P}\sum_{k=1}^{P} \frac{|\boldsymbol{w}^H(f)\boldsymbol{d}(f,\theta_k)|^2}{\boldsymbol{w}^H(f)\boldsymbol{w}(f)} \quad (3)$$

By imposing a minimum desired value β on the mean white noise gain, the regularized filter coefficients can then be computed by solving the following least-squares problem,

$$\min J_{LS}(\boldsymbol{w}(f)) \qquad s.t. \qquad WNG_m(\boldsymbol{w}(f)) \geq \beta. \quad (4)$$

The closed-form solution to Eqs. (2) and (4) can be found in [1]. The regularized method has been used for a planar array of 24 microphones (Figure 1) for 24 horizontal directions and has been evaluated as successful for a chosen subset of 6 directions in terms of localization, spectral coloration and overall performance [2].

High spatial resolution synthesis with the VAH in the horizontal plane

When considering the *non-regularized* least-squares cost function in Eq. (2), the best performance is achieved for the case that P, i.e. the number of directions included in the calibration process (referred to as calibration directions), is smaller than or equal to the number of microphones. The synthesis error will then be zero at the same P directions. However, if the synthesis is performed at an intermediate direction between two calibration directions, the synthesis error will typically be large because this intermediate direction has not been included in the calibration. To examine this, the steering vectors of the array shown in Figure 1 were simulated as relative delays $\tau_n(\theta_n)$ between the n-th microphone and the center of the array as

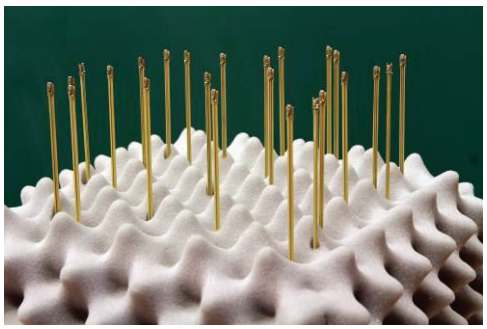

Figure 1: Planar microphone array with 24 microphones on a 20cm × 20cm ground plate [1].

$$\boldsymbol{d}(f,\theta_k) = \left[e^{-2\pi f\tau_1(\theta_k)j}, e^{-2\pi f\tau_2(\theta_k)j} \dots e^{-2\pi f\tau_N(\theta_k)j}\right]^T \quad (5)$$

To compute the filter coefficients P=24 horizontal calibration directions (corresponding to a 15° resolution) were considered and the calibration was performed separately for the HRTFs of the left and the right ear. The used HRTFs were taken from an HRTF database measured for the KEMAR artificial head [3] and were preprocessed (smoothing in spectral and spatial domains) according to [4]. The resulting filter coefficients were used to synthesize HRTFs on a 2°-grid of horizontal synthesis directions θ'_k (superset of the calibration directions). The performance of the synthesis is evaluated based on the spectral distortion (SD), which is defined as the deviation between the magnitude spectra of the synthesized HRTFs and the desired HRTFs for each frequency and direction as

$$SD(f,\theta'_k) = 20lg\left|\frac{H(f,\theta'_k)}{HRTF(f,\theta'_k)}\right| \text{ dB} \quad (6)$$

The spectral distortion for the synthesis with non-regularized filter coefficients (solutions of Eq.(2)) is shown in the upper part of Figure 2. As expected, the synthesis is perfect for the calibration directions (to be recognized as horizontal lines of zero synthesis error), but shows on the other hand large spectral distortion at intermediate directions. The spectral distortion for the synthesis with regularized filter coefficients (solutions of Eq. (4)) with $10\lg(WNG_m) \geq 0$dB is shown in the lower part of Figure 2. As can be observed, incorporating WNG regularization improves the performance at the intermediate directions, but deteriorates the performance at the calibration directions. The synthesis error seems to be redistributed over all directions. Since the mean WNG constraint aims at increasing the overall robustness against deviations in microphone positions or characteristics, it does not offer a direct solution to control the synthesis error at an intermediate direction between two calibration directions. Also, including more directions in the calibration does not provide a direct way to control the synthesis error.

