Evaluation of a human sound localization model based on Bayesian inference

Roberto Barumerli, Piotr Majdak, Robert Baumgartner, Michele Geronazzo, Federico Avanzini

EVALUATION OF A HUMAN SOUND LOCALIZATION MODEL BASED ON BAYESIAN INFERENCE

Roberto Barumerli[1] **Piotr Majdak**[2] **Robert Baumgartner**[2]
Michele Geronazzo[3] **Federico Avanzini**[4]
[1] Dept. of Information Engineering, University of Padova, Italy
[2] Acoustics Research Institute, Austrian Academy of Sciences, Austria
[3] Dyson School of Design Engineering, Imperial College London, United Kingdom
[4] Dept. of Computer Science, University of Milano, Italy
barumerli@dei.unipd.it

ABSTRACT

In cognitive sciences, Bayesian inference has been effectively applied to describe various aspects of perceptual decision making. In the field of spatial hearing, while most of the sound localization models rely on deterministic methods to predict the perceived directional estimates, few attempts have been made to represent the human sound localization mechanism as a probabilistic process. Here, a Bayesian modeling approach for localization of static sound sources in the acoustic free field sticks out [Reijniers *et al.* (2014), Biol Cybern 108(2):169-81], offering a fully spherical localization estimate for a given binaural stimulus. The present work evaluated the quality of that model by quantitatively comparing its predictions to the actual results obtained in psychoacoustic sound localization studies with static sources. For white noise stimuli, the model showed a similar performance to that obtained in localization experiments. We found, however, a mismatch between the predictions and the actual psychoacoustic results for sound sources with band-limited or rippled spectra or for modified head-related transfer functions. The reasons for the deviations will be discussed and suggestions for potential improvements will be outlined.

1. INTRODUCTION

Auditory models are an interesting approach to formally describe the human hearing system from the periphery up to the central nervous system. Such models deepen the knowledge on how the hearing system processes and elaborates the acoustic information. While the peripheral processing is relatively well understood, much more research is required to understand how the nervous system is elaborating and analyzing the acoustic information. This work focuses on the human ability to estimate the direction-of-arrival (DoA) of a static sound source which comprises both the azimuth and the elevation angle. In order to estimate the DoA from the acoustic features many models in the literature rely on a deterministic decision stage which cannot resemble the stochasticity of the nervous system (i.e. including also subcortical neural structures). Furthermore, few attempts have been made to represent this randomness as a probabilistic process. A promising tool is the Bayesian inference which can easily integrate the psychoacoustic literature's findings into a probability model. Hence, the aim of this work is twofold: (i) reproduce the auditory model proposed in [1] and then (ii) evaluate outcomes against previous psychoacoustic results [2–5] to see how well the model resembles the real data. Several simulations have been performed and the results were compared based on the original perceptual metrics. Furthermore, we added the predictions of the Baumgartner model [2] to compare the polar-angle judgments. This functional model has been developed with the aim to reproduce the listener spectral auditory processing by accounting for his acoustic and non-acoustic specificity. The second model has been tested to match actual subject's performance [6] demonstrating how auditory modelling can help to uncover the processes behind the human hearing. While the predictions of model in [2] are restricted to the polar dimension, this comparison is important to understand which model's assumptions can contribute to mimic the real predictions.

This work is organized as follow: Sect. 2 explains the model formulation, Sect. 3 reports the psychoacoustic experiments with their results and finally, Sect 4 discusses the main outcomes and limitations of the adopted model.

2. THE MODEL

The implementation of the model proposed in [1] is available in the Auditory Modeling Toolbox (AMT) [1] . The original manuscript reports every detail of the mathematical formulation but our implementation adopted slightly different assumptions to relate the original formulation with physiological and psychoacoustic grounds. The model aims to extract the azimuth and elevation angle $\boldsymbol{\theta} = (\alpha, \epsilon)$ from the acoustic information by following a template matching procedure, as assumed being implemented in the human brain [3]. *Internal templates* are computed for each of the available directions. Then a distance between the templates and the *internal realization* of the sound source is calculated. Finally, the *decision stage* enables the estimation of the source direction.

