Audio Engineering Society

# Convention Paper

Presented at the 148th Convention
2020 May 25 – 28, Vienna, Austria



# Predicting Directional Sound-Localization of Human Listeners in both Horizontal and Vertical Dimensions

Roberto Barumerli[1], Piotr Majdak[2], Jonas Reijniers[3], Robert Baumgartner[2], Michele Geronazzo[4], and Federico Avanzini[5]

[1]*Dept. of Information Engineering, University of Padova, Italy*
[2]*Acoustics Research Institute, Austrian Academy of Sciences, Austria*
[3]*Dept. of Engineering Management, University of Antwerp, Belgium*
[4]*Dept. of Architecture, Design, and Media Technology, Aalborg University, Denmark*
[5]*Dept. of Computer Science, University of Milano, Italy*

Correspondence should be addressed to Roberto Barumerli (barumerli@dei.unipd.it)

**ABSTRACT**

Measuring and understanding spatial hearing is a fundamental step to create effective virtual auditory displays (VADs). The evaluation of such auralization systems often requires psychoacoustic experiments. This process can be time consuming and error prone, resulting in a bottleneck for the evaluation complexity. In this work we evaluated a probabilistic auditory model for sound localization intended as a tool to assess VAD's abilities to provide static sound-localization cues to listeners. The outcome of the model, compared with actual results of psychoacoustic experiments, shows the advantages and limitations of this systematic evaluation.

## 1 Introduction

Computational auditory models, describing the hearing system from the periphery up to a certain cognitive function, can fundamentally enhance the understanding of how the hearing system senses the acoustic environment [1] and be applied to effectively assess spatial qualities provided by virtual auditory displays (VADs). From the vast inventory of various spatial audio qualities (see SAQI, [2]), a basic and best understood one is the perception of sound direction. In VADs, proper directional sound-localization cues are important because they enable users' correct orientation and navigation. Evaluation of those cues can be done through psychoacoustic experiments, which are usually time-consuming and, thus, limited to small sample sizes. In order to obtain a faster VAD evaluation for a larger population, auditory localization models can be applied [3, 4]. Existing models separately investigate sound localization either in the horizontal [5] or vertical [6] dimension, showing insights into the processes underlying each dimension. Bayesian inference is a promising tool to more formally model the neural uncertainty jointly affecting both dimensions, and thus to more precisely replicate the actual human performance found in auditory localization experiments. Hence, the aim of this study is to evaluate a Bayesian modeling approach [7]

by comparing its predictions with psychoacoustic results [8] uncovering and demonstrating under which conditions the model can provide useful applications. While the authors of the original model stated that their work focused on the *what* rather than the *how*, this manuscript also tries to explain how the model formulation can be linked to the behavioral components of the human sound-localization process. This effort is reported in Section 2. The Section 3 explain how the model has been tested and Section 4 closes with the results and a discussion of the advantages and pitfalls to adopt this model as a tool to parametrize the human behavior.

## 2 Methods

We implemented the model proposed in [7]. While the original work reports in great detail the mathematical formulation, our implementation had to be based on some assumptions that can be slightly different from the original because we attempted to relate their formulation with physiological and psychoacoustic grounds. Our implementation of the model is available in the Auditory Modeling Toolbox (AMT)[1].
The model aims to extract the azimuth and elevation angle $\theta = [\alpha, \varepsilon]$ from the acoustic field by following a template matching procedure, as assumed being implemented in the human brain [9]. Based on this hypothesis *internal templates* for each of the available directions are constructed. These *templates* allow the computation of a distance between them and the *internal realization* of the sound under evaluation. The *decision stage* estimates the sound's incidence angle based on these distances.
The model comprises four critical components: (i) the feature space; (ii) the internal noise; (iii) the internal templates; and (iv) the decision stage.

