Optimizing the microphone array size for a virtual artificial head

MINA FALLAHI[1,*], MATTHIAS BLAU[1], MARTIN HANSEN[1], SIMON DOCLO[2], STEVEN VAN DE PAR[2], AND DIRK PÜSCHEL[3]

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

As an alternative to traditional artificial heads, individual head-related transfer functions (HRTFs) can be synthesized with a virtual artificial head (VAH) consisting of a multi-microphone array in combination with filter-and-sum signal processing. The accuracy of the synthesis depends, amongst others, on the number of microphones in the array and on its topology (array size and microphone positions). In this study the effect of microphone array size on the synthesis accuracy was investigated. Five simulated microphone arrays of different sizes were used to synthesize individual HRTFs in the horizontal plane. Objective results in terms of spectral distortion and ILD deviation as well as subjective results with 10 participants showed that array size has a major effect and that the synthesis accuracy can be improved by carefully choosing an appropriate array size.

INTRODUCTION

An established method to include the spatial properties of the sound within a binaural reproduction is the use of so-called artificial heads. With anthropometric characteristics of an average human head and torso, an artificial head preserves the spatial information in the sound field, which is crucial for sound source localization. However, the non-individual anthropometric geometries of these artificial heads often lead to perceptible deficiencies. Alternatively, a microphone array can be used as a filter-and-sum beamformer to synthesize individual head-related transfer functions (HRTFs). The major advantage of this system, referred to as a virtual artificial head (VAH), is the possibility to modify the individual calculated filter coefficients and to process the same recording post-hoc for individual HRTFs, both statically as well as dynamically (using head tracking). Another potential benefit of the VAH is its flexibility due to the smaller size and weight. One decisive aspect for the accuracy of the VAH is the choice of microphone array topology (including its size and microphone positions). Rasumow *et al.* (2016) developed a VAH as a planar

[*] Corresponding author: mina.fallahi@jade-hs.de

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 6: Adaptive Processes in Hearing, August 2017, Nyborg, Denmark. Edited by S. Santurette, T. Dau, J. C.-Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. The Danavox Jubilee Foundation, 2017. © The Authors. ISBN: 978-87-990013-6-1.

microphone array consisting of 24 microphones and showed that a regularized least-squares cost function approach could be used to synthesize individual binaural HRTFs accurately in the horizontal plane (c.f., also, Rasumow *et al.*, 2011, 2017). In accordance with this approach and with some modifications for increasing the spatial resolution of the VAH (c.f. Fallahi *et al.*, 2017)), the present study investigated the effect of the microphone array size on the accuracy of the VAH. Five microphone arrays of different sizes were simulated and used to synthesize individual HRTFs. After a brief review of the applied methods, the objective and perceptual evaluation of synthesis with different arrays sizes will be discussed.

HIGH SPATIAL RESOLUTION LEAST-SQUARES FILTER-AND-SUM BEAMFORMER

The desired directivity pattern $D(f,\Theta_k)$ of, e.g., an individual HRTF at either the left or right ear as a function of frequency f and direction Θ_k can be synthesized with the VAH as a filter and sum beamformer. Considering the $N\times 1$ steering vector $\mathbf{d}(f,\Theta_k)$ which describes the frequency and direction dependent transfer function between the source at direction Θ_k and the N microphones of the microphone array, the synthesized directivity pattern $H(f,\Theta_k)$ of the VAH is defined as

$$H(f,\Theta) = \mathbf{w}^H(f)\mathbf{d}(f,\Theta) \qquad \text{(Eq. 1)}$$

The $N\times 1$ vector $\mathbf{w}(f) = [w_1(f), w_2(f), ..., w_N(f)]^T$ contains the complex-valued filter coefficients for each microphone which can be derived by minimizing a narrowband least-squares cost function J_{LS}, defined as the sum over P directions of the squared absolute differences between the synthesized and desired directivity pattern, i.e.,