[1] https://amtoolbox.sourceforge.net

10.48465/fa.2020.0704

The processing pipeline is composed of four elements: (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 source's spatial cues. The feature space (see Eqs. 2) is constructed by computing: the interaural time difference (ITD, Eq. 2a) and the combination 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 $\mathbf{T}_{itd}^{\boldsymbol{\varphi}}$ values were computed based on a threshold method [7] and then converted into the just noticeable difference (JND). The log-spectral magnitudes were derived by filtering the binaural source with an all-pole implementation of the Gammatone filterbank [8], with 30 frequency channels each separated by 1 equivalent rectangular bandwidth (ERB) [9] within $[0.3, 15]$ kHz. The computed magnitudes are then limited to a minimal value, resembling the absolute hearing threshold at the respective frequency.

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

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

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

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

2.2 Internal Realization

Eq. 3 defines the representation of the binaural stimulus with direction $\boldsymbol{\theta}$. Uncertainties that are due to the limited precision of our hearing system [6] are assumed to be Gaussian distributed with zero mean. The quantification of each variance was derived from the psychoacoustic literature, if available, or set manually.

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

2.3 Internal Templates

The model determines the internal belief through the a-posteriori probability $P(\mathbf{X}_{\boldsymbol{\theta}}\,[\boldsymbol{\delta}]\,|\&\boldsymbol{\varphi})$ by relying on Bayes' formula. The computation of the likelihood (Eq. 4) is done for every template or direction, $\boldsymbol{\varphi}$, given the internal realization $\mathbf{X}_{\boldsymbol{\theta}}[\boldsymbol{\delta}]$. The covariance matrix is a diagonal matrix whose elements correspond to the variances of the internal noise in (Eq. 3). The denominator of Eq. 5 can be assumed constant while the prior probability, $P(\boldsymbol{\varphi})$, is uniformly distributed, meaning that every direction is equally probable. Further details are reported in [10] while a visual example is reported in Fig. 1.

$$P(\mathbf{X}_{\boldsymbol{\theta}}[\boldsymbol{\delta}]|\boldsymbol{\varphi}) \propto \exp\left\{-\frac{1}{2}(\mathbf{X}_{\boldsymbol{\theta}}[\boldsymbol{\delta}] - \mathbf{T}_{\boldsymbol{\varphi}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}_{\boldsymbol{\theta}}[\boldsymbol{\delta}] - \mathbf{T}_{\boldsymbol{\varphi}})\right\} \tag{4}$$

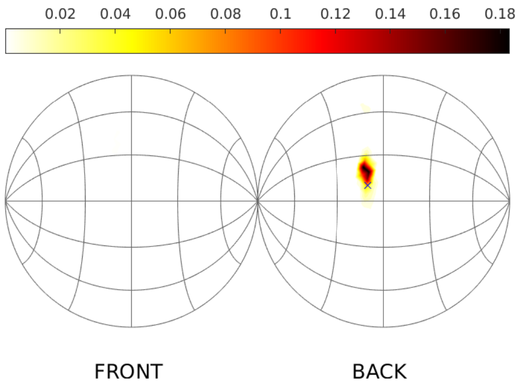

Figure 1. Internal belief obtained for the target direction $\boldsymbol{\theta} = (-14°, 171°)$. The symbol × shows the target direction and the color-coded areas the computed a-posteriori probability.

$$P(\boldsymbol{\varphi}|\mathbf{X}_{\boldsymbol{\theta}}[\boldsymbol{\delta}]) = \frac{P(\mathbf{X}_{\boldsymbol{\theta}}[\boldsymbol{\delta}]|\boldsymbol{\varphi})P(\boldsymbol{\varphi})}{P(\mathbf{X}_{\boldsymbol{\theta}}[\boldsymbol{\delta}])} \tag{5}$$