### 2.1 Feature space

The feature space approximates the neural representation of the acoustic input. Prior to the computation of the actual features, our implementation converted the angular grid to be uniformly sampled by using the spherical-harmonics interpolation with Tikonov regularization. The feature space (see Eqs. 2) aggregates the interaural time difference (ITD, Eq. 2a) with the sum of the log-spectra of both the head-related impulse responses (HRIRs), $\mathbf{H}_{L,R}$, and the source, $\mathbf{S}$, (Eqs. 2b and 2c). The method for the ITD and log-spectra computation methods were partially reported in the original paper. To compute $\mathbf{T}^{\varphi}_{itd}$, we adopted the threshold method with a tenth-order, low-pass Butterworth filter, $fc = 4\ kHz$ and activation level of $-10\ dB$ [10]. The ITD is then converted into the just noticeable difference (JND) in order to estimate its noise distribution from pyschoacoustic data. Finally, to compute the log-spectral magnitudes we applied an all-pole implementation of the Gammatone filterbank [11], with 30 frequency channels each separated by 1 equivalent rectangular bandwidth (ERB)[12] within $[0.3, 15]$ kHz. The computed magnitudes are then limited to a minimal value, according to the equal loudness curves, approximating the absolute hearing threshold.

$$\mathbf{T}_{\varphi} = [\mathbf{T}^{\varphi}_{itd}, \mathbf{T}^{\varphi}_{-}, \mathbf{T}^{\varphi}_{+}] \tag{1}$$

$$\mathbf{T}^{\varphi}_{itd} = itd(\varphi) \tag{2a}$$

$$\mathbf{T}^{\varphi}_{-} = \mathbf{H}_L(\varphi) - \mathbf{H}_R(\varphi) \tag{2b}$$

$$\mathbf{T}^{\varphi}_{+} = \mathbf{S} + [\mathbf{H}_L(\varphi) + \mathbf{H}_R(\varphi)]/2 \tag{2c}$$

### 2.2 Internal Realization

When listening to a sound source, our hearing system transforms the acoustic input with limited precision [13]. Furthermore, each subject shows individual variations in how efficiently the internal information is used [14]. The model addresses these uncertainties by adding an ensemble of stochastic noise sources to the computed feature. Eq. 3 reports the case for a sound source with the direction of arrival $\theta$. The component $\delta$ is expanded in Eq. 4. These uncertainties have different meaning but all of them are assumed to be Gaussian distributed with zero mean. The quantification of the variance of each element was derived from the psychoacoustic literature, if available, or set manually. Here follows a description of each noise source:

- $\delta_{itd}$: error of the ITD measurement. Derived from a psychoacoustic ITD-JND experiment and represented by a Gaussian distribution with zero mean and variance $\sigma^2_{itd}$.

[1] see http://amtoolbox.sourceforge.net

- $\delta_I$: error of the hearing system to measure spectral magnitudes. Left and right channels are assumed to be equal. This noise source is represented by a multivariate Gaussian distribution with diagonal covariance matrix $\Sigma_-$. The constant $\sigma_I^2$ was derived from an intensity discrimination experiment with broad band noise. .
- $\delta_S$: error of the source's imperfect estimation. Composed by two different noise generators: the first, with variance $\sigma_S^2$, models the subject's memory on the sound source spectra, the second, with variance $\sigma^2$, mimics the cross-talk between the adjacent element of the filter-bank. $\sigma_S$ was imposed to be the same as in $\delta_I$ and $\sigma^2$ was chosen arbitrarily.

$$\mathbf{X}_\theta[\delta] = \mathbf{T}_\theta + \delta \quad \text{with} \quad \delta = [\delta_{itd}, \delta_-, \delta_+] \tag{3}$$

$$\delta_{itd} \sim \mathcal{N}(0, \sigma_{itd}^2) \tag{4a}$$
$$\delta_- \sim \mathcal{N}(0, \Sigma_-), \quad \Sigma_- = 2\sigma_I^2 \cdot \mathbf{I} \tag{4b}$$
$$\delta_+ \sim \mathcal{N}(0, (\Sigma_+), \quad \Sigma_+ = (\sigma_I^2/2 + \sigma_S^2) \cdot \mathbf{I} + \sigma^2 \cdot 1) \tag{4c}$$

