Localization in Elevation with Non-Individual Head-Related Transfer Functions: Comparing Predictions of Two Auditory Models

Roberto Barumerli
Dept. of Information Engineering
University of Padova
Email: barumerli@dei.unipd.it

Michele Geronazzo
Dept. of Architecture, Design
and Media Technology
Aalborg University
Email: mge@create.aau.dk

Federico Avanzini
Dept. of Computer Science
University of Milano
Email: federico.avanzini@di.unimi.it

***Abstract*—This paper explores the limits of human localization of sound sources when listening with non-individual Head-Related Transfer Functions (HRTFs), by simulating performances of a localization task in the mid-sagittal plane. Computational simulations are performed with the CIPIC HRTF database using two different auditory models which mimic human hearing processing from a functional point of view. Our methodology investigates the opportunity of using virtual experiments instead of time- and resource- demanding psychoacoustic tests, which could also lead to potentially unreliable results. Four different perceptual metrics were implemented in order to identify relevant differences between auditory models in a selection problem of best-available non-individual HRTFs. Results report a high correlation between the two models denoting an overall similar trend, however, we discuss discrepancies in the predictions which should be carefully considered for the applicability of our methodology to the HRTF selection problem.**

I. Introduction

Spatial hearing defines the perceptual ability to localize sound sources in space. In particular, mammals – and thus humans – continuously analyze the acoustic scene retrieving and monitoring surrounding source positions. This process is performed based on the two-channel binaural sound stream which is filtered by subject physicality: sound waves diffract and interact with torso, head and external ears, causing listener-dependent temporal and spectral transformations [1]. The resulting effects provide meaningful cues about sound source locations in an egocentric view. Binaural cues heavily influence azimuth and lateral localization that is evaluated mostly by mean of *interaural time difference* (ITD), and *interaural level difference* (ILD).

On the other hand, spectral cues are primary cues for elevation perception, and *head-related transfer function* (HRTF) contains such relevant information; HRTF measurements summarize the direction-dependent acoustic filtering of a free-field point source due to the head, torso, and pinna [1]. Knowledge of such a complex process is needed for the development of accurate and realistic artificial sound spatialization in several application domains, including immersive virtual and augmented reality, gaming, 3DTV and cinema, etc.

HRTFs are deeply connected to listener anthropometry, electing individual HRTFs the *ground truth* condition for sound spatialization [2]. However, the acquisition process of individual HRTF sets is time consuming, requires a complex hardware setup, and the measurement protocol varies for each laboratories leading to different results [3]. This unpractical solution has driven the researchers to propose alternative methods to provide a perceptually plausible HRTF set which approximates individual variations of each listener. The most common method uses a generic HRTF set, typically dummy-head measurements [2]; otherwise, the selection approaches employ some perceptually motivated metrics to provide the most effective HRTF set taken from a database of pre-recorded measurements [4], [5]; HRTF selection methods can dramatically improve especially elevation perception, where human localization abilities are known to be less accurate than in the lateral plane [6].

In this paper, we analyze the best choice for HRTF selection according to auditory model predictions, thus implementing two series of virtual experiments able to compute perceptual errors in two virtual worlds, i.e. in the theoretical frameworks of the two auditory models. In order to perform such analysis, we adopted a public available database,[1] known as *CIPIC HRTF Database* [7]. Hence, the selection problem is faced by using a metric built upon two different auditory models from Baumgartner *et al.* [8], [9], which enables the evaluation of the the best non-individual HRTF set for all CIPIC subjects in the dataset. The adopted models derive from a *functional* model, proposed by Langendijk and Bronkhorst [10], which describes the processing sequence performed by the human auditory system in localizing a static audio source. This is achieved by simulating the transformations undergone by the audio signal from the outer ear up to the cochlea in the form of a deterministic process.

