Constrained optimization for binaural sound reproduction using a virtual artificial head

Mina Fallahi[1], Matthias Blau[1], Martin Hansen[1], Simon Doclo[2], Steven van de Par[2], Dirk Püschel[3]

[1] *Jade Hochschule Oldenburg, Institüt für Hörtechnik und Audiologie*

[2] *Carl von Ossietzky Universität Oldenburg, Dept. für medizinische Physik und Akustik und Exzellenzcluster Hearing4All*

[3] *Akustik Technologie Göttingen*

Email: mina.fallahi@jade-hs.de

Introduction

Artificial heads are used as an established binaural recording method to capture the spatial properties of sound fields. However, due to their non-individual anthropometric geometries, these artificial heads often lead to perceptible deficiencies. As an alternative, individual Head Related Transfer Functions (HRTFs) can be synthesized with a microphone array-based filter-and-sum beamformer, referred to as a Virtual Artificial Head (VAH) [1]-[3]. The main advantage of a VAH is the possibility to adapt the same recording post hoc to different individual HRTFs by simply applying individually calculated filter coefficients. Furthermore, head tracking may be employed during the individualized binaural reproduction. A VAH based on a regularized least-squares cost function was proposed in [1] and was further improved in [4] by imposing boundaries on the synthesis error, resulting in a constrained optimization problem. However, the feasibility of finding a solution satisfying all constraints depends on a variety of parameters such as array topology and the used constraints. After providing a brief review of the methods proposed in [1] and [4] this study investigates the effect of relaxing the constraints on the constrained optimization performance. The results are discussed based on objectively calculated and perceptually measured outcomes.

Least-squares beamformer with constraints on WNG and Spectral Distortion

The objective of a VAH is to synthesize individual (left or right) HRTFs with directivity pattern $D(f,\Theta_k)$ using a filter-and-sum beamformer, where f denotes frequency and Θ_k denotes direction k. The synthesized directivity pattern of this beamformer, $H(f,\Theta_k)$, is defined as:

$$H(f,\Theta_k) = \mathbf{w}^H(f)\mathbf{d}(f,\Theta_k). \tag{1}$$

The $N \times 1$ steering vector $\mathbf{d}(f,\Theta_k)$ describes the free-field acoustic transfer function between a source at direction Θ_k and the N microphones of the array. The $N \times 1$ vector $\mathbf{w}(f)$ contains the complex-valued filter coefficients for each microphone. These filter coefficients can be computed by minimizing a narrow-band least-squares cost function, defined as the sum of the squared absolute differences between the desired and the synthesized directivity patterns across all P directions, i.e.

$$J_{LS}(\mathbf{w}(f)) = \sum_{k=1}^{P} |H(f,\Theta_k) - D(f,\Theta_k)|^2. \tag{2}$$

To increase the robustness against deviations in the microphone positions and characteristics and against microphone self-noise, a White Noise Gain (WNG) constraint is typically imposed. In contrast to many beamformers, e.g. super directive beamformers, it has been shown in [1] that for synthesizing HRTFs it is advantageous to constrain the *mean* White Noise Gain, WNG_m, which is defined as the ratio between the mean output power of the beamformer over all considered directions and the output power of spatially uncorrelated noise, i.e.

$$\mathrm{WNG}_m(\mathbf{w}(f)) = \frac{1}{P}\sum_{k=1}^{P} \frac{|\mathbf{w}^H(f)\mathbf{d}(f,\Theta_k)|^2}{\mathbf{w}^H(f)\mathbf{w}(f)}. \tag{3}$$

A closed-form solution for minimizing J_{LS} subject to a mean WNG constraint has been derived in [1], ensuring more robustness at the cost of a larger synthesis error. To achieve a small synthesis error for a large number of directions, a new method was proposed in [4], which additionally imposes constraints on the Spectral Distortion (SD) at directions θ_k, $k = 1, 2, ..., p$, by setting an upper and lower limit, L_{Up} and L_{Low}, i.e. for all k

$$L_{\mathrm{Low}} \leq SD(f,\theta_k) = 10\lg\frac{\left|\mathbf{w}^H(f)\mathbf{d}(f,\theta_k)\right|^2}{|D(f,\theta_k)|^2}\mathrm{dB} \leq L_{\mathrm{Up}}. \tag{4}$$

A minimum desired value β was set for the WNG_m as an additional constraint to guarantee a certain level of robustness:

$$10\lg\left(\frac{1}{p}\sum_{k=1}^{p} \frac{|\mathbf{w}^H(f)\mathbf{d}(f,\theta_k)|^2}{\mathbf{w}^H(f)\mathbf{w}(f)}\right)\mathrm{dB} \geq \beta. \tag{5}$$

Minimizing J_{LS} subject to the inequality constraints in Eq.(4) and (5) was done by applying an iterative Interior-Point optimization method, using the solution proposed in [1] as the initial value for the iterative method.