### 2.3 Internal Templates

After the computation of all the available templates (Eq. 1) and the internal realization (Eq. 3), the model represents the internal belief by relying on the Bayes' theorem. The elaboration starts by calculating the likelihood (Eq. 5) for every template or direction, $\varphi$, given the internal realization $\mathbf{X}_\theta[\delta]$. Since the likelihood can lead to a biased estimate, the dependence of the single template is removed by computing the a-posteriori probabilities (Eq. 7). While the denominator, $P(\mathbf{X}_\theta[\delta])$, can be assumed as a constant factor, the prior probability, $P(\varphi)$ is stated to be uniformly distributed, or in other words, every direction is equally probable. A visual example is shown in Fig. 1.

$$P(\mathbf{X}_\theta[\delta]|\varphi) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \times \exp\left\{ -\frac{1}{2} (\mathbf{X}_\theta[\delta] - \mathbf{T}_\varphi)^T \Sigma^{-1} (\mathbf{X}_\theta[\delta] - \mathbf{T}_\varphi) \right\} \tag{5}$$

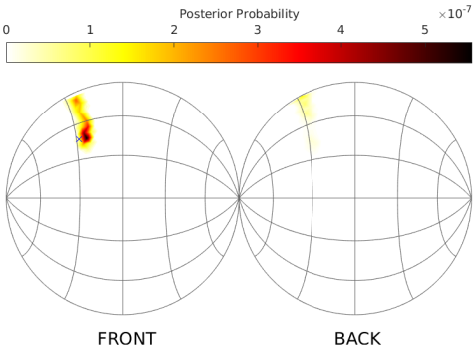


**Fig. 1:** Simulated internal belief of a localization experiment for the target direction $\theta = (-34°, 45°)$. The symbol $\times$ reports the true direction.

$$\Sigma = \begin{bmatrix} \sigma_{itd}^2 & 0 & 0 \\ 0 & \Sigma_- & 0 \\ 0 & 0 & \Sigma_+ \end{bmatrix} \tag{6}$$

$$P(\varphi|\mathbf{X}_\theta[\delta]) = \frac{P(\mathbf{X}_\theta[\delta]|\varphi) P(\varphi)}{P(\mathbf{X}_\theta[\delta])} \tag{7}$$

### 2.4 Decision Stage

As the last step, the model adopts the maximum a-posteriori (MAP) estimator (Eq. 8) to extract the estimated azimuth and elevation angles from the internal templates.

$$\hat{\varphi} = \arg\max_\varphi P(\varphi|\mathbf{X}_\theta[\delta]) \tag{8}$$

## 3 Experiments

In [7], the model was evaluated by reporting graphically the mean spherical error and the local bias, for various conditions. The analysis covers different aspects of the model, i.e. different subjects and different values of internal noises.

### 3.1 Comparison with actual behavioral data

Our model predictions were compared with the outcome of the behavioral localization experiment, [8], for the same conditions[2]. The behavioral experiment tested the performance of five trained listeners on localizing Gaussian white-noise bursts of 500*ms* duration presented from all directions. The listeners were wearing a head-mounted display and asked to manually point at the perceived sound-source direction. For the predictions, every subject has been represented by individual set of directional transfer functions (DTFs). The general match the human performance was done by adjusting the internal noise variances in the model.

### 3.2 Metrics

Our work aims to understand if the model can be adopted to create a virtual version of an actual user. Thus, our implementation translated the spherical estimations into the interaural-polar coordinate system on which human subjects seem to rely on [15]. Furthermore, our experiments were evaluated with the metrics used in [9] and [8]. These are: lateral bias, lateral RMS error, elevation bias, local RMS polar error, and quadrant-error rate.

## 4 Results

### 4.1 Implementation

The predicted localization errors are shown in Fig. 2. These results correspond to Fig. 4 from [7], providing clear evidence for a valid re-implementation. This can also be concluded for other conditions tested in [7], the reproduction of which is provided in the AMT (function `exp_reijniers2014`).

### 4.2 Comparison with behavioral data

Table 1 shows the predicted localization errors for various metrics. These metrics were also used in [8] (their Tab. 4) and are reprinted here. The comparison shows that when using the internal noise as proposed in [7], the predicted errors are consistently smaller than those found in the behavioral experiment [8]. The size of such predictions corresponds to a super-human localization performance.