The main aim of our contribution is to compare predictions in localization with non-individual HRTFs of both auditory models, focusing on best-possible HRTF selections according

[1] http://sofacoustics.org/data/database/cipic/

ISBN 978-90-827970-1-5 © EURASIP 2018

to these two simulated scenarios. Following a structural modeling approach [11], sensitivities on localization performances are analyzed in the mid-sagittal plane, both with and without torso acoustic contribution.

II. Materials

The virtual experiments were conducted by means of the *Auditory Modeling Toolbox (AMT)*[2] which includes the implementations of both the auditory models by Baumgartner *et al.*, as well as the functions for directly working with HRTF database wrapped into *SOFA* format.[3] Figure 1 depicts the work-flow of our methodology.

A. The CIPIC HRTF Database

This database consists on 45 HRTF sets measured on different subjects, including the KEMAR mannequin. The dataset comprises *head-related impulse responses* (HRIR) for 1250 directions for each ear and subject, measured at a sampling rate of $f_s = 44.1$kHz, with 16 bit resolution and length 4.5 ms. HRTFs are derived from the Fourier transform of the HRIRs. The database also includes some anthropometric measures of the subjects.

In order to study the elevation error, the interval $\phi \in [-45°, 45°]$ in the mid-sagittal plane was considered: for the remaining elevation angles the spectral details led to a very poor localization performances even in real life conditions [12]. The dataset was further reduced to 31 subjects by discarding those with high predicted localization errors: this was achieved by removing those subjects with less than three elevation-dependent notches in the HRTF magnitude responses (see Geronazzo *et al.* [12] for a detailed analysis of this issue).

B. Auditory models

The tools for inter-subject perceptual evaluation are based on two auditory models for sound localization in sagittal planes proposed by Baumgartner *et al.* [8], [9]. These models are implemented into the *AMT* with the scripts `baumgartner2013` and `baumgartner2014`, respectively.

In particular, these models simulate virtual experiments quantifying a perceptual metric on localization for *stationary broadband* auditory stimuli. These perceptual metrics, originally introduced by Langendijk [10], compare the *target* sound, processed to obtain an internal representation,[4] with an internal *template*, resulting in a probabilistic prediction of polar angle response. The template encloses the process in learning the correspondence between spectral features and direction of arrival of an acoustic event [13]. The two models share the same overall structure, where the latter extends the former by covering the modeling of the cochlea and mimicking human perception in more detail (see the following subsections for a more detailed description). For both models, the first step consists in converting HRTF measurements into *directional transfer functions* (DTFs), which incorporate directional cue of listener acoustics. DTFs are computed for both template and target, and are further processed to obtain the internal representations: a gamma-tone filter bank with a frequency spacing of one equivalent rectangular bandwidth (ERB) is thus applied.

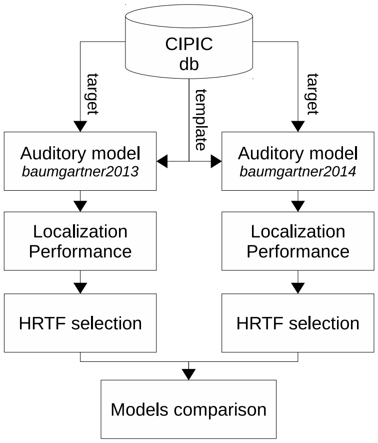


Fig. 1: Structure of the virtual experiments and model comparison.

1) `baumgartner2013`: the model processes each frequency band with a half-wave rectifier and a low-pass filter simulating the inner hair cells. Each band is averaged in time by means of the root-mean-square (RMS) amplitude, providing an internal representation of the sound [6], [8].

Each combination of available elevation angles is compared by computing the *standard deviation* (SD) of the inter-spectral differences between internal representations of target and template. A probabilistic approach is used to map this into the predicted response probability: for each target angle, template angle, and ear, the SD is mapped to a *similarity index* (SI) using a Gaussian function. The SI represents the response probability for the response angle, in degrees.a For a Gaussian function with zero mean, its standard deviation denotes the *uncertainty* (U), modeling the loss of precision due to perceptual process [6].

