Audio Engineering Society

Conference Paper

Presented at the AES 2026 International Conference on
Audio for Virtual and Augmented Reality and Immersive Games
2026 June 30 – July 3, Paris, France

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Effect of the template on the predicted sagittal-plane sound-localization performance

Felix Perfler[1] and Piotr Majdak[1]

[1]*Acoustics Research Institute, Austrian Academy of Sciences, Dominikanerbastei 16, 1010, Vienna, Austria*

Correspondence should be addressed to Felix Perfler (felix.perfler@oeaw.ac.at)

ABSTRACT

In the state-of-the art models of sagittal-plane sound localization, template head-related transfer functions (HRTFs) are used to reflect the listener's internal calibration of auditory space decoding, and thus determine the prediction quality. The effect of the template HRTFs has not yet been investigated directly. Here, a model was calibrated separately to two HRTF measurements of the same listeners and its predictions were compared to behavioral localization responses of these listeners obtained in three listening conditions: two acoustically measured HRTF sets (those used for model calibration), and an additional condition (unseen during the calibration) used to test the model's ability to generalize. We analyzed the quadrant error rates (QE) and local polar errors (PEs) from eight listeners. The predicted errors were similar in both calibration conditions and increased in the unseen condition. The quality of the predictions, however, varied significantly with the template, more for PE than for QE, slightly preferring one template over the other when predicting the unseen condition. Our findings suggest that small differences in HRTFs used for the template may influence the prediction quality, especially when applied to unseen listening conditions.

1 Introduction

Accurate localization in the sagittal plane is strongly shaped by monaural spectral cues introduced by the torso, head, and pinnae [21, 13, 1]. These direction-dependent spectral patterns are represented by head-related transfer functions (HRTFs), through which essential information for estimating sound-source position along the sagittal planes [18] is provided, especially at higher frequencies where the effect of the pinna is most pronounced [1, 14, 9].

Computational localization models are important because they enable rapid evaluation of expected perceptual consequences without running a full behavioral localization experiment for every listener, stimulus set, or signal-processing variant. This is particularly relevant because behavioral experiments remain expensive and time-consuming, while recent work has accelerated the development of HRTF personalization methods based on visual anatomy, anthropometric cues, feature prediction, parametric pinna representations, and 3D geometry [2, 3, 17, 24, 16] and of HRTF upsampling methods designed to reconstruct dense directional sampling from sparse measurements [8, 15, 12]. In addition, large community resources such as the SONICOM HRTF dataset and its recent extension have made data-driven benchmarking increas-

ingly practical [6, 19]. In this context, auditory models provide a useful first-pass benchmark for new HRTF individualization and upsampling algorithms before committing to extensive perceptual testing, and they complement emerging community benchmarks such as the SONICOM / IEEE Listener Acoustic Personalisation (LAP) Challenge [5].

Computational models of sagittal-plane localization aim to capture the physical and neural auditory mechanisms underlying sound localization. In the sagittal-plane sound-localization model [4], positive spectral gradients are extracted from target and template HRTFs and converted into response probabilities across polar angle. By linking neurophysiological evidence for spectral-gradient coding [20, 23], from which perceptual metrics such as quadrant error rate (QE) and local polar error (PE) are derived, a useful framework for the study of cue-dependent localization is provided.

The model assumes that the template HRTF used for comparison matches the listener's internal template. This assumption is broadly consistent with evidence that localization can degrade with nonindividualized HRTFs and adapt when spectral cues are systematically altered [22, 7]. What remains unclear is whether model predictions depend on the choice of template. This question is well addressed by a dataset [11], which includes, per subject, two acoustically measured HRTF sets (**HRTF A**, dataset label **binOwn**, and **HRTF B**, dataset label **binOwnB**) and loudspeaker crosstalk-cancellation playback driven by the second measurement (**HRTF X**, dataset label **ctcOwnB**).

With that dataset, two template-specific model variants were compared: one calibrated with **HRTF A** as template and one calibrated with **HRTF B**. Both variants were then evaluated across all three target conditions against behavioral data. In this design, the contribution of template selection to the localization performance predicted by the model is isolated.