2.4 Decision Stage

The model relies on the maximum a-posteriori (MAP) estimator (Eq. 6) to estimate the azimuth and elevation angles, $\hat{\boldsymbol{\varphi}} = (\hat{\alpha}, \hat{\epsilon})$, from the internal belief.

$$\hat{\boldsymbol{\varphi}} = \arg\max_{\boldsymbol{\varphi}} P(\boldsymbol{\varphi}|\mathbf{X}_{\boldsymbol{\theta}}[\boldsymbol{\delta}]) \tag{6}$$

3. THE EXPERIMENTS

The outcome of the simulation of the evaluated model, also named Reijniers' model, is compared against the results of three different actual experiments. [2] Furthermore, this step was possible thanks to the AMToolbox which contains the implementation of both models and it made available the 23 Head-Related Transfer Function (HRTF) datasets of different subjects which were used for all the simulations. From now on, we are going to refer to the Reijniers' model with spherical angle (SA) localization model and to the Baumgartner's one as saggital-plane (SP) localization model.

3.1 Non-individual spatial filtering

This simulation replicated the work from Middlebrooks [3]. Both models were tested to predict the effect of localizing sound sources filtered by non-individual HRTF datasets. The original experiment tested eleven listeners localizing Gaussian noise bursts with a duration of 250 ms. The subjects were tested with their own set of HRTFs and also by up to 4 sets of HRTFs from other subjects (21 cases in total). Since these combinations and the actual directions were not stated in the original work, for our simulations,

[2] The original data were digitized from the original papers by relying on the software *Engauge Digitizer* [11].

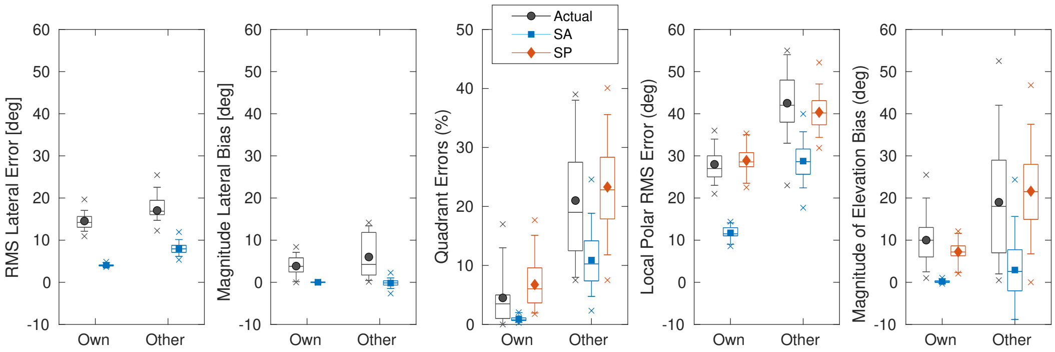

Figure 2. Performances over the polar and lateral dimensions for the direction estimation with the individual, *Own*, and non-individual, *Other*, HRTFs. The actual data are re-plotted from [3]. Lateral estimations for the Baumgartner's model are not available.

we evaluated both models on all directions available and all HRTF combinations were considered. Actual and predicted performances were measured with the same metrics: RMS Lateral Error, Lateral Bias, Quadrant Error (QE), Local Polar RMS Error and Elevation Bias. While the lateral metric covered the entire horizontal plane, the polar analysis was restricted in the $[-30°, 30°]$ interval. The original study clearly demonstrated that the performances degrade when a subject is not listening with his own HRTF. Under this condition also the models increased their uncertainties in the estimations. The results are reported in Fig. 3.1. The SA model showed the best performance across all the conditions while SP reported comparable errors with the actual results.

