Audio Engineering Society

Conference e-Brief 61

Presented at the Conference on
Spatial Reproduction
2018 August 6 – 9, Tokyo, Japan

Individual binaural reproduction of music recordings using a virtual artificial head

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

[1]*Institut für Hörtechnik und Audiologie, Jade Hochschule, Oldenburg, Germany*
[2]*Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany*
[3]*Akustik Technologie Göttingen, Göttingen, Germany*

Correspondence should be addressed to Mina Fallahi (mina.fallahi@jade-hs.de)

ABSTRACT

As an alternative to traditional artificial heads, a virtual artificial head (VAH) comprising a microphone array-based beamformer can be used to capture the spatial properties of a listener's head and ears in the sound field. The advantage of a VAH is the possibility to adapt the same recording post hoc to individual Head Related Transfer Functions (HRTFs) and to use head-tracking in the binaural reproduction. Here, a narrow-band least-squares cost function was minimized to calculate the filter coefficients for the VAH with additional constraints on the spectral accuracy and beamformer robustness. Different versions of these constraints were applied to two simulated microphone array topologies and their effect on the resulting binaural reproduction is discussed based on objective results and perceptual experiments with spatially distributed music sources. The results show that, by choosing an appropriate array topology and properly defined constraints, the VAH can be used for a perceptually convincing reproduction of music content.

1 Introduction

In real life situations, humans are able to perceive sound sources in the surrounding three-dimensional space by processing the signals arriving at the ears. The source signal is not only affected by reflections in a reverberant environment, but also by the reflections and diffractions caused by the listener's head, torso and external ear, which provide the listener important cues for acoustical localization [1]. The objective of binaural technology is to preserve these spatial cues during sound reproduction via headphones. A well-established binaural recording method is the use of a so-called artificial head, which is a replica of an average human head and torso, equipped with microphones in the ear canals. Although a substantial amount of spatial information can be preserved in the recordings made with an artificial head, the non-individual anthropometry of these artificial heads often leads to perceptible artifacts such as front-back confusions or in-head localization.

As an alternative binaural recording method, a microphone array can be used to synthesize individual Head Related Transfer Functions (HRTFs) based on filter-and-sum beamforming [2]-[4]. The main advantage of this approach, referred to as Virtual Artificial Head (VAH), is that the same recording can be individual-

ized post hoc for different HRTFs by applying the individually calculated filter coefficients to the recorded microphone signals. This also makes it possible to incorporate head rotations during the reproduction. In addition, a VAH offers a high flexibility due to its small size and weight.

A VAH based on a regularized least-squares cost function was proposed in [5] and further improved in [6] by imposing constraints on the synthesis error, resulting in a constrained optimization problem. In this study four different cases of constraints were investigated for two simulated microphone array topologies. After a brief review of the methods proposed in [5] and [6], the optimized beamformers are presented and the results are discussed based on objectively calculated outcomes as well as perceptual experiments with music content.

2 Methods

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

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

where the $N \times 1$ steering vector $\mathbf{d}(f,\theta_k)$ denotes the free-field acoustical transfer function between the source at direction θ_k and the N microphones of the array and the vector $\mathbf{w}$ contains the complex-valued filter coefficients for each microphone. In order to synthesize directivity patterns $D(f,\theta_k)$ of individual HRTFs (left or right), these filter coefficients can be calculated by minimizing a narrow-band least-squares cost function J_{LS}, defined as the sum of the squared absolute differences between the synthesized and the desired directivity patterns over all P directions, i.e.

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

The desired directivity patterns to be synthesized by VAH in this study were individual HRTFs measured in the horizontal plane with 7.5° resolution, i.e. $P = 48$. To ensure a small synthesis error for these directions, additional constraints were imposed on the Spectral Distortion (SD) for each direction θ_k, $k = 1,2,...,P$, by setting an upper and lower limit, L_{Up} and L_{Low}, i.e. for all k:

$$L_{\mathrm{Low}} \leq SD(f,\theta_k) = 10\lg \frac{|\mathbf{w}^H(f)\mathbf{d}(f,\theta_k)|^2}{|D(f,\theta_k)|^2}\mathrm{dB} \leq L_{\mathrm{Up}}. \tag{3}$$

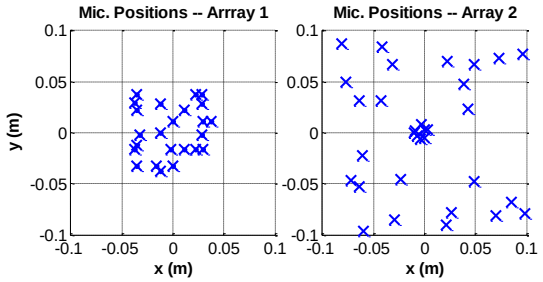

Fig. 1: Microphone positions of both considered array topologies. Left: Planar array with 24 microphones as described in [7] ("Array 1"). Right: Array with 32 microphones ("Array 2"), consisting of 24 outer microphones and 8 microphones close to the center of the array.