[2]The acquired dataset has been incorporated into the AMT.

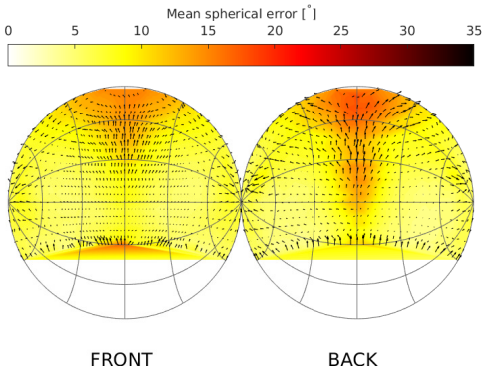


**Fig. 2:** Predicted localization errors for the baseline condition, simulating 100 different subjects localizing 500 signals from each direction of the spherical grid. Superimposed arrows indicate the size and direction of the error averaged across all subjects and signals.

In order to tune the model to provide human-like performance, we had to increase the levels of the internal noises by a factor of three. The corresponding predictions and the noise levels are also shown in Tab. 1. The prediction with the optimized noise levels well correspond to those observed in actual psychoacoustic experiments.

## 5 Discussion

In this work, we reproduced the auditory model presented in [7] and compared its predictions with results of a behavioral experiment presented in [8]. As such a model plays an important role in systematic evaluations of novel methodologies for VADs [3, 4], it is essential to mimic human behavior as precisely as possible. We demonstrate here, however, that the model from [7] outperforms the actual sound-localization performance of humans, yielding as super-human performance (see Tab. 1). This mismatch was also found when comparing the model estimations with other psychoacoustic outcomes (i.e. band limited sources or non-individual HRIR) [16]. Finally, we were able to obtain predicted performance metrics similar to those of humans, but for that had to substantially increased all the model's internal noises.

From the psychoacoustic perspective, the implications of such a modification of model parameters are unclear.

**Table 1:** Comparison between the metrics from [8]. All elements are averaged over all directions.

| Metrics | Actual from [8] | Our predictions | |
|---|---|---|---|
| | | $\delta$ | $3\delta$ |
| Lateral Bias [°] | -0.15 | 0.03 | 0.36 |
| Lateral RMS error [°] | 12.25 | 4.33 | 11.90 |
| Elevation bias [°] | -4.33 | -0.30 | -5.80 |
| Local RMS polar error [°] | 32.73 | 12.95 | 33.51 |
| Quadrant error [%] | 7.83 | 0.91 | 21.40 |

Our current understanding of the processes involved in sound localization might be limited (see the approaches in [1]) and, consequently, it seems to be difficult to determine the best formal approach which can represent the human behavior [17]. Still, the model from [7] sets the foundations to work with the Bayesian framework providing the advantage to combine and compare interdisciplinary knowledge within a non-deterministic but reproducible approach.

In order to more clearly link the processes involved in sound localization with various model stages, the model's assumptions need to be refined. Several issues may play a role. First, the human auditory pathway separates the elaboration of the horizontal and the polar dimensions [15], whereas the model estimates the source position by merging the acoustic information into a binaural sum and difference vector, regardless of the spatial dimensions. Second, while the derivation of the noise variances is mathematically sound (see supplementary material of [7]), our results suggest that it is not trivial to derive such uncertainties. In particular, the internal noise, represented by an ensemble of Gaussian random variables, may not adequately mirror the brain processes involved in sound localization. Third, the exploitation of a-priori knowledge might help because listeners seem to exhibit biases in a sound-localization task [18]. Fourth, a different evaluation of the posterior distribution might help because the MAP estimator defines an ideal observer while listeners probably seem to rely on a heuristic method to estimate the sound direction [18], [6].

## 6 Outlook

Future work may consider a new formulation of the feature space in order to integrate more pyschoacoustic findings, and a listener-specific parametrization in order to account for listener-specific performance. Such extensions, combined with the flexibility of the mathematical framework may, in the future, enable a systematic evaluation of even other spatial qualities, such as sound externalization or distance perception, which are significant to the development of realistic VADs.