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

In order to increase the robustness of the system against deviations in microphone positions or characteristics (c.f. Rasumow *et al.*, 2011; Doclo *et al.*, 2003), Rasumow *et al.* (2016) derived a closed form solution for the minimization of the least-squares cost function subject to some regularization constraints, however at the cost of a general loss of accuracy. With the aim of maintaining the accuracy of synthesis for a high number of directions Fallahi *et al.* (2017) suggested a constrained optimization approach, minimizing the least-squares cost function for directions Θ_k , $k = 1,2,...,P$ subject to constraints set on the mean white noise gain (WNG$_\text{m}$, Rasumow *et al.*, 2016) and spectral distortion (*SD*) at synthesis directions θ_k , $k = 1,2,...,p$ by setting an upper and lower limit, L_up and L_low, for the synthesis error at each one of these directions, i.e., for all k

$$L_\text{low} \le SD(f,\theta_k) = 10\lg\frac{|\mathbf{w}^H(f)\mathbf{d}(f,\theta_k)|^2}{|D(f,\theta_k)|^2}\text{dB} \le L_\text{up} \qquad \text{(Eq. 3)}$$

The minimization of J_{LS} subject to inequality constraints described above was solved by applying the interior-point method, using the results of the closed form solutions by Rasumow *et al.* (2016) as the initial values for this iterative optimization algorithm.

INFLUENCE OF MICROPHONE ARRAY SIZE ON THE VAH SYNTHESES

The main goal of the current study was to investigate the effect of the array size on the synthesis accuracy. Adopting the topology of the microphone array shown in Fig. 1 (20 cm×20 cm planar array with 24 microphones, c.f. Rasumow *et al.*, 2011), the original array as well as downsized copies of it, namely 50%-, 37.5%- and 25%-size arrays were simulated. In addition, a combination of the 50%-size and 25%-size arrays was considered as well (named as 'Mix.' in the following), by taking the positions of the eight outermost microphones of the 50%-size array and the innermost positions of the 25%-size array for the rest. Using the constrained optimization approach described above, a set of individually measured horizontal HRTFs with 5° azimuthal resolution were synthesized with these arrays. L_{low} and L_{up} were set to -1.5 dB and 0.5 dB respectively, leading to a maximum range of interaural level difference (ILD) deviation of 2 dB at each of the 72 synthesized directions.

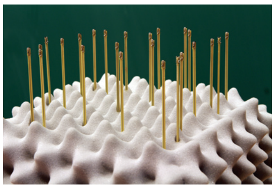

Fig. 1: Virtual artificial head: planar microphone array with 24 microphones (Rasumow *et al.*, 2016).

The results for spectral deviation at the left ear and the resulting ILD deviation are shown in Fig. 2. As can be seen, the constraints set on spectral error could not be met for all directions, especially not at high frequencies. For a given microphone array with a fixed size this could be due to more spatial details contained in the HRTF directivity patterns at higher frequencies as well as aliasing effects. For a smaller array, although the complexity of the HRTF's directivity pattern at the contralateral side (e.g., 270° for the left ear) could still be a challenge, the aliasing effects for the ipsilateral side could be shifted to higher frequencies, as can be seen in the simulation results of smaller arrays at higher frequencies in comparison to larger arrays (Fig. 2, left). At the same time, however, a reduced array size corresponds to an increased wavelength relative to the array size which leads to the widening of the synthesized directivity pattern (c.f. Ward *et al.*, 2001). At low frequencies this might not introduce a major problem since the HRTFs have a mostly omnidirectional directivity pattern. But in the mid-frequency range of about 1 to 4 kHz, the directivity pattern starts to get more complicated while the wavelength might still be large enough to prevent the beamformer from reaching sufficient damping at the contralateral side, leading to the increased spectral distortions and ILD deviations of smaller arrays in the mid-frequency range. The question now arises whether the resulting spectral distortions are perceptually relevant and which array size should be preferred.