2 Methods

2.1 Sagittal-Plane Localization Model

The sagittal-plane localization model [4] was used. Response probabilities over polar angle are predicted by the model from monaural spectral gradients of target and template directional transfer functions.

For each frequency band b, the log-power spectrum in dB is

$$\xi[b] = 10\log_{10}\left(\frac{1}{N}\sum_{n}(x * g_b)^2[n]\right), \tag{1}$$

where x is the signal and g_b is the b-th gammatone filter. Positive spectral gradients are extracted as

$$\tilde{\xi}[b] = \max(\xi[b] - \xi[b-1], 0). \tag{2}$$

Template-target dissimilarity for polar angle θ is computed with an L1 distance,

$$d[\theta] = \sum_{b=2}^{N_b} |\tilde{\xi}_{\text{template}}[\theta, b] - \tilde{\xi}_{\text{target}}[b]|, \tag{3}$$

and mapped to similarity via

$$\varsigma[\theta] = 1 - \left(1 + e^{-\Gamma(d[\theta]-S_l)}\right)^{-1}, \tag{4}$$

with free parameters sensitivity S_l and selectivity Γ. Binaural combination follows [4], and sensorimotor scatter was fixed to $\varepsilon = 17^\circ$.

2.2 Dataset and Conditions

Anonymized data from eight normal-hearing listeners (NH12, NH14, NH15, NH57, NH62, NH64, NH68, NH72) from [11] were analyzed. Three conditions were used. Here, "matched" denotes playback using the listener's own measured HRTFs for the respective condition:

- **HRTF A**: matched binaural playback using the primary acoustic HRTF measurement.
- **HRTF B**: matched binaural playback using a second acoustic HRTF measurement of the same listener.
- **HRTF X**: loudspeaker playback using crosstalk cancellation based on the second individualized HRTF measurement, corresponding to the **ctcOwnB** condition in [11].

Details of the loudspeaker setup, crosstalk-cancellation processing, and regularization are described in [11]. All HRTFs were converted to directional-transfer functions (DTFs).

2.3 Calibration Protocol for Template-Effect Analysis

To isolate template effects, listener-specific sensitivity was calibrated separately for two template-specific models, while $\Gamma = 6$ and $\varepsilon = 17°$ were fixed:

- **Model A**: template = **HRTF A**, fit to **HRTF A** behavioral data.
- **Model B**: template = **HRTF B**, fit to **HRTF B** behavioral data.

For both models, sensitivity was estimated with a listener-wise grid search over $S \in [0.01, 2.00]$ in steps of 0.01 using the same objective combining QE and PE as well as the same chance normalization as in [4].

2.4 Hypothesis and Statistical Analysis

We hypothesized that localization performance would be largely invariant to acoustic template choice, consistent with the behavioral results from [11]. This hypothesis was tested by means of the predicted QEs and PEs as well as their deviations from those obtained in the behavioral test,

$$\Delta\mathrm{QE}_{l,c} = \mathrm{QE}^{\mathrm{model}}_{l,c} - \mathrm{QE}^{\mathrm{behavioral}}_{l,c}, \quad (5)$$

$$\Delta\mathrm{PE}_{l,c} = \mathrm{PE}^{\mathrm{model}}_{l,c} - \mathrm{PE}^{\mathrm{behavioral}}_{l,c}, \quad (6)$$

for listener l and condition c. Inferential testing was performed with one-factor repeated-measures ANOVAs across target conditions for the behavioral data and with 2×3 repeated-measures ANOVAs (TemplateModel $\times$ TargetCondition) for the model predictions and errors, followed by post hoc model contrasts within each target condition using MATLAB `multcompare`. Post hoc ps were Holm-corrected across the three conditions for each metric.