3.2 Band-limited sound sources

Here we evaluated the effect of time-variant and band-limited sound sources. The original experiment [4] which work relied on a speech corpus composed of 260 monosyllabic words, with a band width of $[0.3, 16]$ kHz and an average duration of 710 ms. The samples were filtered with a low-pass filter, $fc = 8$ kHz, with different attenuation in the stop band: $0, 20, 40, 60, 80$ dB. The performances of five trained subjects were recorded and a Gaussian broad band noise provided the baseline condition. Furthermore, the estimated directions were analyzed by means of the absolute lateral and polar errors and QE. As the main outcome, the study reported a degraded performances for the polar and quadrant errors when moving towards the extreme stop-band attenuation while the lateral error did not show significant effects (see Fig. 3). While for the SA model only the median plane was accounted, the simulations performed with SP accounted for all available directions. The model from Baumgartner *et al.* represented the actual data by decreasing the performances when the attenuation increased. Instead, the simulations for SA resulted in a constant accuracy disregarding of the specific condition.

3.3 Sound source with rippled spectrum

In this experiment we reproduced the work done by Macpherson and Middlebrooks [5] probing the localization performance for spectrally rippled noises. The aim was to investigate how a non-flat source spectrum can disrupt the subject's performances considering that the psychoacoutic literature reports that human subjects rely on spectral features to estimate the polar-angle dimension [12]. The spectral ripples were generated in the frequency band $[1, 16]$ kHz with a sinusoidal spectral shape in the log-magnitude domain. The conditions considered different ripple depths, defined as the peak-to-peak difference, and ripple densities, defined as the period of the sinusoidal shape along the logarithmic frequency scale. The actual experiment tested six trained subjects in a dark, anechoic chamber listening to the stimuli via loudspeakers. The sounds were 250 ms long and they were positioned across the whole lateral dimension and within $[-60°, 60°]$ and $[120°, 240°]$ for the polar angle. For the polar error, the lateral interval was limited to $[-30°, 30°]$. Furthermore, the polar error definition relied on an selective procedure by computing two regressions, separating between front and back. The polar error was then retrieved only if the difference of the estimation with the regression line was greater than $45°$. The results are reported in Fig. 4. Since the spatial grid was not reported in the original work, we simulated five repetitions of the all the available directions in the accounted interval. The first outcome of the study was that by increasing the ripple depth the listeners, performances gradually decreased. When accounting for the ripple density, the performances worsened up to 4 ripples/octave. The actual results and our simulations are shown in Fig. 4. The SP model underestimated the performance in the baseline condition, but it followed the trends of the remaining conditions well. Instead, our implementation of the SA model reproduced the baseline while it reported super-human performance for the other with no effect of ripple density variations while it improved when the depth increased because for low depth values the integrated hearing threshold reduces its internal information which decreases its discrimination capabilities.

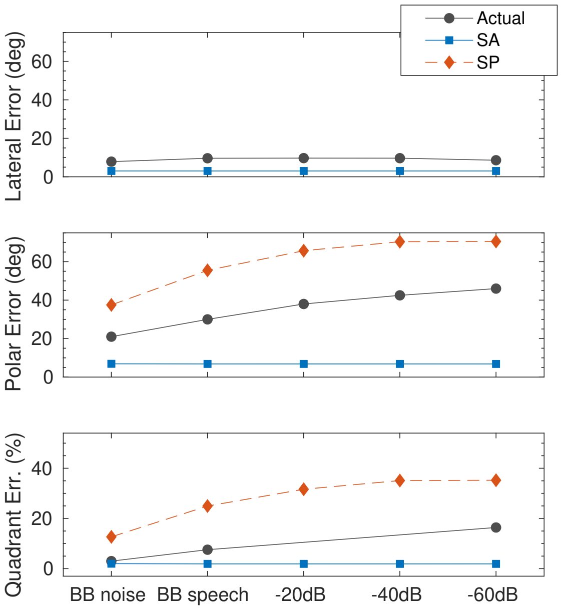

Figure 3. Human and models, performances with band-limited speech samples. The results are from [4] and the absolute lateral (top) and polar (middle) angle errors and QE (bottom) were averaged over directions and listeners. The stimulus are broad-band (BB) noise, BB speech and $\{-20, -40, -60\}$ dB represents the stop-band attenuation of the processed BB speech samples with a low pass filter, $fc = 8$ kHz.