New approach: constrained optimization

In order to increase the synthesis accuracy on a finer grid of horizontal directions, the new approach presented here uses a constrained optimization algorithm. Such algorithms

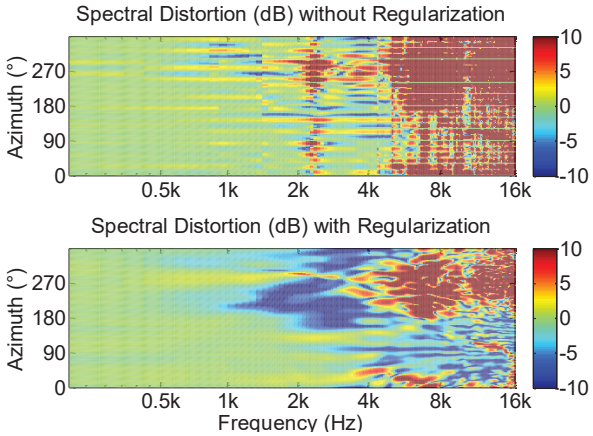

Figure 2: Spectral Distortion of synthesized HRTFs at 180 directions in the horizontal plane. Microphone array was calibrated for 24 of these 180 directions, without regularization (above), and with a regularization of $10\lg(WNG_m) \geq 0$dB. The results are shown only for the left ear.

optimize a cost function with respect to a variable, subject to some constraints set on this variable. For this constrained optimization case, the least-squares cost function of Eq. (2) was defined as the sum of the (squared) error over 180 directions ($\theta_k = 0°, 2°, \dots, 358°$) and was minimized subject to the constraint that the synthesis error at each of the 180 directions ($\theta'_k = 0°, 2°, \dots, 358°$) remains within a defined allowable range:

$$L_{Low} \leq 10lg\frac{|w^H d(\theta'_k)|^2}{|D(\theta'_k)|^2} \text{ dB} \leq L_{Up} \quad (7)$$

Please note that the dependence on frequency is not shown for more clarity. To guarantee a certain level of robustness an additional constraint was set on the average WNG_m such that

$$10\,lg\left(\frac{1}{180}\sum_{k=1}^{180}\frac{|\boldsymbol{w}^H\boldsymbol{d}(\theta_k)|^2}{\boldsymbol{w}^H\boldsymbol{w}}\right)\text{dB} \geq 0 \quad (8)$$

The minimization of the cost function in Eq. (2) subject to the constraints in Eq.(7) and Eq.(8) leads to an optimization problem with inequality constraints which has been solved here by applying the interior-point algorithm using the function **fmincon** in the MATLAB Optimization Toolbox [5]. This algorithm is based either on solutions of the Karush-Kuhn Tucker Problems via linear approximations of the problem or on trust-region methods, depending on the current iteration. Since the iterative algorithms should be supplied with an initial guess of the variables, the results of the previous method (solutions to Eq. (4)) were used here as the initial values.

The question arises how to define the allowable range of synthesis error. It is not strictly clear how much spectral deviation within a single frequency bin would lead to audible coloration or localization artifacts. However, starting with boundaries for the resulting interaural deviations, the acceptable amount of monaural spectral distortion can be

redefined so that the deviations in the resulting interaural features could be kept in a desirable range. According to different perceptual experiments the Just Noticeable Difference (JND) for the deviation in the Interaural Level Difference (ILD) varies between 0.6 dB and 2 dB, depending on the used stimuli [6]. In case of a broadband stimulus (clicks) a JND of 1.5 dB is reported in [6]. However, this value was obtained from measurements for frontal incidences in the median plane. Since the JND for the ILD deviation can get larger at lateral directions, as a compromise, here the JND of 2 dB was considered for all directions. Therefore, in order to meet a decision on the upper and lower boundaries L_{Up} and L_{Low} in Eq. (7) it was considered, that the resulting deviation in the ILDs at each of the directions $\theta'_k = 0°, 2°, \ldots, 358°$ should not exceed 2 dB, such that

$$\left| 10lg \frac{|\boldsymbol{w}_L^H \boldsymbol{d}(\theta'_k)|^2}{|\boldsymbol{w}_R^H \boldsymbol{d}(\theta'_k)|^2} \text{ dB} - 10lg \frac{|D_L(\theta'_k)|^2}{|D_R(\theta'_k)|^2} \text{ dB} \right| \leq 2\text{dB} \qquad (9)$$

