DOI: 10.61782/fa.2025.0824

FORUM ACUSTICUM
EURONOISE 2025

SPIKING NEURAL NETWORKS FOR SOUND LOCALIZATION: A NEW PERSPECTIVE ON AUDITORY SPATIAL PERCEPTION

Qin Liu[1*] **Simon Laurent**[2] **Hannes Wüthrich**[2] **Hervé Lissek**[1]

[1] Laboratory of Wave Engineering, EPFL, Lausanne, Switzerland
[2] Sonova AG, Zurich, Switzerland

ABSTRACT

Mammals can localize sounds by using the information in their auditory neuronal system. Conventional models use auditory cues (e.g., interaural time differences (ITDs) and interaural level differences (ILDs), spectral cues, etc.) to perform the sound localization. These cues are extracted from binaural and monaural signals or inferred from neuronal spikes. In this article, we proposed a new model that predicts sound source directions from the firing rates of auditory neurons directly, bypassing the auditory cue extraction. This model incorporates auditory peripheral processing, spiking neural networks (SNNs), and deep neural networks (DNNs) to achieve auditory spatial perception. The peripheral processing transfers the binaural signals into auditory nerve fiber (ANF) spikes; the SNN calculates the firing rate of the medial superior olive (MSO) based on the ANF firing rate. The DNN performs nonlinear regression to predict azimuth and elevation angles based on the ANF and MSO firing rates. Preliminary results demonstrate that the SNN-DNN system can accurately estimate source direction.

Keywords: *Sound Localization, Auditory Perception, Spiking neural network, Deep neural network.*

**Corresponding author*: *qin.liu@epfl.ch.*

1. INTRODUCTION

Mammals can localize sounds in space by using the auditory system to decode sound signals received by their two ears. The basic principles for the decoding process in human spatial hearing have been investigated for decades, both in terms of auditory physical metrics and neuronal processing [1, 2]. From the standpoint of physics, the sound source position is related to auditory cues that can be directly extracted from the binaural signals, such as interaural time differences (ITDs), interaural level differences (ILDs), and monaural spectral cues [3, 4]. The low-frequency sound localization relies on ITDs, while high-frequency localization depends on ILDs and envelope ITDs. Specifically, when sound frequencies are below 1,400 Hz, localization is primarily based on ITDs. At higher frequencies, between 4,000 and 20,000 Hz, ILDs become the dominant cue due to the increasing influence of head shadow. The greatest difficulty in localization occurs in the range between 1,500 and 4,000 Hz, where neither ITDs nor ILDs provide reliable information. [2, 5].

The perception of sound sources in elevation primarily relies on monaural spectral cues, which are shaped by the diffraction of sound waves over the head and torso [6]. The pinna, especially, is crucial for discerning the vertical position of high-frequency sounds (above 3000 Hz) through what is known as the pinna effect [7, 8].

From the perspective of neuroscience, neurons encode position information about sound sources in the spikes and their firing rates and transmit it through the auditory system. The auditory system contains the peripheral processing parts (i.e., outer, middle, and inner ear) and brain regions (i.e., superior olivary nuclei, inferior colliculus (IC), and auditory cortex) [7, 9, 10]. The

binaural sounds are converted into electrical signals (action potentials) by hair cells in the cochlea of the inner ear. The neuron impulses are conducted to the central auditory pathway by auditory nerve fibers (ANFs) [11, 12]. These fibers are responsible for encoding information regarding sound stimuli and transmitting it from the cochlea to the cochlear nucleus (CN) of the brainstem [8]. The initial stage of neural integration for binaural information takes place within the superior olivary complex (SOC) of the brainstem [13]. These organs combine encoded auditory cues into a comprehensive spatial framework, enabling humans to perceive the full spatial context, including sound direction and distance.