4. DISCUSSION

Comparisons between two auditory models and the actual data from three different experiments were performed. The localization experiments tested the effects of filtering with non-individualized HRTF (Sect. 3.1), and sources spectrum distortions, by means of filtering speech samples (Sect. 3.2) and spectral ripples (Sect. 3.3). Only in one case the novel model showed a good agreement with the literature while super-human performances have been reported for all other cases. The match was found for the baseline condition of the experiment from Macpherson and Middlebrooks [5] and it can be explained through the fact that the actual subjects were trained showing better performance than the average population. The deviations can be due to the mathematical formulation and uncertainties quantification within the Reijniers' model since the work from Baumgartner *et al.* reported similar trends with the real data in most of the conditions. Although the formulation of SA is interesting for its simplicity, it is clear that mimicking the human hearing process, as in SP, from a functional perspective can help to simulate the actual results. The super-human performances can also be related to the mathematical methods of SA which introduce an high discrimination between the internal realization and the templates thanks to the adoption of probability theory. Moreover, the SA model returned very similar performances across different conditions in contrast to the large differences observed in human behaviour. This mismatch can be explained by four major issues of the model. First, the quantification of the internal noises does not result into similar error predictions for the baseline conditions of the experiments in Sect 3.1 and 3.2. Second, the computation of the feature space relies on the log-magnitude average of the HRTF amplitude over a limited set of frequency bands. While this aggregation allows an efficient implementation, it is reducing the variability on which the human brain has to deal with in the real case. Third, the sound source spectrum is also averaged over time and then added to both the target and template. Afterwards, this information is ruled out when the Bayes formula is computed, see Eq. 4, removing the source spectral cues from the estimation pipeline. Finally, the adoption of the MAP estimator made the model an ideal observer which a human subject is not [13].

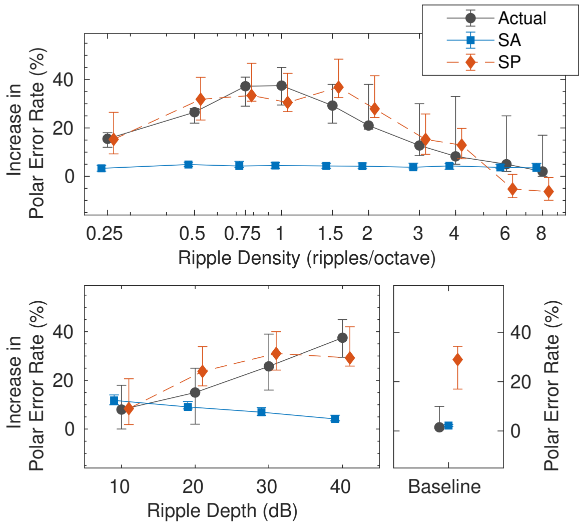

Figure 4. Polar errors of spectral ripples. The actual results are reported from [5]. The ripple density of 1 ripple/octave (bottom) or a depth of 40 dB (top) was kept constant. The polar error on the different condition is computed as difference with the baseline condition (bottom-right). The plots report the medians with the respective quartile intervals.