The subscripts L and R refer to the synthesized or desired HRTFs of the left and right ears respectively. The combination of the requirement in Eq. (9) and the monaural spectral distortion in Eq. (7) leads to the result that the distance between the upper and lower error boundaries L_{Up} and L_{Low} should be 2dB. Among the unlimited number of possible combinations, the values 0.5 dB and -1.5 dB were chosen for L_{Up} and L_{Low} respectively, being more strict at the upper boundary and allowing more deviation at the lower boundary.

Simulation results and discussion

As simulation results in Figure 3 show, the spectral distortions of synthesized HRTFs on a 2°-resolution grid of horizontal directions now exhibit a clear improvement compared to previous results, with monaural deviations between -1.5dB and 0.5dB and a maximum ILD deviation of 2dB for all directions up to about 5kHz. In addition the condition of an average WNG of at least 0 dB is met for these frequencies as well.

From the results it can be seen that for higher frequencies the algorithm is no longer able to meet the desired constraints defined through Eqs. (7-9). This could be due to more spatial details contained in the HRTF directivity patterns at high frequencies as well as due to the effects of spatial aliasing for the given microphone array topology. In order to further achieve the desired solution from the constrained optimization, one could modify the constraints on the allowable spectral deviation at higher frequencies, for example through the relaxation of the constraints at perceptually less relevant directions (e.g. lateral directions for the contralateral side) to promote the optimization process. Alternatively, one could modify the microphone positions to improve the results. To examine the effect of microphone positions, the array shown in Figure 1 was simulated with microphone positions being halved in x- and y- coordinates with no further changes in the array topology, resulting in an array half of the size of the original microphone array. Simulation results with the same

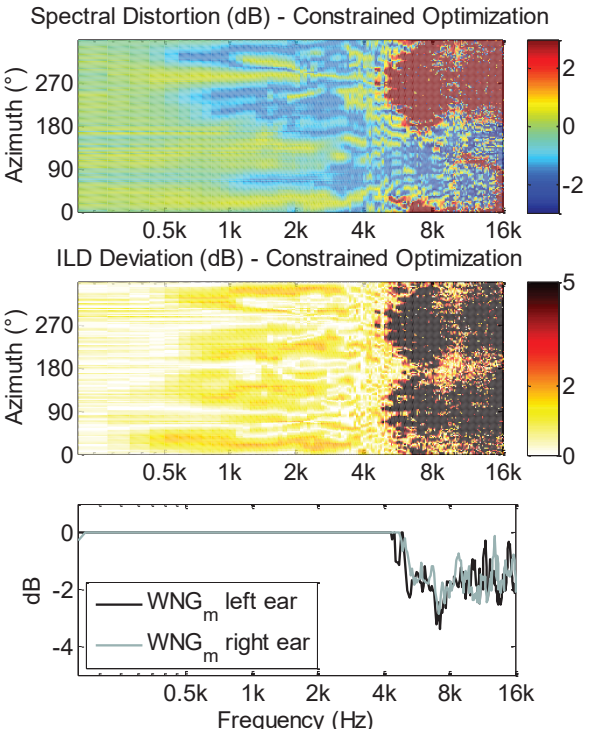

Figure 3: Spectral Distortion for the left ear (top), resulting ILD deviation (middle) and averaged WNG (bottom) for the synthesis of 180 horizontal directions, with filter coefficients calculated via constrained optimization for the simulated array shown in figure 1.