From the combined viewpoint of physics and neuroscience, the SOC plays a crucial role in determining the spatial position of sound sources due to its specialized nuclei that decode key auditory spatial cues. The medial superior olive (MSO) appears to be primarily responsible for extracting ITDs from low-frequency auditory signals, while the lateral superior olive (LSO) processes ILDs [13–15]. The SOC sends its output to the IC in the middle brain for further integration. Monaural spectral cues—particularly the pinna-induced spectral notches and peaks above 3 kHz—are encoded through spatial patterns of ANF firing rates across characteristic frequencies (CFs) [16], reflecting the tonotopic mapping of spectral information along the basilar membrane.

Drawing on physical auditory metrics and their neural correlates, various localization models have been developed to characterize the processes underlying spatial hearing. These models can generally be divided into two approaches: metrics-based models that focus on physical signal properties and neural-based models that simulate the auditory neuron system to decode auditory cues. Together, they provide significant theoretical and practical contributions to understanding auditory localization.

Regarding metrics-based models, most approaches focus on auditory cue extraction and template matching. [17] proposed a quantitative model for predicting elevation angles. This model suggests that listeners compare the spectral information of incoming sounds to a individual reference of spectral templates derived from their Head-Related Transfer Functions (HRTFs) which is related with their long term memory. The elevation is then determined by finding the best matches in this template: [2]. A Bayesian-based model was proposed by [18], which uses ITD and spectral cues to predict the two-dimensional direction of sound arrival. The performance of the model in [18] aligns well with actual human localization performance since the model parameters are based on individualized psychoacoustic experiments. An extension of this work, presented by [19], compares the efficacy of using magnitude profiles versus gradient profiles as monaural spectral cues, finding that positive spectral gradients yield superior localization accuracy.

Among the aforementioned classical metrics-based models, the gammatone filterbank is used in peripheral auditory processing to mimic and simplify the complex behavior of the cochlea [4, 19, 20]. Neural-based models, on the other hand, often use non-linear spiking neural networks instead of these linear gammatone filters to try to replicate cochlear processing more specifically. A spike-generation model was proposed by [7, 11], whose input is binaural sound and output is the firing rate of ANF. The ability to simulate inner and outer hair cell dysfunction in the *Zilany model* provide the feasibility of introducing hearing loss into the auditory localization model. A physiological accuracy-improved model was proposed by [21], which better explains the statistics of spontaneous activity in ANF by involving adaptive redocking dynamics, but is more computationally complex. Additionally, considering the relationship between the firing rates of the MSO and the ITD, as well as the LSO and ILD, neural-based models have been developed to decode these spatial cues from neural spikes for localization. This approach is believed to more closely resemble the actual processes occurring within our auditory system.

A MSO spiking neural network (SNN) was proposed by [22] for sound localization, where HRTFs data from adult cats are used as the input binaural signals, and an MSO model is used to extract the ITDs from the delayed spikes for localizing the azimuth angles. The localization accuracy achieves 91.82% with an error tolerance of $\pm 10°$. An excitatory-inhibitory model using the ANF firing rates difference between left and right hemispheric to predict the extent of laterality was proposed by [15], but this model is limited to high-frequency stimuli and not extended to predict azimuth angles. A sagittal-plane localization model proposed by [8] simulates the impact of sensorineural hearing loss based on the peripheral processing model from [7]. The positive spectral gradients of ANF firing rates are used as the auditory metrics for predicting elevation. A physiologically plausible spiking neuron network (SNN) model based on the opponent-channel hypothesis was developed by [13], incorporating the auditory periphery model by [7]. The model simulates the role of excitatory and inhibitory inputs in encoding ITDs through opponent-channel coding. Two approaches