## References

[1] Dietz, M., Lestang, J.-H., Majdak, P., Stern, R. M., Marquardt, T., Ewert, S. D., Hartmann, W. M., and Goodman, D. F. M., "A framework for testing and comparing binaural models," *Hearing Research*, 360, pp. 92–106, 2018, ISSN 0378-5955, doi:10.1016/j.heares.2017.11.010.

[2] Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S., "A Spatial Audio Quality Inventory (SAQI)," *Acta Acustica united with Acustica*, 100(5), pp. 984–994, 2014, ISSN 16101928, doi:10.3813/AAA.918778.

[3] Geronazzo, M., Spagnol, S., and Avanzini, F., "Do we need individual head-related transfer functions for vertical localization? The case study of a spectral notch distance metric," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7), pp. 1243–1256, 2018, ISSN 2329-9290, doi:10.1109/TASLP.2018.2821846.

[4] Marelli, D., Baumgartner, R., and Majdak, P., "Efficient approximation of head-related transfer functions in subbands for accurate sound localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7), pp. 1130–1143, 2015, doi:10.1109/TASLP.2015.2425219.

[5] Dietz, M., Ewert, S. D., and Hohmann, V., "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Communication*, 53(5), pp. 592–605, 2011, ISSN 01676393, doi:10.1016/j.specom.2010.05.006.

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

[7] Reijniers, J., Vanderelst, D., Jin, C., Carlile, S., and Peremans, H., "An ideal-observer model of human sound localization," *Biological Cybernetics*, 108(2), pp. 169–181, 2014, ISSN 1432-0770, doi:10.1007/s00422-014-0588-4.

[8] Majdak, P., Goupell, M. J., and Laback, B., "3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Attention, Perception, & Psychophysics*, 72(2), pp. 454–469, 2010, ISSN 1943-3921, 1943-393X, doi:10.3758/APP.72.2.454.

[9] Middlebrooks, J. C., "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *The Journal of the Acoustical Society of America*, 106(3), pp. 1493–1510, 1999, ISSN 0001-4966, doi:10.1121/1.427147.

[10] Katz, B. F. G. and Noisternig, M., "A comparative study of interaural time delay estimation methods," *The Journal of the Acoustical Society of America*, 135(6), pp. 3530–3540, 2014, ISSN 0001-4966, doi:10.1121/1.4875714.

[11] Lyon, R. F., "All-pole models of auditory filtering," *Diversity in auditory mechanics*, pp. 205–211, 1997.

[12] Moore, B. C. J. and Glasberg, B. R., "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, 74(3), pp. 750–753, 1983-09, ISSN 0001-4966, doi:10.1121/1.389861.

[13] Verhulst, S., Altoè, A., and Vasilkov, V., "Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss," *Hearing Research*, 360, pp. 55–75, 2018, ISSN 0378-5955, doi:10.1016/j.heares.2017.12.018.

[14] Majdak, P., Baumgartner, R., and Laback, B., "Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization," *Frontiers in psychology*, 5, p. 319, 2014, doi:0.3389/fpsyg.2014.00319.

[15] Morimoto, M. and Aokata, H., "Localization cues of sound sources in the upper hemisphere," *Journal of the Acoustical Society of Japan (E)*, 5(3), pp. 165–173, 1984, doi:10.1250/ast.5.165.

[16] Barumerli, R., Majdak, P., Baumgartner, R., Geronazzo, M., and Avanzini, F., "Evaluation of a human sound localization model based on Bayesian inference," in *Forum Acusticum*, 2020.

[17] Barumerli, R., Geronazzo, M., and Avanzini, F., "Localization in Elevation with Non-Individual Head-Related Transfer Functions: Comparing Predictions of Two Auditory Models," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2539–2543, IEEE, 2018, ISBN 978-90-827970-1-5, doi:10.23919/EUSIPCO.2018.8553320.

[18] Ege, R., Opstal, A. J. V., and Van Wanrooij, M. M., "Accuracy-Precision Trade-off in Human Sound Localisation," *Scientific Reports*, 8(1), p. 16399, 2018, ISSN 2045-2322, doi:10.1038/s41598-018-34512-6.