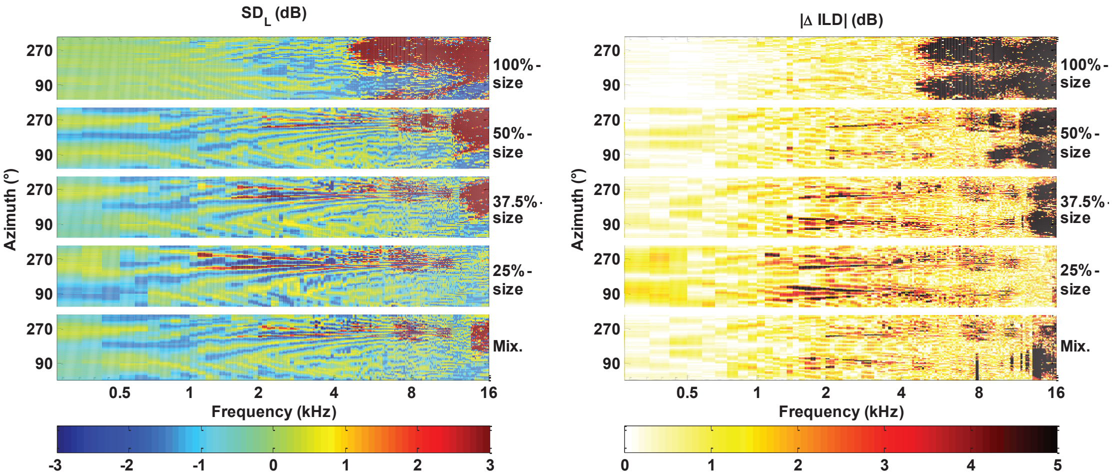

Fig. 2: Left: Spectral distortion (left ear). Right: ILD deviation between original and synthesized HRTFs with different array sizes.

EXPERIMENTAL PROCEDURE

In order to evaluate the perceptual quality of different array sizes, a listening experiment was performed. Prior to the listening test, individual horizontal HRTFs of 10 subjects were acquired with a 5° azimuthal resolution in a non-anechoic room (c.f. Koehler *et al.*, 2014), followed by individual measurements of the headphone transfer functions (HPTFs). The measured HRTFs were then smoothed both in frequency and spatial domain (c.f. Rasumow *et al.*, 2014) before being synthesized with different simulated microphone arrays. Three short bursts of pink noise with a spectral content of $300\text{Hz} \leq f \leq 16000\text{Hz}$ were chosen as the test signal. Each noise burst lasted $\frac{1}{3}$ s with 1-ms onset-offset ramps followed by $\frac{1}{6}$ s of silence. This test signal was then convolved with the individually measured and synthesized HRTFs and filtered individually with the inverse individual HPTF and presented via headphones.

Ten subjects, five of them with extensive experience with binaural psychoacoustical experiments, participated in the experiment. Participants were instructed to rate the binaural signals generated with synthesized HRTFs of different array sizes with respect to the reference (binaural reproduction with original HRTFs). Subjects performed the ratings in three separate sessions for three different aspects: spectral coloration, localization, and overall performance, giving their ratings on a continuous scale between bad, poor, fair, good, and excellent. In order to limit the total number of repetitions of the experiment to a feasible amount, five directions for the virtual source