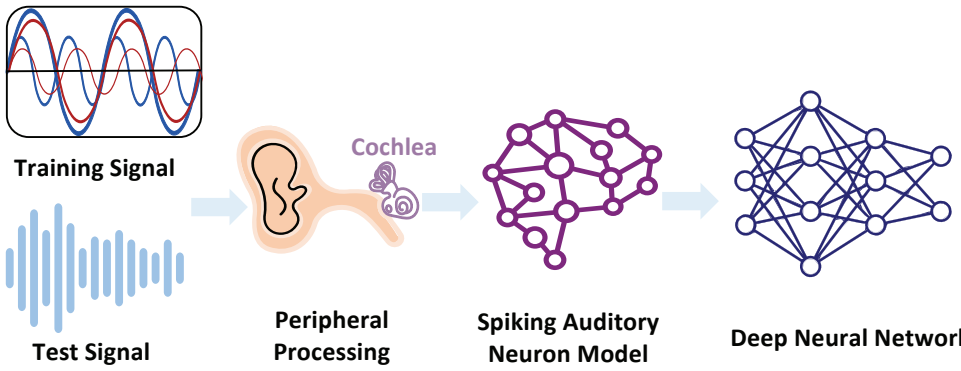

Figure 1. Overview of the proposed auditory localization model. Binaural signals are first processed by a peripheral auditory model to generate ANF spikes, which are then converted into MSO spikes using a spiking auditory neuron model. A DNN predicts lateral and polar angles directly from the firing rates of MSO and ANF, respectively. The training signals consist of multiple pure tones and white noise, assigned to azimuth and elevation prediction, respectively. The testing signals are composed of white noise.

were employed to extract ITDs: a linear two-channel decoder and a trained artificial neural network (ANN) [13]. These methods demonstrated precise ITD detection across a wide frequency range, including complex speech signals. The findings provide insights into the biophysical mechanisms of ITD encoding and present computational frameworks for understanding auditory spatial perception in natural environments. However, these SNN-based models are still only using those SNNs to extract auditory cues to resolve the sound localization task.

Deep neural networks (DNNs) have been employed to advance the understanding of auditory perception, particularly in sound localization [1, 12]. Peripheral auditory models were integrated into the DNNs to approximate the neural coding of auditory signals. These models simulated the fine-grained temporal precision of ANF spikes and their degradation, revealing the necessity of phase-locked spike timing for tasks like sound localization [12]. The models were shown to replicate human sound localization accuracy and robustness under noisy and complex auditory conditions.

In this paper, a hybrid model is introduced, which integrates an SNN with a DNN to achieve sound localization. The proposed model introduces a new perspective on spatial perception by demonstrating that sound direction-related spatial information can be directly extracted from the spiking activity of ANF and MSO neurons, bypassing traditional physical auditory cue (e.g., ITD/ILD) extractions. The model is preliminarily validated with broadband signals, which laid the foundation for further research.

2. METHODS

2.1 Model structure

The proposed localization model has two main steps: calculating the firing rate dataset and predicting the angles through DNN, which skips the step of extracting auditory cues. The model architecture is shown in Fig. 1. In the training phase, pure tone signals at multiple CFs and amplitudes were employed for azimuth angle prediction, whereas white noise signals were used for elevation angle prediction. During testing, white noise signals were utilized to estimate both azimuth and elevation angles.

Peripheral auditory processing is implemented by the *Zilany model* [7]. The Zilany model has a middle-ear compensation filter, a nonlinear basilar membrane model, and a neural transduction stage that simulates the conversion of signals from inner hair cells to ANF. The model takes binaural time-domain signals as input and outputs ANF spikes. The high-spontaneous-rate ANF spikes are then transferred into MSO spikes by a SNN [13]. In Encke's study, this model was used to decode ITDs from both single-frequency and speech signals [13]. Finally, the DNN uses firing rates from ANF and MSO populations to predict elevation and azimuth angles, respectively,