In addition, a constraint was imposed on the *mean* White Noise Gain, WNG_m, defined as the ratio between the mean output power of the beamformer over all P directions and the output power of spatially uncorrelated noise [5], i.e.

$$\mathrm{WNG}_m = 10\lg\left(\frac{1}{P}\sum_{k=1}^{P}\frac{|\mathbf{w}^H(f)\mathbf{d}(f,\theta_k)|^2}{\mathbf{w}^H(f)\mathbf{w}(f)}\right)\mathrm{dB} \geq \beta. \tag{4}$$

The constraint on the WNG_m serves to increase the robustness against microphone self noise or deviations in microphone positions and characteristics. An iterative Interior Point optimization algorithm was used to minimize J_{LS} in Eq.(2) subject to the inequality constraints in Eq.(3) and (4) using the solution proposed in [5] as the initial values for the iterative method.

Since the microphone array topology and size influence the synthesis accuracy [7], two different array topologies as shown in Fig. 1 (Array 1 [7] and Array 2) were simulated . In addition, the constrained optimization performance depends on how tight the constraints are set [8]. Therefore, we consider a "Fixed" constraint case and three modifications. For the Fixed case, fixed values L_{Up} = 0.5 dB and L_{Low} = -1.5 dB were used, aiming at maximum Interaural Level Difference (ILD) deviations of 2 dB for all horizontal directions (7.5° resolution). In addition, a fixed value β = 0 dB was used for the minimum desired WNG_m. The three modifications of the Fixed case consist of either relaxing the L_{Low} at the contralateral directions (referred to as

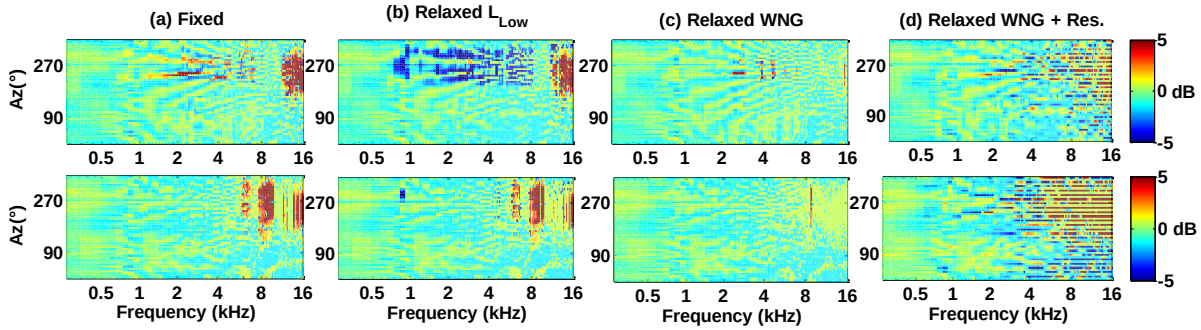

Fig. 2: Resulting Spectral Distortion (SD) at the left ear for the synthesis with Array 1 (top row) and Array 2 (bottom row) and for four different constraints.

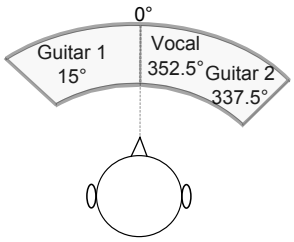

Fig. 3: Spatial setup of the virtual musical scene at $\theta = 0°$.