3 Results

3.1 Model Calibration

Distinct sensitivity sets for Model A and Model B were obtained by template-specific fitting. For the eight analyzed listeners (NH12, NH14, NH15, NH57, NH62, NH64, NH68, and NH72), the sensitivities reported by [4] were $S^{\mathrm{ref}} = \{0.25, 0.58, 0.55, 0.96, 0.98, 0.84, 0.75, 0.79\}$. Relative to these sensitivities, no exact match was obtained for any listener. This difference was expected because different behavioral data were used in the present optimization than in the calibration of [4]. The resulting sensitivities were $S^A = \{0.28, 0.62, 0.63, 1.17, 0.60, 0.60, 0.53, 0.83\}$ and $S^B = \{0.01, 0.60, 0.52, 1.13, 0.50, 0.73, 0.77, 0.68\}$ in the same listener order.

3.2 Group Analysis

Model performance was first inspected descriptively across the three conditions. The localization responses of exemplary listener NH64 are shown in Figure 1.

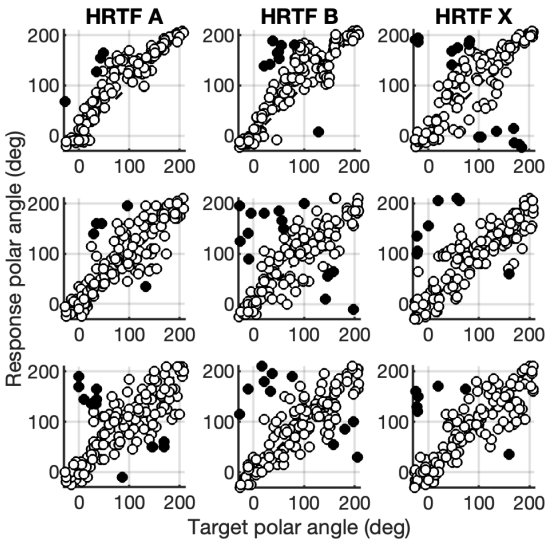


Fig. 1: Localization responses of a listener as an example (NH64). The columns show responses when listening with **HRTF A**, **HRTF B**, and **HRTF X**, respectively. Top row: Responses from the behavioral test. Middle row: Responses predicted with Model A. Bottom row: Responses predicted with Model B. Local responses are indicated by open circles; responses outside the $\pm90°$ range are indicated by filled circles.

Behavioral and predicted QEs and PEs are summarized in Figure 2, and the corresponding behavioral-vs-model scatter plots are shown in Figure 3.

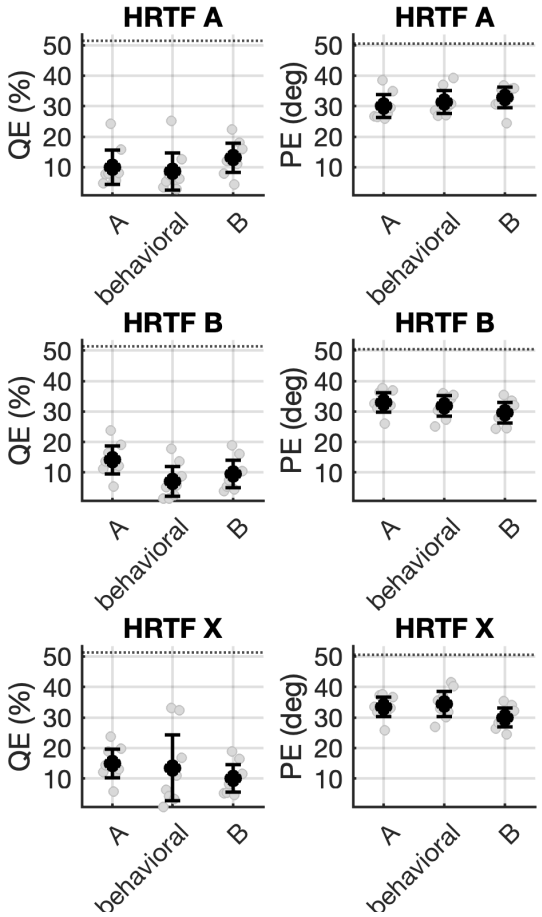


Fig. 2: Localization errors (QE and PE) of all ($N = 8$) listeners. The rows show conditions when listening to **HRTF A**, **HRTF B**, and **HRTF X**, respectively. Each panel shows predicted errors (Models A and B) and errors from the behavioral test. Gray symbols show listener-specific errors. Black symbols show means ±95% CI across all listeners. The dotted horizontal lines show the chance performance (QE = 51.4% and PE = 50.5°).