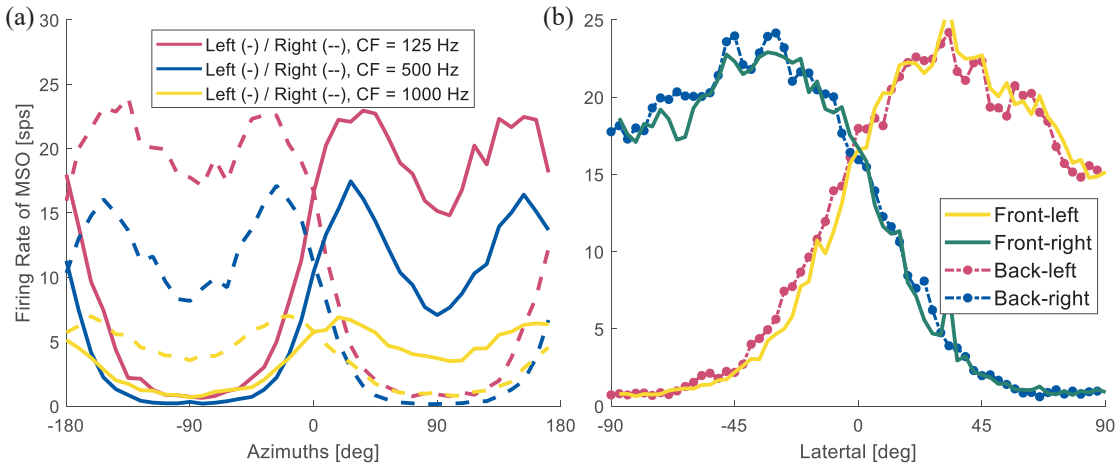

Figure 2. The relationship between azimuth angle and averaged MSO firing rate. The MSO firing rate was derived at a specific CF from a sine tone at same frequency. (a) MSO firing rates from the left (solid lines) and right (dashed lines) show strong left-right symmetry across azimuth angles ranging from –180° to 180°, under CFs = 125 Hz, 500 Hz, 1000 Hz. (b) MSO firing rates exhibit front-back symmetry under CF = 125 Hz when flipping the azimuth angles.

through nonlinear regression. We describe the DNN architecture in detail in the following subsection.

The non-linear relationship between averaged MSO firing rate and sound source directions is shown as Fig. 2. The MSO firing rates were calculated at a specific CF in response to a sine tone, convolved with HRTFs. The azimuth angle is defined as follows: 0° means that a sound source is in front of the listener; negative values indicate positions to the left side, while positive angles correspond to positions to the right side; ±180° represents a sound source behind the listener. In order to more clearly exhibit the relationship between azimuth angle and firing rate, an interaural-polar coordinate system [23] was adopted to flip the back azimuth angle to the front. In this system, spatial locations are represented by two angles: the lateral angle, indicating horizontal position relative to the median plane, and the polar angle, denoting vertical and front-back position around the interaural axis of the listener inside the sagittal plane. The lateral angle ranges from -90° at the left side of the listener, through 0° at the median plane, to +90° at the right side. The polar angle ranges from 90° to 270°, with 0° directly in front at eye level, 90° at the top (above the listener), and 180° right behind the listener.

Both ears exhibit clear and left-right symmetric modulation patterns as a function of azimuth under lower frequency, illustrated in Fig. 2. The maximal responses are recorded when the sound source is on the opposite side of the ear under consideration—negative azimuth angles yield high firing rates in the right ear, while positive azimuth angles stimulate high firing rates in the left ear. As the CF increases, the magnitude of modulation significantly diminishes, indicating reduced directional sensitivity of MSO neurons at higher frequencies, as shown in Fig. 2(a). This observation aligns well with the fact that ITD cues primarily work at lower frequencies for sound localization. The firing rates exhibit a highly symmetrical pattern when comparing front and back hemispheres. Specifically, the average firing rate curves from the back hemisphere closely overlap with those from the front when the azimuth angles are folded, shown in Fig. 2(b).