Impact of the constraints on the constrained optimization performance

In this study, the impact of the values chosen for L_{Up}, L_{Low}, and β in Eq.(4) and (5) on the constrained optimization performance was investigated since the performance obviously depends on how strict the constraints

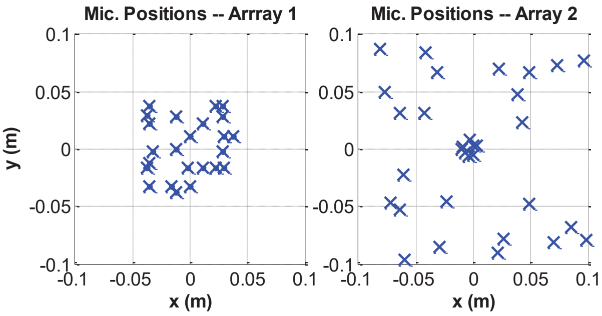


Figure 1: Microphone positions of both considered array topologies. Left: Planar array with 24 microphones as described in [5] ("Array 1"). Right: Array with 32 microphones ("Array 2"), consisting of 24 outer microphones and 8 microphones close to the center of the array.

are set. Since the microphone array topology and size [5] also influence the synthesis accuracy, two different array topologies as shown in Fig. 1 (Array 1 [5] and Array 2) were simulated. As an objective measure for the constrained optimization performance, the Success Rate (SR) was defined as the percentage of narrow-band optimizations at 170 Hz $\leq$ f $\leq$ 16 kHz for which all constraints could be satisfied.

We consider the case (referred to as Fixed) with fixed values L_{Up} = 0.5 dB and L_{Low} = -1.5 dB, aiming at maximum Interaural Level Difference (ILD) deviations of 2 dB, for 48 equidistant horizontal directions (i.e. 7.5° resolution), and with a fixed value $\beta = 0$ dB for the minimum desired WNG$_m$. The resulting SD and WNG$_m$ at the left ear when synthesizing $p = 48$ horizontal HRTFs with both simulated arrays are shown in Fig. 2a. It can be observed that for Array 1 only a success rate of 24% is achieved, while for Array 2 (with different size and a larger number of microphones) a much larger success rate of 67% is achieved, however still with room for improvement. In order to increase the success rate of the constrained optimization, three different constraint modifications were applied (c.f. Table 1) either by relaxing the constraints or by reducing the number of constraints. When relaxing the spectral distortion constraint in Eq.(4), the aim was to only allow for a decrease of the resulting spectral distortion (and no increase) because positive (narrow band) spectral components will generally be easier detectable than negative spectral components. Therefore, L_{Up} remained unchanged and modifications were applied only to L_{Low}.

The first modification (referred to as Relaxed L_{Low}) was to relax L_{Low} at contralateral directions, defined here as 200° $\leq \theta_{cl} \leq$ 340° and 20° $\leq \theta_{cl} \leq$ 160° for the left and the right ear, respectively. At each contralateral direction L_{Low} was reduced as a function of the difference between the amplitude of the desired directivity pattern D at this direction and the maximum amplitude of the directivity pattern ($|D|_{max}$), i.e.:

$$L_{\text{Low}}(f, \theta_{cl}) = -1.5 - \alpha(|D(f)|_{max} - |D(f, \theta_{cl})|), \quad (6)$$

where the factor α determines how much L_{Low} is reduced. Starting with $\alpha = 0$, α was incremented in steps of 0.1 (with an upper limit of 0.6 to limit the computation time) until the constraints could be satisfied for all directions. The second modification (referred to as Relaxed WNG) was to relax the minimum desired value of WNG$_m$. Starting at $\beta = 0$ dB, β was reduced in steps of 1 dB (with β_{min} = -13 dB as the minimum value) until all constraints could be satisfied. The third modification combined the relaxation of β with a reduction of the number of constraints by decreasing the spatial resolution of the directivity pattern from 7.5° to 15°(referred to as Relaxed WNG + Res.).