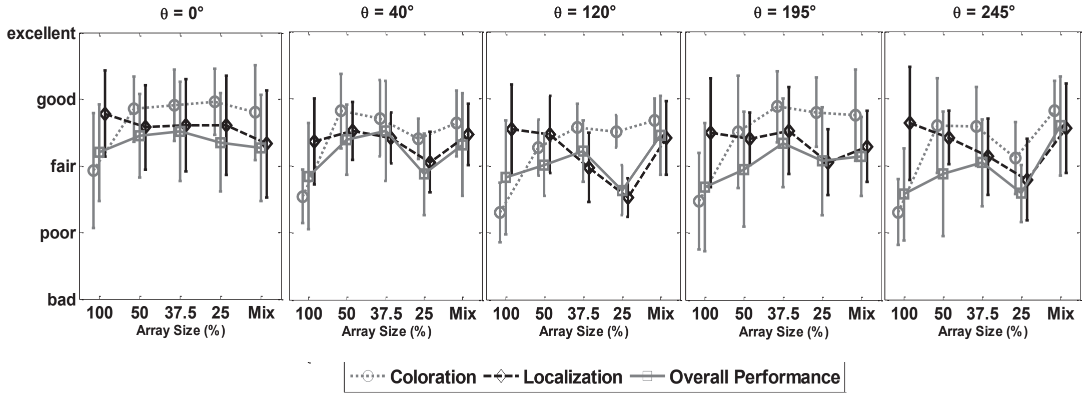

Fig. 3: Mean evaluations over all 10 subjects regarding different array sizes (x-axis) for different aspects and source directions.

were chosen, including two cases with the largest and two cases with small variations in the resulting ILD deviation caused by different array sizes (θ = 120°, 245° and θ = 40°, 195°), and the frontal direction (θ = 0°). Each direction appeared three times for each given aspect. The directions were presented in a randomized order.

PERCEPTUAL EVALUATION – RESULTS AND DISCUSSION

The results of the perceptual evaluations over all subjects with regard to different aspects and source direction are shown in Fig. 3 as mean and standard deviation across subjects. As can be seen, the results vary not only with different array sizes and source directions but also with subjects which can be due to different internal scales used by subjects or individual differences in the HRTFs.

According to the perceptual results, the mean evaluations of spectral coloration improved generally for a smaller array. This effect could be noticed at all of the five tested directions. Objective results (see Fig. 2 for one participant as an example) had shown that for a reduced-size array spectral distortions at the contralateral side start to get prominent in the mid-frequency range of 1 to 4 kHz, whereas at this frequency range the spectral distortion for the largest array (100%-size array) remained within the allowed range (defined in Eq. 3), however increased drastically above ca. 4 kHz. Considering the two frequency ranges $1\,\text{kHz} \le f \le 4\,\text{kHz}$ and $4\,\text{kHz} \le f \le 16\,\text{kHz}$, the ratings for spectral coloration of all participants vs. the absolute spectral distortion averaged across frequency for these two frequency ranges are shown in Fig. 4a at θ = 245° as an example. It seems that the increased spectral distortion of the smaller arrays at mid-frequency range did not influence the ratings on spectral coloration (Fig. 4a, top). Moreover, the prominent increased spectral distortion of the largest array at frequencies above 4 kHz coincided with the generally reduced coloration ratings for this array (Fig. 4a, bottom).