This symmetrical firing rate behavior across lateral angles causes front-back confusion during sound localization. To resolve this ambiguity, spectral information is necessary. In the proposed model, positive spectral gradients derived from the ANF firing rate are used to predict polar angles. If the estimated polar angle exceeds 90°, the sound comes from the back hemisphere.

2.2 Deep neural network structure

A DNN model is designed for auditory localization, integrating ANF and MSO firing rate in a two-stage structure.

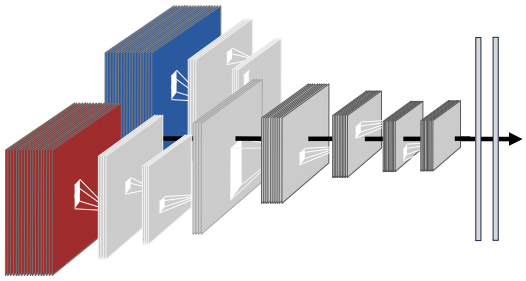

Figure 3. The schematic diagram of proposed DNN structure.The blue and red blocks represent left and right ear inputs, respectively, composed of ANF and MSO firing rates across CFs. These inputs are processed through two parallel CNN branches. The outputs from the left and right branches are merged to a joint binaural representation which consists the ANF and MSO parts. The polar angle is predicted from the ANF part of this representation, while lateral angle is predicted from the MSO part.

The first step employs ANF firing rate across different CFs (700-18000 Hz) and MSO firing rate (CFs: 125-1000 Hz) through convolutional neural networks (CNNs) to get a joint binaural representation. This representation consists ANF part and MSO part. In the second stage, non-linear regression are utilized to predict lateral angles based on MSO part, and polar angles through ANF part.

The ANF firing rate data, structured as temporal sequences across time steps, and MSO firing rates are processed through two parallel CNN branches corresponding to the left(blue) and right(red) ears, as illustrated in Fig. 3. Each branch consists of multiple convolutional layers designed to extract spectral-temporal features, followed by batch normalization and rectified linear unit (ReLU) activations. Dropout regularization is subsequently applied to prevent overfitting. Max-pooling layers are used to reduce the dimensionality of feature representations and to preserve important spatial and temporal information at the same time. The outputs from the left and right branches are merged along the feature dimension to form a joint binaural representation, which consists of both ANF and MSO components.

The second processing stage utilizes a multilayer perceptron (MLP) to estimate both lateral and polar angles. This dual-angle prediction framework employs different input pathways: the MLP processes MSO component features for lateral angle estimation and ANF parts for polar angle prediction. The regression network consists of several hidden layers utilizing hyperbolic tangent sigmoid activation functions to capture non-linear relationships between MSO firing rates and lateral angles. The network is trained by minimizing mean squared error (MSE). The dataset is divided into training, validation, and testing subsets to thoroughly evaluate generalization performance.

This two-stage DNN mimics human auditory processing mechanisms, polar angle prediction through ANF firing rates at high frequencies which resolves the front-back confusion, while lateral localization partly depends on ITD information encoded in MSO firing rates at low frequencies.

3. RESULTS

The training datasets of ANF and MSO fring rate is generated by the peripheral model and SNN, respectively. For the ANF dataset, the inputs to the peripheral model are 2520 distinct 3000 ms long white noise signals (252 positions × 10 sound pressure level). The firing rate is calculated at 35 different CFs. For the MSO dataset, the inputs for SNN are 2520 distinct 300 ms long sine tones (252 positions × 10 sound pressure level × 20 frequencies). The pure-tone frequencies are matched to the CFs of the corresponding MSO neurons.