For all considered constraint modifications the resulting SD and WNG$_m$ at the left ear as well as the success rate are shown in Fig. 2 b-d. Besides the fact that Array 2 generally led to higher success rates, the constraint modifications considerably improved the optimization performance for both arrays compared to the Fixed case. This improvement was, however, either at the cost of increased negative SD at the contralateral side (down -9 dB for the Relaxed L_{Low} case), less robustness (WNG$_m$ of less than -10 dB for the Relaxed WNG and Relaxed WNG + Res. cases, especially for Array 1), or high positive and negative SDs at intermediate directions which were not considered in the constrained optimization when the resolution was reduced (Relaxed WNG + Res. case). For the latter case, spectral deviations exceeded 16 dB and -30 dB, which can be clearly observed as dark horizontal lines in the resulting SDs in Fig. 2d.

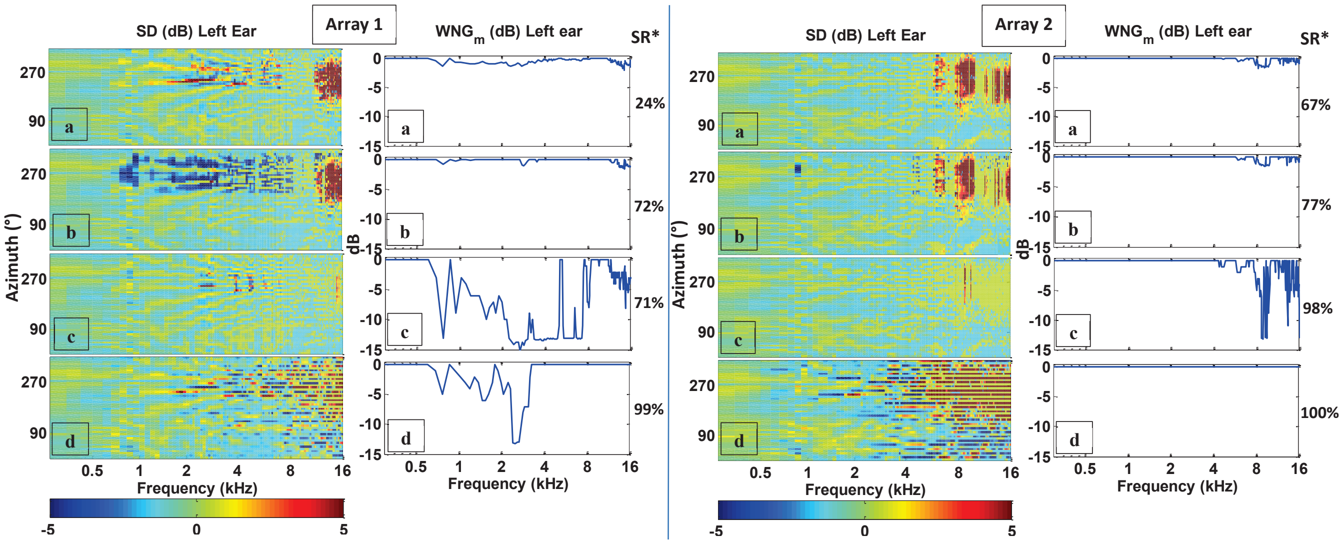
In the next step, a subjective listening test was performed to evaluate the perceptual quality of the synthesis with these different constraints.