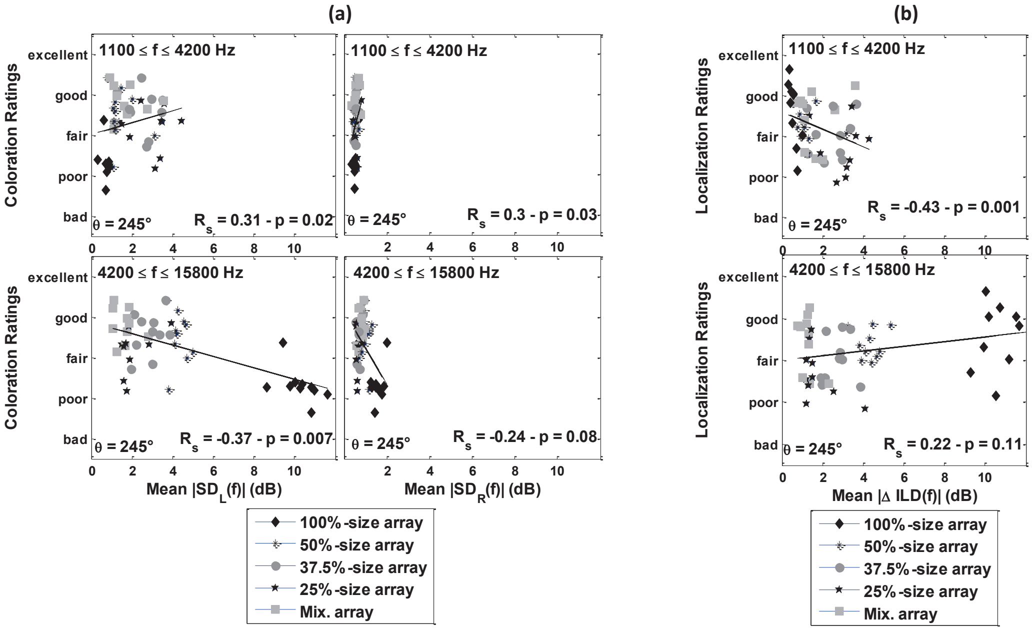

Fig. 4: (a) Evaluations on spectral coloration vs. mean spectral distortion at the left and right ears. (b) Evaluations on localization vs. mean ILD deviations. $\theta = 245°$, $1 \leq f \leq 4$ kHz (upper row) and $4 \leq f \leq 16$ kHz (lower row). Mean $|.(f)|$ = average over the depicted frequency range.

Smaller array sizes led in average to decreased localization ratings especially at $\theta = 120°$ and $245°$. The increased ILD deviation in the mid-frequency range was apparently more relevant for the decline in the ratings for smaller arrays (Fig. 4b, top), whereas the effect of ILD deviations in the frequency range $4\,\text{kHz} \leq f \leq 16\,\text{kHz}$ seemed not to be as important for the localization (Fig. 4b, bottom).

The mean evaluations on overall performance lay almost for all cases at the lower edge of coloration and localization ratings (Fig. 3). This could indicate that the overall perception of the synthesis depended both on coloration as well as on localization cues. In other words, the accuracy of synthesis should be preserved both for spectral coloration and localization cues. The ratings on overall performance show a general increase towards arrays with the middle-range size (37.5%-size array or the Mix-array) confirming the compromise between localization artifacts of smaller arrays and coloration artifacts of larger arrays.

In order to analyse whether at least one of the microphone arrays for a fixed direction and perceptual aspect led to significantly different evaluations the Friedman test was applied. The p-values for the given aspect and source direction are listed in Table 1. Considering the Bonferroni correction for the 3 repetitions of each direction, the p-values of conditions indicating a significant effect of the array size ($p \leq \frac{0.05}{3}$) are

	θ = 0°	θ = 40°	θ = 120°	θ = 195°	θ = 245°
Spectral Coloration	**0.0068**	**0.0113**	**0.0003**	**0.0029**	**0.0002**
Localization	0.4435	0.0192	**0.0025**	**0.0146**	**0.0001**
Overall Performance	0.06917	0.0976	**0.0011**	0.1108	**0.0157**

Table 1: p-values according to Friedman test for different source directions and perceptual aspects. p-values indicating a significant effect of array size ($p \leq \frac{0.05}{3} = 0.0167$) are depicted as bold numbers.

shown as bold numbers. According to the results, array size seemed to have a significant effect on the coloration ratings at all of the five considered directions due to the difference in ratings for the 100%-size array compared to the other arrays. Different array sizes seemed to affect the evaluations regarding localization mostly at θ = 120°, 195°, and 245°. The effect was significant due to decreased ratings given to the 25%-size array. The effect of array size on the evaluation of overall performance was significant at θ = 120° and 245°, due to either different ratings given to the 100%-size array or 25%-size array, compared to the other arrays. This confirms that participants chose the spectral and localization cues differently as the critical cue for giving overall ratings.