Relaxed L_{Low}), or relaxing the minimum desired value of WNG_m (referred to as Relaxed WNG), or relaxing the minimum desired value of WNG_m and reducing the spatial resolution from 48 to 24 directions corresponding to 15° resolution instead of 7.5° (referred to as Relaxed WNG + Res.). A more detailed description of these constraint cases can be found in [8]. The effect of the different constraint cases on the resulting SD at the left ear when synthesizing $P = 48$ horizontal HRTFs with both arrays is shown in Fig. 2. It has been shown in [8] that the discussed modifications improve the optimization performance considerably, leading to more optimizations with all constraints satisfied. This improvmenet however comes at the cost of deliberately allowing increased negative SD at the contralateral side ($200° \leq \theta \leq 340°$ for the left ear) for the Relaxed L_{Low} case (Fig. 2b) or increased positive and negative SD at directions which were excluded from the optimization for the Relaxed WNG + Res. case (visible as dark horizontal lines in Fig. 2d). Please note that the relaxation of the constraint on WNG_m will lead to a loss of robustness, the effect of which however was not considered in this study.

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

3 Perceptual evaluation

For the subjective listening test, individually measured horizontal HRTFs with 7.5° resolution and individually measured Headphone Transfer Functions (HPTFs) of 10 subjects were used (simulation results shown in Fig. 2 are based on HRTFs of one of these subjects). The measured HRTFs were smoothed in the spatial and spectral domains [9] and then synthesized with Array 1 and Array 2 applying the four aforementioned constraint cases. The test signal was filtered either with the individually measured HRTFs or with one of the four synthesized HRTFs, and subsequently with the inverse individual HPTFs before being presented via headphone. Binaural signals generated with measured HRTFs for the KEMAR artificial head were also presented as an anchor signal. Participants rated the binaural signals generated with synthesized HRTFs or the anchor signal compared to the reference signal (binaural signals generated with individually measured HRTFs) with regard to overall audio quality (without being instructed, whether the focus should be on spatial or spectral cues). The ratings were given on a 9-point scale which covered the labels bad, poor, fair, good, and excellent in four equidistant steps.

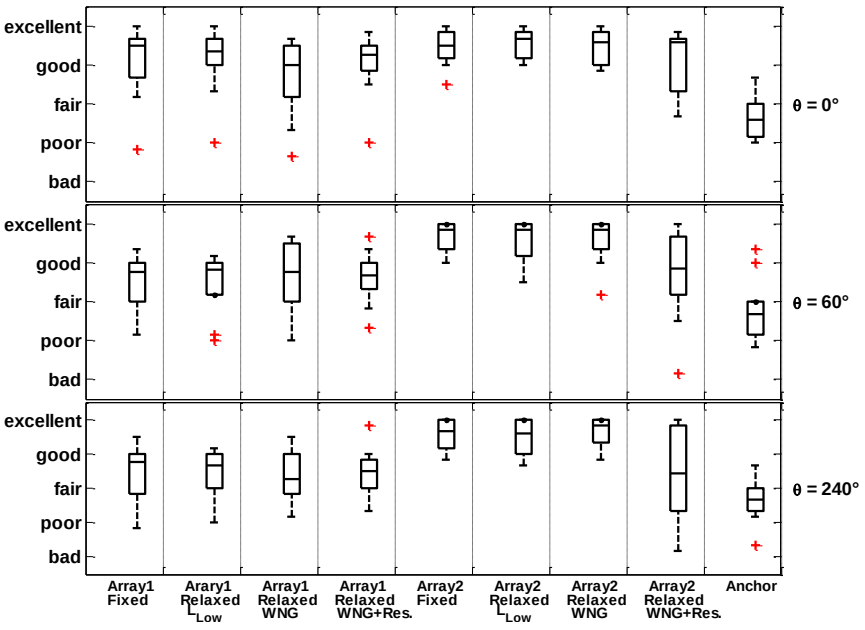

Fig. 4: Results of perceptual evaluations for 10 subjects regarding different constraints applied to Array 1 and Array 2 as well as the anchor signal for three different directions of the musical scene.

The test signal consisted of a piece of music recorded separately for a male voice, and two guitars. A virtual musical scene was created for $\theta = 0\,^\circ$ by filtering each recorded part with the (synthesized or measured) HRTFs at the directions shown in Fig. 3, i.e. 15° and 337.5° for the first and second guitar, respectively, and 352.5° for the vocal part. The musical scene was also generated for two other directions by rotating the musical scene shown in Fig. 3 to $\theta = 60\,^\circ$ (front left) and $\theta = 240\,^\circ$ (back right). Each direction of the music scene was presented three times in a randomized order. A segment of 1 minute duration of the test signal was presented in a continuous loop and the participants could switch freely between reference and synthesized versions during playback.