60 lateral angles ($\theta \in (-90:3:90)$) and 60 polar angles ($\phi \in (0:3:180)$) were selected from the HRTF of Knowles Electronic Manikin for Acoustic Research (KEMAR), resulting in 360 unique positions. In the initial model, only the positions in upper hemisphere was considered. The dataset of HRTFs was partitioned into three distinct subsets: a training set (70%), a validation set (15%), and a test set (15%), resulting 252 positions for calculating the aforementioned firing rate training dataset. ANF firing rates were calculated for 35 logarithmically spaced frequencies between 700 Hz and 18,000 Hz, while MSO firing rates were computed for 20 logarithmically spaced frequencies between 125 Hz and 1,000 Hz. The 10 different levels are randomly selected between 50 and 70 dB.

To test the preliminary performance of the proposed model, white noise signals are used as input stimulus. The localization error of the lateral angle under different polar angle conditions is shown in Fig. 4(a). Due to the left-right symmetry of the lateral angle, the experimental results are averaged for the left half and the right half, and the polar angles are selected as 10°, 30°, 60° and 120°.

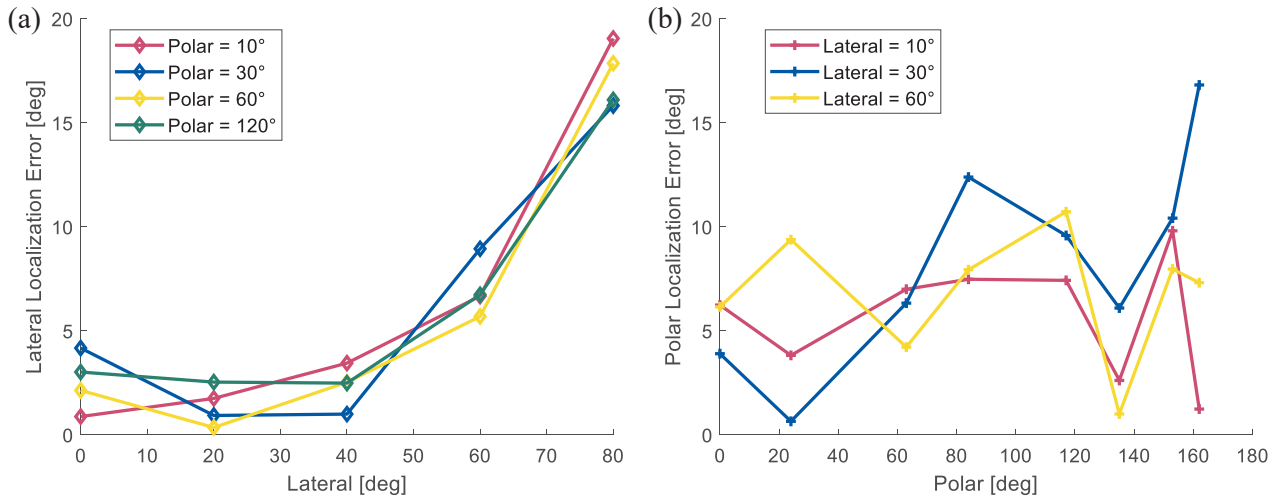

Figure 4. The localization error of the proposed model evaluated using broadband speech stimuli. Errors are averaged across left and right hemispheres due to lateral symmetry. (a) Lateral localization error under different polar angle conditions (10°, 30°, 60°, and 120°). (b) Polar localization error under different lateral angle conditions (10°, 30°, and 60°).

The localization error of a polar angle under different lateral angles is shown in Fig. 4(b). The lateral angles are selected as 10°, 30°, and 60°.

It can be seen that when the lateral angle is less than 40° in Fig. 4, the localization error is less than 5 degrees under all polar angle conditions. The localization error increases as the lateral angle approaches the left or right side (lateral = ±90°), which is consistent with human behavior. In contrast, no clear trend is observed between polar localization error and either polar or lateral angle.