Table 1: Description of different constraints

Name	Description
Fixed	β=0 dB, for all directions: L_{Up} = 0.5 dB , $L_{\text{Low}} = -1.5$ dB, resolution = 7.5°
Relaxed L_{Low}	β=0 dB, for all directions: L_{Up} = 0.5 dB , for contralateral directions* relax L_{Low} according to Eq.(6) (α = 0 : 0.1 : 0.6) until all constraints are satisfied, resolution = 7.5°
Relaxed WNG	for all directions: L_{Up} = 0.5 dB , $L_{\text{Low}} = -1.5$ dB, start with β=0 dB and reduce it by 1dB until all constraints are satisfied. Stop by $\beta_{min} = -13$ dB, resolution = 7.5°
Relaxed WNG + Res.	as Relaxed WNG, but resolution = 15°
* Contralateral directions: 200° $\leq \theta_{cl} \leq$ 340° and 20° $\leq \theta_{cl} \leq$ 160° for the left and the right ear, respectively	

Perceptual evaluation and discussion

For the subjective listening test individually acquired horizontal HRTFs with 7.5° azimutal resolution and individually measured Headphone Transfer Functions



*SR (Success Rate): Percentage of narrowband optimizations (170 Hz ≤ f ≤ 16 kHz) with all constraints satisfied.

Figure 2: Resulting Spectral Distortion (SD), mean White Noise Gain (WNG$_m$) and Success Rate (SR) at the left ear for the synthesis with Array 1 (left) and Array 2 (right). Four different constraints: (a) Fixed, (b) Relaxed L_{Low}, (c) Relaxed WNG, (d) Relaxed WNG + Res. (See Table 1 for a detailed description)

(HPTFs) of 10 subjects were considered (simulation results shown in Fig. 2 refer to the HRTFs of one of these subjects). The measured HRTFs were first smoothed in the frequency and spatial domain [6] and then synthesized with Array 1 and Array 2, applying the different constraints described in Table 1. The simulated microphone arrays were considered as perfectly robust, i.e. no deviations in microphone characteristics or microphone noise were considered. The test signal was filtered either with the individually measured HRTFs or with the synthesized HRTFs, and subsequently with the inverse individual HPTF prior to headphone presentation. Participants rated the binaural signals generated with the synthesized HRTFs with respect to the reference (binaural reproduction with individually measured HRTFs). Binaural signals generated with HRTFs measured for the KEMAR artificial head were also presented as an anchor signal.

Two attributes, i.e. spectral coloration and localization, were evaluated on a continuous scale, equidistantly marked with bad, poor, fair, good, and excellent. For rating the spectral coloration, three directions $\theta = 0°$, 60° and 247.5° were considered and the test signal consisted of three short bursts of pink noise (each lasting $\frac{1}{3}$ s followed by $\frac{1}{6}$ s of silence) with a spectral content of 300 Hz $\leq f \leq$ 16 kHz. For evaluating the localization, again three directions were considered. For each direction a sound source moving over a course of seven subsequent positions was presented, either from 22.5° to -22.5° (0° ±22.5°), from 37.5° to 82.5° (60° ±22.5°), or from 217.5° to 262.5° (240° ± 22.5°). From each source position, a single pink noise pulse was presented, with the same spectrum and length as described above. For both attributes, each direction appeared three times in a randomized order.

The results of the perceptual evaluations are shown in Fig. 3. For the Fixed case the median for evaluations of spectral coloration and localization lie between fair and good for both arrays. For the Relaxed L_{Low} case, the increased negative contralateral SD of Array 1 at ca. 1 kHz $\leq f \leq$ 8 kHz led to slightly lower ratings for spectral coloration, while it did not lead to lower ratings for localization. Array 2 suffered less from modifications on L_{Low} due to its topology (larger size and more microphones) and led more often to optimizations with all constraints satisfied at $f \leq$ 8 kHz. For the Relaxed WNG case, the ratings for both attributes showed no specific differences to the Fixed case for Array 1, even though the success rate was improved compared to Fixed case (see Fig. 2). For Array 2, the ratings improved slightly at both evaluated lateral directions. For the Relaxed WNG+Res case, there was a pronounced deterioration of the ratings for spectral coloration at θ=247.5° and the ratings for localization at all directions for Array 2, resulting in ratings comparable or even worse than the anchor signal. This is however not unexpected, since every second direction, including θ=247.5°, was excluded from the constrained optimization in this case. Due to the smaller size of Array 1, the resulting spectral distortions at these excluded directions appear first at higher frequencies compared to Array 2, especially for the ipsilateral side, so that the ratings for Array 1 showed a smaller deterioration than for Array 2.