Across the three target conditions, only modest changes were observed in the behavioral errors (consistent with [11]), whereas a stronger dependence on the selected template was observed in the model predictions, with the clearest separation appearing for PE. Beyond the matched-template cases, the non-fitted conditions were more informative for generalization. For the two acoustic cross-measurement cases, no uniform advantage was seen: for **HRTF A**, behavioral means of 8.6% QE and 31.3° PE were compared with 13.1% QE and 32.8° PE from Model B, whereas for **HRTF B**, behavioral means of 7.1% QE and 31.8° PE were compared with 14.1% QE and 33.0° PE from Model A. Thus, Model B in **HRTF A** was closer to behavior for QE, whereas Model A in **HRTF B** was slightly closer for PE. The most informative generalization case was **HRTF X**, because neither model was fitted to that condition. There, behavioral means of 13.5% QE and 34.5° PE were compared with 14.9% QE and 33.5° PE from Model A and with 10.1% QE and 30.0° PE from Model B, so the direct means were closer to behavior for Model A. The corresponding Pearson correlations in Figure 3 were nevertheless higher for the Model B→**HRTF A** case than for the Model A→**HRTF B** case (QE: $r = 0.75$ vs. $r = 0.63$; PE: $r = 0.73$ vs. $r = 0.58$), and in **HRTF X** they were again higher for Model B (QE: $r = 0.97$ vs. $r = 0.51$; PE: $r = 0.79$ vs. $r = 0.40$). Taken together, these descriptive comparisons showed that mean closeness to the behavioral data and preservation of listener-wise differences did not favor the same model in every condition.

3.3 Statistical Error Analysis

The behavioral errors (QEs and PEs) did not differ significantly across target conditions (QE: $F(2,14) = 3.100$, $p = 0.106$; PE: $F(2,14) = 3.692$, $p = 0.064$), consistent with [11]. For the model predictions, significant TemplateModel × TargetCondition interactions were found for QE ($F(2,14) = 19.411$, $p = 0.0029$, $\eta_p^2 = 0.735$) and PE ($F(2,14) = 14.301$, $p = 0.0064$, $\eta_p^2 = 0.671$). The difference between Model A and Model B was significant for PE in all three conditions (all $p_{\text{Holm}} = 0.0495$), but not for QE (all $p_{\text{Holm}} = 0.0724$). Because Model A was fitted to **HRTF A** behavioral data and Model B was fitted to **HRTF B** behavioral data, the matched conditions were not the most informative test of generalization. In the non-fitted conditions, no single model generalized better across all comparisons.

Error differences were then inspected to determine whether the template effect reflected systematic over- or underestimation relative to behavior (Figure 4).

No single global over- or underestimation was observed. For PE, Model A underestimated behavior in **HRTF A** and **HRTF X** but overestimated it in **HRTF**