CONCLUSION

In this study the effect of microphone array size on the accuracy of HRTF synthesis with a virtual artificial head was investigated. Simulation results for five different array sizes (planar arrays with 24 microphones, approximately quadratic with side lengths ranging from 20 cm to 5 cm) indicated that there are noticeable differences in the resulting monaural and binaural features (spectral distortion and ILD deviation) between original and synthesized HRTFs for different array sizes. While spectral distortions especially at the ipsilateral side could be shifted to higher frequencies by choosing a smaller array, spectral distortion increased in the mid-frequency range of $1\,\mathrm{kHz} \leq f \leq 4\,\mathrm{kHz}$ at the contralateral side for smaller arrays, leading to increased ILD deviations at these frequencies. Furthermore, experimental results showed that the array size had a significant effect on the perceived spectral coloration and source localization. In particular, large spectral distortions introduced by the largest array at frequencies above 4 kHz affected the perceived spectral coloration. Contralateral spectral distortions at the mid-frequency range appearing for smaller arrays did not affect the perceived coloration, but led to decreased localization ratings. These ratings presumably resulted from ILD deviations at these frequencies. The overall evaluation of different arrays sizes confirmed the importance of accuracy both with respect to spectral and localization cues. In general, the overall ratings were the highest for microphone arrays of mid-range size (37.5%-size array or the ‘Mix’ combination of 50%- and 25%-size arrays) since the deficiencies of larger and smaller arrays could be balanced for these array sizes. A further investigation should analyze the interaural phase differences resulting from different array sizes, both regarding their perceptual relevance and also regarding an effective incorporation of phase contraints for the minimization of the cost function J_{LS}.

ACKNOWLEDGMENTS

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

REFERENCES

Doclo, S., and Moonen, M. (**2003**). “Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics,” IEEE T. Signal Proces., **51**, 2511-2526. doi: 10.1109/TSP.2003.816885

Fallahi, M., Hansen, M., Doclo, S., van de Par, S., Mellert, V., Püschel, D., and Blau, M. (**2017**). “High spatial resolution binaural sound reproduction using a virtual artificial head,” Fortschritte der Akustik, DAGA 2017, Kiel, Germany, pp. 1061-1064.

Köhler, S., Blau, M., van de Par, S., and Rasumow, E. (**2014**). “Simultane Messung mehrerer HRTFs in nicht reflexionsarmer Umgebung,” Fortschritte der Akustik, DAGA 2014, Oldenburg, Germany, pp. 202-203.

Rasumow, E., Blau, M., Hansen, M., Doclo, S., van de Par, S. Mellert, V., and Püschel, D. (**2011**). “Robustness of virtual artificial head topologies with respect to microphone positioning errors,” Proc. Forum Acusticum, Aalborg, pp. 2251-2256.

Rasumow, E., Blau, M., Hansen, M., van de Par, S., Doclo, S., Mellert, V., and Püschel, D. (**2014**). “Smoothing individual head-related transfer functions in the frequency and spatial domains,” J. Acoust. Soc. Am., **135**, 2012-2025. doi: 10.1121/1.4867372

Rasumow, E., Hansen, M., van de Par, S., Püschel, D., Mellert, V., Doclo, S., and Blau, M. (**2016**). “Regularization approaches for synthesizing HRTF directivity patterns,” IEEE T. Audio Speech, **24**, 215-225. doi: 10.1109/TASLP.2015.2504874

Rasumow, E., Blau, M., Doclo, S., van de Par, S., Hansen, M., Püschel, D., and Mellert, V.(**2017**). “Perceptual evaluation of individualized binaural reproduction using a virtual artificial head,” J. Audio Eng. Soc., **65**, 448-459. doi: 10.17743/jaes.2017.0012

Ward, D.B., Kennedy, R.A., and Williamson, R.C. (**2001**). “Constant directivity beamforming,” in *Microphone Arrays, Signal Processing Techniques and Applications*, Eds. M. Brandstein and D. Ward (Berlin Heidelberg: Springer-Verlag), pp. 3-18.