Despite the reported limitations, we believe that the SA model's structure successfully interpreted the Bayesian approach to resemble the human perception. Hence, relying on the formulation of Reijniers *et al.* we propose here some elements that can be addressed to improve the model. From our point of view the feature space should describe how the human ear is transducing the acoustic field into neural information and, consequently, how this information is processed in a functional perspective. While the model adopted a multidimensional feature space to wrap the binaural information, the human hearing adopts a hierarchical organization of the information [2]. For instance, the psychoacoustic literature reports that listeners rely mainly on ITD to estimate the lateral direction [14] and to weight the monaural spectral features [2]. Also the probabilistic formulation of the model can be addressed. The representation of the noise sources should be separated since the original model adopts the same quantities to define the noise of the external ear, Eq. 3, and the cognition uncertainties $\mathbf{\Sigma}$. Moreover, the decision system should be replaced with a heuristic method since it appears from the literature that human localization task is not following the MAP estimator [13].

5. CONCLUSION

Our implementation of the auditory localization model proposed by Reijniers et *al.* [1] was tested and evaluated under different conditions. The performances of this model and the one proposed by Baumgartner *et al.* [2] were compared against the actual data of previous experiments [3], [4], [5]. While our implementation of the model did not resembled the real data by showing super-human performancs, we believe that its Bayesian formulation can help to mimic the human behaviour by integrating uncertainty factors into the model. Hence, we underlined some issues that might help moving towards an accurate representation. Future work will attempt to build upon these critical elements relying of on more robust probabilistic modelling techniques.

6. REFERENCES

[1] J. Reijniers, D. Vanderelst, C. Jin, S. Carlile, and H. Peremans, “An ideal-observer model of human sound localization,” *Biological Cybernetics*, vol. 108, pp. 169–181, Apr. 2014.

[2] R. Baumgartner, P. Majdak, and B. Laback, “Modeling sound-source localization in sagittal planes for human listeners,” *The Journal of the Acoustical Society of America*, vol. 136, no. 2, pp. 791–802, 2014.

[3] J. C. Middlebrooks, “Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency,” *The Journal of the Acoustical Society of America*, vol. 106, pp. 1493–1510, Aug. 1999.

[4] V. Best, S. Carlile, C. Jin, and A. van Schaik, “The role of high frequencies in speech localization,” *The Journal of the Acoustical Society of America*, vol. 118, pp. 353–363, July 2005.

[5] E. A. Macpherson and J. C. Middlebrooks, “Vertical-plane sound localization probed with ripple-spectrum noise,” *The Journal of the Acoustical Society of America*, vol. 114, pp. 430–445, July 2003.

[6] P. Majdak, R. Baumgartner, and B. Laback, “Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization,” *Frontiers in Psychology*, vol. 5, Apr. 2014.

[7] B. F. G. Katz and M. Noisternig, “A comparative study of interaural time delay estimation methods,” *The Journal of the Acoustical Society of America*, vol. 135, pp. 3530–3540, June 2014.

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

[9] B. C. J. Moore and B. R. Glasberg, “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,” *The Journal of the Acoustical Society of America*, vol. 74, pp. 750–753, Sept. 1983.

[10] R. Barumerli, P. Majdak, J. Reijniers, R. Baumgartner, M. Geronazzo, and F. Avanzini, “Predicting directional sound-localization of human listeners in both horizontal and vertical dimensions,” in *Audio Engineering Society Convention 148*, Audio Engineering Society, 2020.

[11] M. Mitchell, B. Muftakhidinov, T. Winchen, B. van Schaik, A. Wilms, kylesower, kensington, Z. Jedrzejewski-Szmek, T. G. Badger, and badshah400, “markummitchell/engauge-digitizer: Version 12.1 Directory dialogs start in saved paths,” Nov. 2019.

[12] M. Geronazzo, S. Spagnol, and F. Avanzini, “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*, vol. 26, pp. 1247–1260, July 2018.

[13] R. Ege, A. J. V. Opstal, and M. M. Van Wanrooij, “Accuracy-Precision Trade-off in Human Sound Localisation,” *Scientific Reports*, vol. 8, p. 16399, Dec. 2018.

[14] S. Carlile, S. Delaney, and A. Corderoy, “The localisation of spectrally restricted sounds by human listeners,” *Hearing Research*, vol. 128, pp. 175–189, Feb. 1999.