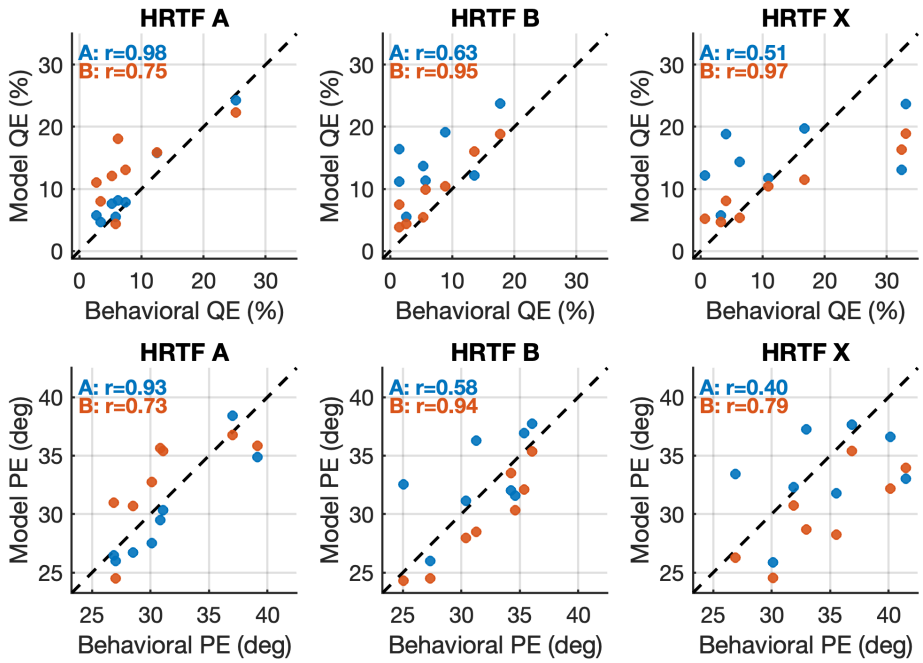


Fig. 3: Behavioral-vs-model scatter plots for QE and PE across listeners and HRTF conditions. Each point represents one listener. The panels compare behavioral localization errors with the corresponding predictions from Model A and Model B. Pearson correlation coefficients are shown in the individual panels.

B, whereas Model B overestimated behavior in **HRTF A** and underestimated it in **HRTF B** and **HRTF X**. For QE, both models overestimated behavior in **HRTF A** and **HRTF B**, while in **HRTF X** Model A still overestimated behavior but Model B underestimated it. Thus, the direction of the bias depended on target condition, and this pattern was clearer for PE than for QE.

Taken together, a mismatch between behavioral stability and model sensitivity to template choice was shown by these analyses. Whereas comparatively stable behavioral QE and PE were observed across **HRTF A**, **HRTF B**, and **HRTF X**, the model predictions showed condition-dependent template effects. Template choice affected PE more consistently than QE, and the error analysis showed that this mainly reflected condition-dependent over- and underestimation rather than a uniform advantage of one template. One possible explanation for the higher Pearson correlations often obtained with Model B, despite the lack of a consistent advantage in mean closeness to behavior, is that the later **HRTF B** measurement benefited from improvements in the measurement procedure. If so, the present results may indicate that the sagittal-plane localization model [4] is sensitive to small HRTF differences that preserve listener-wise structure but have only limited consequences for mean behavioral performance. This interpretation remains speculative and would need to be tested directly.

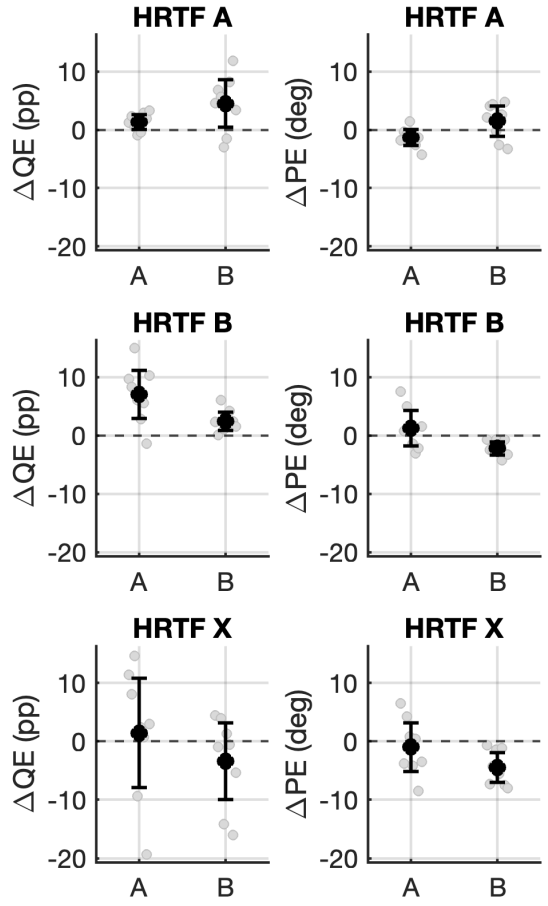


Fig. 4: Error differences between predicted and behavioral localization errors. Rows show **HRTF A**, **HRTF B**, and **HRTF X**; columns show QE and PE. Gray symbols indicate individual listeners; black symbols show means ±95% CI.