4. CONCLUSION

This study presents a hybrid model that directly decodes sound source position from spiking activity in the auditory system. By leveraging both ANF and MSO firing rates, the model eliminates the need for traditional auditory cue extraction and instead relies on a DNN to infer sound location through a two-stage process. Preliminary results indicate that the SNN-DNN hybrid system can accurately predict source direction. This approach opens new perspectives for modeling auditory perception under realistic and impaired conditions. Future work will extend the applicability of the proposed model to noisy acoustic environments and simulate auditory deficits such as age-related sensorineural hearing loss by adjusting model parameters and training datasets. This will make the model a promising tool for revealing real-world spatial perception.

5. ACKNOWLEDGMENTS

This work was funded by the Swiss Innovation Promotion Agency (Innosuisse) with grant number [103.590 IP-LS] in collaboration with Sonova AG.

6. REFERENCES

[1] A. Francl and J. H. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nature human behaviour*, vol. 6, no. 1, pp. 111–133, 2022.

[2] J. C. Middlebrooks, "Sound localization," *Handbook of clinical neurology*, vol. 129, pp. 99–116, 2015.

[3] M. Dietz, S. D. Ewert, and V. Hohmann, "Lateralization of stimuli with independent fine-structure and envelope-based temporal disparities," *The Journal of the Acoustical Society of America*, vol. 125, no. 3, pp. 1622–1635, 2009.

[4] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speak-

ers from binaural signals," *Speech Communication*, vol. 53, pp. 592–605, 2011.

[5] A. Carlini, C. Bordeau, and M. Ambard, "Auditory localization: A comprehensive practical review," *Frontiers in Psychology*, vol. 15, p. 1408073, 2024.

[6] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[7] M. S. A. Zilany, I. C. Bruce, and L. H. Carney, "Updated parameters and expanded simulation options for a model of the auditory periphery," *The Journal of the Acoustical Society of America*, vol. 135, pp. 283–286, 01 2014.

[8] R. Baumgartner, P. Majdak, and B. Laback, "Modeling the effects of sensorineural hearing loss on sound localization in the median plane," *Trends in Hearing*, vol. 20, 2016.

[9] P. W. Anderson and P. Zahorik, "Auditory/visual distance estimation: accuracy and variability," *Frontiers in psychology*, vol. 5, p. 96823, 2014.

[10] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *ACTA Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[11] M. S. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2390–2412, 2009.

[12] M. R. Saddler and J. H. McDermott, "Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing," *Nature Communications*, vol. 15, no. 1, pp. 1–29, 2024.

[13] J. Encke and W. Hemmert, "Extraction of interaural time differences using a spiking neuron network model of the medial superior olive," *Frontiers in Neuroscience*, vol. 12, 2018.

[14] L. Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology*, vol. 61, pp. 468–486, 1947.

[15] J. Klug, L. Schmors, G. Ashida, and M. Dietz, "Neural rate difference model can account for lateralization of high-frequency stimuli," *The Journal of the Acoustical Society of America*, vol. 148, pp. 678–691, 08 2020.

[16] A. J. Kolarik, B. C. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss," *Attention, Perception, & Psychophysics*, vol. 78, pp. 373–395, 2016.

[17] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annual review of psychology*, vol. 42, no. 1, pp. 135–159, 1991.

[18] 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, 2014.

[19] R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, and R. Baumgartner, "A bayesian model for human directional localization of broadband static sound sources," *Acta Acustica*, vol. 7, p. 12, 2023.

[20] 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.

[21] I. C. Bruce, Y. Erfani, and M. S. Zilany, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Implications of limited neurotransmitter release sites," *Hearing Research*, vol. 360, pp. 40–54, 2018. Computational models of the auditory system.

[22] B. Glackin, J. A. Wall, T. M. McGinnity, L. P. Maguire, and L. J. McDaid, "A spiking neural network model of the medial superior olive using spike timing dependent plasticity for sound localization," *Frontiers in Computational Neuroscience*, vol. 4, 2010.

[23] M. Morimoto and H. Aokata, "Localization cues of sound sources in the upper hemisphere.," *The Journal of The Acoustical Society of Japan (e)*, vol. 5, pp. 165–173, 1984.