In summary, these results show that although relaxing the constraints may improve the success rate, it may also negatively affect the perceptual evaluations. In particular, reducing the horizontal resolution does not seem to be a good idea. Moreover, the higher ratings for Arrary 2 compared to Array 1 in this study suggest that the choice of microphone array topology is also very important. The Relaxed WNG case, yielding both a high success rate (over 95%) and improved perceptual ratings for Array 2 in this study, seems to be a good option. Nevertheless, it should be noted that the arrays were not

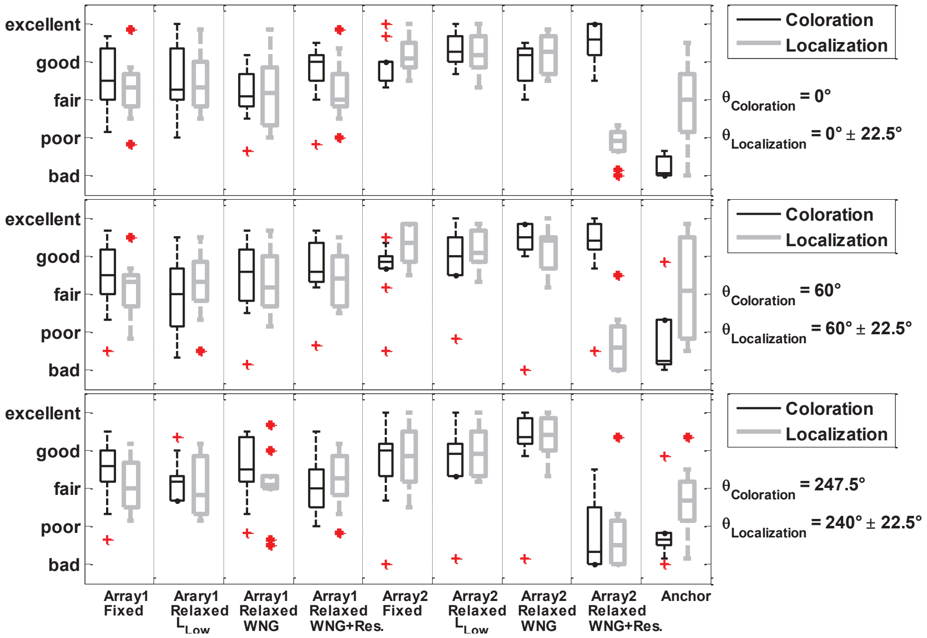


Figure 3: Results of perceptual evaluations for 10 subjects regarding different constraints applied to Array 1 and Array 2 as well as the anchor signal. Two perceptual attributes (Coloration and Localization) are evaluated for three different source positions.

evaluated with respect to their robustness against microphone deviations and microphone noise, for which a subsequent investigation is required.

Conclusion

In this study a constrained optimization method was applied to synthesize individual horizontal HRTFs using a microphone array-based virtual artificial head, with constraints set on the spectral distortion and mean white noise gain WNG_m. Four different constraints were applied to two array topologies to investigate their effect on the optimization performance. Simulations and perceptual results confirm that a proper relaxation of the constraints, especially a relaxed constraint on the WNG_m, in combination with an appropriate array topology can lead to satisfying more constraints without deteriorating (sometimes even improving) the perceptual evaluations. Further investigations are essential to evaluate the performance of constrained optimization with relaxed constraints on the WNG_m in terms of robustness.

Acknowledgments

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

References

[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), pp. 215-225, 2016

[2] Sakamoto, S., Hongo, S., Okamoto, T., Iwaya, Y., Suzuki, Y.: Sound-space recording and binaural presentation system based on a 252-channel microphone array. *Acoust. Sci & Tech.* 36(6), pp.516-526, 2015

[3] Atkins, J.: Robust beamforming and steering of arbitrary beam patterns using spherical arrays. *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA), New Paltz, NY, USA, Oct. 2011*, pp. 237-240

[4] Fallahi, M., Hansen, M, Doclo, S., van de Par, S., Mellert, V., Püschel, D., Blau, M.: High spatial resolution binaural sound reproduction using a virtual artificial head. *Fortschritte der Akustik, DAGA 2017, Kiel*, pp. 1061-1064

[5] Fallahi, M., Blau, M, Hansen, M., Doclo, S., van de Par, S., Püschel, D.: Optimizing the microphone array size for a virtual artificial head. *Proc. of International Symposium on Auditory and Audiological Research, Nyborg, Aug. 2017*

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