4 Conclusion

We investigated how the choice of an HRTF set as a template affects the predictions of the sagittal-plane localization model [4]. We calibrated the model to two HRTF sets of the same listener and compared the predictions to the localization responses obtained with these two sets. In order to see how the models generalize to conditions not involved in the model calibration, we also predicted responses for a third listening condition. In our study, we used the HRTFs and behavioral data from [11].

Although the behavioral QEs and PEs were similar across the two HRTF sets, the model predictions varied with the choice of the HRTF set as the template. The differences were clearer for PE than for QE. When comparing the predictions obtained for the third listening condition, we did not find a clear advantage for either template. However, when considering only the correlation with the behavioral data, Model B generalized better, especially in the unseen HRTF X condition. This suggests that small differences in the template HRTFs may influence model predictions, especially when applied to unseen listening conditions.

Our results demonstrate that the predictive power of the sagittal-plane localization model depends on the HRTFs selected as the template (and thus, the calibration procedure). Further work is required to better understand how HRTF-template effects generalize to unseen sounds used to simulate localization experiments. Further work is also needed to determine how the calibration procedure can be improved, for example by integrating multiple HRTF measurements considered as the template.

5 Funding

F. Perfler is supported by a DOC Fellowship (A 27288) of the Austrian Academy of Sciences at the Acoustics Research Institute. Partial funding was received from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017743 (project "SONICOM" – Transforming auditory-based social interaction and communication in augmented/virtual reality (AR/VR)).

6 Data Availability

The localization data are from [11] and are available in the Auditory Modeling Toolbox (AMT). Code to reproduce the analyses is provided as `exp_perfler2026` in the AMT 1.7.

References

[1] Algazi, V. R., Avendano, C., & Duda, R. O. (2001). Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3), 1110–1122. doi: 10.1121/1.1349185.

[2] Huang, X., Wang, Y., Liu, Y., Ni, B., Zhang, W., Liu, J., & Li, T. (2023). AudioEar: Single-view ear reconstruction for personalized spatial audio. *Proceedings of the AAAI Conference on Artificial Intelligence*. doi: 10.48550/arXiv.2301.12613.

[3] Arbel, L., Ananthabhotla, I., Ben-Hur, Z., Alon, D. L., & Rafaely, B. (2024). On HRTF notch frequency prediction using anthropometric features and neural networks. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi: 10.1109/ICASSP48485.2024.10447610.

[4] Baumgartner, R., Majdak, P., & Laback, B. (2014). Modeling sound-source localization in sagittal planes for human listeners. *The Journal of the Acoustical Society of America*, 136(2), 791–802. doi: 10.1121/1.4887447.

[5] Geronazzo, M., Picinali, L., Hogg, A. O. T., Barumerli, R., Poole, K. C., Daugintis, R., Pauwels, J., Ntalampiras, S., McLachlan, G., & Brinkmann, F. (2025). Technical report: SONICOM / IEEE listener acoustic personalisation (LAP) challenge - 2024. *TechRxiv*. doi: 10.36227/techrxiv.173153187.72930965/v2.

[6] Engel, I., Daugintis, R., Vicente, T., Hogg, A. O. T., Pauwels, J., Tournier, A. J., & Picinali, L. (2023). The SONICOM HRTF dataset. *Journal of the Audio Engineering Society*, 71(5), 241–253. doi: 10.17743/jaes.2022.0066.

[7] Hofman, P. M., Van Riswick, J. G. A., & Van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature Neuroscience*, 1(5), 417–421. doi: 10.1038/1633.