Furthermore, a binaural weighted sum takes into account the contribution of both ears, obtaining a binaural SI. For a target angle, the binaural SI is computed for each template angle, then a normalization phase allows the definition of the probability mass vector (PMV) describing listener's response probability as a function of the response angle for a given incoming sound.

2) `baumgartner2014`: the model computes the positive spectral gradient from the DTF, relying on the role of the dorsal cochlear nucleus (DCN) which is thought to be crucial for sagittal-plane sound localization [14].

The comparison between template and target is performed using the L_1-norm of the positive spectral gradients responses.

[2] http://amtoolbox.sourceforge.net

[3] www.sofaconventions.org

[4] The internal representation is intended to be the elaboration of the audio stimulus with a mathematical description of the functions performed by the human hearing system.

Since the sagittal-plane localization is considered to be a monoaural process [15] the comparison is performed for each ear, separately.

Furthermore, the model maps this distance into a similarity index, as in baumgartner2013, with the aim of defining human uncertainty in localizing sound sources. The mapping is based upon a sigmoid psychometric function with the following parameters: Γ, denoting the degree of selectivity, and S_l denoting listener specific sensitivity.

Hence the binaural weighting is computed according to variations on the perceived lateral location [16].

The sensori-motor mapping is also taken into account simulating the cognitive and kinematic process that a subject performs while pointing to a sound source location. The authors model this component by manipulating the responses using a Gaussian function with scatter parameter ϵ.Finally, as in the 2013 model, the output has the final form of a probability mass vector representing the prediction of the response probability.

III. Simulations

In this work, we adopt the following approach: for both models, we computed the localization errors, considering individual and non-individual HRTF sets as target for all available HRTF sets; we compared differences in the predictions produced by the two models in order to assess their equivalence in a HRTF selection scenario. A graphical representation of this methodology is depicted in Fig. 1.

We assigned to model parameters the same values for all CIPIC subjects. For the 2013 model: uncertainty $U = 2$. This value was obtained by averaging estimated real Us [6], thus reproducing a realistic perceptual error. For the 2014 model: degree of selectivity $\Gamma = 6$ dB, sensitivity $S = 0.7$, sensori-motor scatter $\epsilon = 17°$. These were the default values proposed for the model [9].

Additionally, the same simulations were repeated removing the torso acoustic information: a 1 ms Hanning window centered on the maximum temporal peak was applied to each DTF. It has to be noted that torso does not produce any shadow effect for the considered elevation angles [12].

A. Perceptual metrics

The perceptual error and differences between target and response angles led us to define four metrics. Two of them account for the absolute localization errors firstly introduced by Middlebrooks [13] and further formalized by Geronazzo *et al.* [12]: the *Polar Error* (PE) and the *Quadrant Error* (QE) are defined for every response angle in $[-45°, 45°]$ computing the comparison between the corresponding template angles with all available target angles. The PE metric accounts for localization judgments occurring into the same hemifield of the response angle, thus being an estimate for precision. The QE metric accounts for all the localization judgments affected by front-back confusions and the responses where the error exceeded $90°$. The PE is defined for every j-th elevation response close to the target position:

$$\mathrm{PE_j} = \sqrt{\frac{\sum_{i \in L} (\phi_i - \phi_j)^2 p_j[\phi_j]}{\sum_{i \in L} p_j[\phi_i]}} \tag{1}$$

with $L = \{i \in N : 1 \le i \le N_\phi,\, |\phi_i - \phi_j| \, mod \, 180° < 90°\}$, where ϕ_i, ϕ_j represent the local response and the target position respectively and $p_j[\phi_i]$ denotes the probability mass vector.

Instead the QE error is formalized as:

$$\mathrm{QE_j} = \sum_{i \in NL} p_j[\phi_i] \tag{2}$$

with $NL = \{i \in N : 1 \le i \le N_\phi,\, |\phi_i - \phi_j| \mod 180° \ge 90°\}$, and for the j-th elevation response.