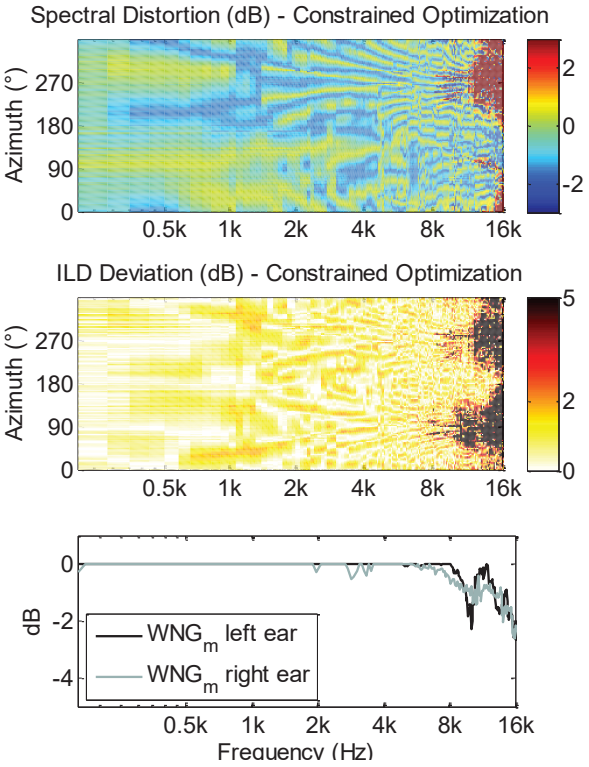

Figure 4: Spectral Distortion for the left ear (top), resulting ILD deviation (middle) and averaged WNG (bottom) for the synthesis of 180 horizontal directions with filter coefficients calculated via constrained optimization for the halved-size microphone array.

constraints as defined in Eqs. (7-9) for the halved-size array are shown in Figure 4. As can be seen, by decreasing thearray extension the same constraints could be met for all directions for higher frequencies further up to about 8 kHz.

Conclusion and outlook

A new approach based on constrained optimization was presented for synthesizing high spatial resolution horizontal HRTFs with the microphone array-based virtual artificial head. Simulation results showed that for the array used in previous work of our group [1], monaural spectral deviations in the synthesis could be kept to less than 2dB up to approx. 5 kHz for a 2° resolution grid, maintaining the resulting interaural level differences below the just noticeable difference [6]. The performance of the optimization algorithm could also be improved to higher frequencies of about 8kHz by using the half size microphone array, establishing a promising basis for forthcoming work. Future studies aim at the optimization of the array topology as well as perceptual evaluation of the synthesis with filter coefficients optimized for individually measured HRTFs. Furthermore, one major goal is to extend the virtual artificial head to capture the 3D sound field, including elevated sound sources as well as horizontally distributed ones. This would confront the beamformer with the challenging task of synthesizing the elevation cues included in form of sharp dips in the HRTF spectrum. An advantage of the virtual artificial head to be explored is the potentiality of changing the look direction of the beamformer by simply adjusting filter coefficients. Including head rotations via head tracking during signal representation will therefore be an important part of the future work.

Acknowledgement

This work was funded by Bundesministerium für Bildung und Forschung under grant no. 03FH021IX5.

Literature

[1] Rasumow, E., Hansen, M., van de Par, S., Püschel, D., Mellert, V., Doclo, S., Blau, M.: Regularization approaches for synthesizing HRTF directivity patterns. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2), 2016

[2] Rasumow, E., Blau, M., Doclo, S., Hansen, M., van de Par, S., Püschel, D., Mellert, V.: Perzeptive Evaluation von individualisierten und generischen binauralen Reproduktionen. In*: DAGA 2015, Nürnberg*, pp. 1097-1098

[3] Thiemann, J., Escher, A., van de Par, S.: Multiple model high-spatial resolution HRTF measurements. In: *DAGA 2015, Nürnberg*, pp. 797-798

[4] Rasumow, E., Blau, M., Hansen, M., van de Par, S., Doclo, S., Mellert, V., Püschel, V.: Smoothing individual head-related transfer functions in the frequency and spatial domains. *J. Acoust. Soc. Am., 135(4)*, pp. 2012-2025, Apr. 2014.

[5] *Mathworks, (2016), Documentation, Optimization Toolbox*, Retrieved February 2, 2017 from https://de.mathworks.com/help/optim/ug/constrained-nonlinear-optimization-algorithms.html#brnpd5f

[6] Blauert, J.: *Spatial hearing: the psychophysics of human sound localization*, Revised ed. Cambridge, Massachusetts: MIT Press, 1997