4 Results and discussion

The results of the perceptual evaluations are shown in Fig. 4. The ratings for Array 2 were in general higher than for Array 1. Median ratings lay between fair and good for Array 1 and between good and excellent for Array 2. One exception was the Relaxed WNG + Res. case for Array 2 where the ratings showed a noticeable deterioration in comparison to ratings for Fixed, Relaxed L_{Low} and Relaxed WNG cases, especially for the two lateral directions $\theta = 60°$ and $\theta = 240°$. This is due to the increased synthesis error at the directions which were excluded from the optimization for this case (Fig. 2d). On the other hand, the relaxed WNG + Res. case did not show any deterioration for Array 1 compared to the three other cases. This is because Array 1 has a smaller size than Array 2 and the spectral distortions at the excluded directions appear first at higher frequencies, especially for the ipsilateral side. Although for the Relaxed L_{Low} case the negative contralateral SD increases (see Fig. 2b), this case did not lead to a major degradation of the subjective ratings. Both arrays showed better ratings compared to the anchor signal (binaural signals generated with non-individual HRTFs).

In order to analyze whether there may be one case with a significantly different rating the Friedman test was applied. The results indicated significant differences for all directions with p = 0.0001 ($\theta = 0°$), p = 5.4e-08 ($\theta = 60°$) and p = 5.7e-08 ($\theta = 240°$). A post-hoc multiple comparison procedure revealed significant ($p<0.05$) differences between the anchor and Array 2 for the three

cases Fixed, Relaxed L_{Low} and Relaxed WNG at all evaluated directions. Significant differences between ratings for the two arrays occured at $\theta = 60^\circ$ between Fixed case for Array 1 and the two cases Fixed and Relaxed L_{Low} for Array 2 and at $\theta = 240^\circ$ between Relaxed WNG case for Array 1 and the two cases Fixed and Relaxed WNG for Array 2. There were no significant differences between the ratings for different cases of only Array 1 or Array 2.
The results for Array 2 with median ratings between good and excellent for the three cases Fixed, Relaxed L_{Low} and Relaxed WNG and a significant difference to the ratings for non-individual HRTFs (anchor signal) indicated that individual HRTFs synthesized with a VAH could be perceptually in accordance with individually measured HRTFs. In addition, higher ratings for Array 2 compared to Array 1 confirmed that the choice of array topology is also very important [7], [8]. However, it should be noted that the microphone arrays were not evaluated with respect to the robustness against microphone deviations and microphone self noise and a subsequent study is necessary to investigate the performance of the suggested method in this study.

5 Summary

For the design of a virtual artificial head, a constrained optimization method was used to synthesize horizontal HRTFs for two microphone array topologies with four different constraints on Spectral Distortion and mean White Noise Gain. The results show that, by using a proper array topology and appropriately defined constraints, perceptually convincing reproductions of music content can be achieved, which are rated significantly better than recordings with a classical artificial head. Further investigations are required to evaluate the performance of the constrained optimization method with respect to robustness and to include more directions (e.g., elevation) into the optimization.

6 Acknowledgments

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

References

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

[2] Rasumow, E., Blau, M., Doclo, S., van de Par, S., Hansen, M., Püschel, D., and Mellert, V., "Perceptual evaluation of individualized binaural reproduction using a virtual artificial head," *J. Audio Eng. Soc.* 65, pp. 448–459, 2017.

[3] Sakamoto, S., Hongo, S., Okamoto, T., Iwaya, Y., and 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.

[4] 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*, pp. 237–240, 2011.

[5] Rasumow, E., Hansen, M, van de Par, S., Püschel, D., Mellert, V., Doclo, S., and Blau, M., "Regularization approaches for synthesizing HRTF directivity patterns," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2), pp. 215–225, 2016.

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

[7] Fallahi, M., Blau, M, Hansen, M., Doclo, S., van de Par, S., and Püschel, D., "Optimizing the microphone array size for a virtual artificial head," *Proc. of International Symposium on Auditory and Audiological Research, Nyborg*, pp. 359–366, 2017.

[8] Fallahi, M., Blau, M., Hansen, M., Doclo, S., van de Par, S., and Püschel, D., "Constrained optimization for binaural sound reproduction using a virtual artificial head," *Fortschritte der Akustik-DAGA Munich*, 2018.

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