[8] Hogg, A. O. T., Jenkins, M., Liu, H., Squires, I., Cooper, S. J., & Picinali, L. (2024). HRTF upsampling with a generative adversarial network using a gnomonic equiangular projection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 2085–2099. doi: 10.1109/TASLP.2024.3375635.

[9] Langendijk, E. H. A., & Bronkhorst, A. W. (2002). Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4), 1583–1596. doi: 10.1121/1.1501901.

[10] Majdak, P., Goupell, M. J., & Laback, B. (2010). 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Attention, Perception, & Psychophysics*, 72(2), 454–469. doi: 10.3758/APP.72.2.454.

[11] Majdak, P., Masiero, B., & Fels, J. (2013). Sound localization in individualized and non-individualized crosstalk cancellation systems. *The Journal of the Acoustical Society of America*, 133(4), 2055–2068. doi: 10.1121/1.4792355.

[12] Masuyama, Y., Wichern, G., Germain, F. G., Ick, C., & Le Roux, J. (2025). Retrieval-augmented neural field for HRTF upsampling and personalization. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi: 10.48550/arXiv.2501.13017.

[13] Middlebrooks, J. C. (1999). Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America*, 106(3), 1493–1510. doi: 10.1121/1.427147.

[14] Middlebrooks, J. C. (1992). Narrow-band sound localization related to external ear acoustics. *The Journal of the Acoustical Society of America*, 92(5), 2607–2624. doi: 10.1121/1.404400.

[15] Masuyama, Y., Wichern, G., Germain, F. G., Pan, Z., Khurana, S., Hori, C., & Le Roux, J. (2024). NIIRF: Neural IIR filter field for HRTF upsampling and personalization. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. doi: 10.48550/arXiv.2402.17907.

[16] Niu, R., Koyama, S., & Nakamura, T. (2025). Head-related transfer function individualization using anthropometric features and spatially independent latent representation. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. doi: 10.48550/arXiv.2508.16176.

[17] Perfler, F., Pausch, F., Pollack, K., Holighaus, N., & Majdak, P. (2025). Parametric model of the human pinna based on Bézier curves and concave deformations. *Computers in Biology and Medicine*, 188, 109817. doi: 10.1016/j.compbiomed.2025.109817.

[18] Pollack, K., Kreuzer, W., & Majdak, P. (2022). Modern Acquisition of Personalised Head-Related Transfer Functions: An Overview. In *Advances in Fundamental and Applied Research on Spatial Audio*, edited by Brian FG Katz and Piotr Majdak. IntechOpen. doi: 10.5772/intechopen.102908.

[19] Poole, K. C., Meyer, J., Martin, V., Daugintis, R., Marggraf-Turley, N., Webb, J., Pirard, L., La Magna, N., Turvey, O., & Picinali, L. (2025). The extended SONICOM HRTF dataset and Spatial Audio Metrics toolbox. *arXiv*. doi: 10.48550/arXiv.2507.05053.

[20] Reiss, L. A. J., & Young, E. D. (2005). Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus. *Journal of Neuroscience*, 25(14), 3680–3691. doi: 10.1523/JNEUROSCI.4963-04.2005.

[21] Shaw, E. A. G. (1974). Transformation of sound pressure level from the free field to the eardrum in the horizontal plane. *The Journal of the Acoustical Society of America*, 56(6), 1848–1861. doi: 10.1121/1.1903522.

[22] Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the*

Acoustical Society of America, 94(1), 111–123. doi: 10.1121/1.407089.

[23] Zakarauskas, P., & Cynader, M. S. (1993). A computational theory of spectral cue localization. *The Journal of the Acoustical Society of America*, 94(3), 1323–1331. doi: 10.1121/1.408160.

[24] Zhao, J., Yao, D., Gu, J., & Li, J. (2024). Efficient prediction of individual head-related transfer functions based on 3D meshes. *Applied Acoustics*, 219, 109938. doi: 10.1016/j.apacoust.2024.109938.