In addition, two more metrics are considered, which are gathered from recent scientific literature [12]: the *Front-Back confusion rate* (FB), and the *Global Polar Error* (GPE). The FB error models the perceptual confusion when a frontal sound source is localized by the listener back side and vice versa; the GPE quantifies the absolute angular localization error with front-back resolution. These metrics were considered in order to further describe static listening conditions with non-individual HRTF set which lead to significant perceptual distortion due to front-back confusion [2]. For the j-th elevation response:

$$\mathrm{FB_j} = \sum_{i \in C} p_j[\phi_i] \tag{3}$$

with $C = \{i \in N : 1 \le i \le N_\phi,\, \phi_i > 120° \text{ if } \phi_j \ge 60° \wedge \phi_i \le 60° \text{ if } \phi_j > 120°\}$. The GPE quantifies the absolute angular localization error not accounting for front-back confusion. For the j-th elevation response, the GPE is defined as:

$$\mathrm{GPE_j} = \frac{\begin{array}{c}\sum_{f \in F} |\phi_f - \phi_j| \left(p_j[\phi_f] + p_{\bar{j}}[\phi_f]\right) \\ + \sum_{b \in B} |\phi_b - \phi_j| \left(p_j[\phi_b] + p_{\bar{j}}[\phi_b]\right)\end{array}}{\begin{array}{c}\sum_{f \in F} p_j[\phi_f] + p_{\bar{j}}[\phi_f] \\ + \sum_{b \in B} p_j[\phi_b] + p_{\bar{j}}[\phi_b]\end{array}} \tag{4}$$

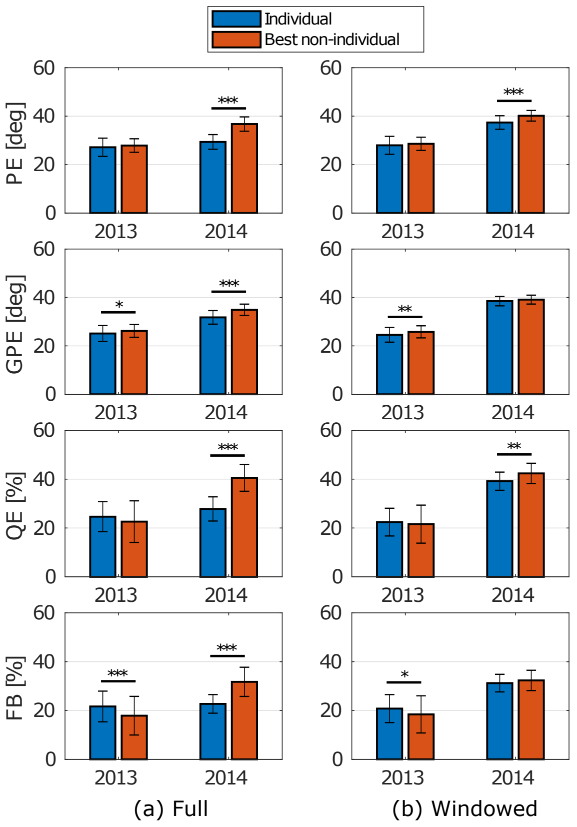
with $F = \{i \in N : 1 \le i \le N_\phi, \phi_i \le 90°\}$ and $B = \{i \in N : 1 \le i \le N_\phi, \phi_i > 90°\}$, where the index $\bar{j}$ has been called the front-back index for to the j-th complementary angle.

Since all metrics are described for a single j-th elevation response, their averages across all responses were taken as a global assessment of each virtual localization experiment.

B. HRTF selection and model comparison

Each of the four metrics was applied to both models, resulting in four *all-against-all* matrices, with row indexes and column indexes spanning template sets and targets sets, respectively.

Then, for each subject localization errors with individual HRTF sets (i.e., error values on the diagonal for each matrix) were compared to the best performance with non-individual HRTF set. Such best performance was found by taking the minimum difference between the template (the individual HRTF set) and all the remaining targets (non-individual HRTF



(a) Full (b) Windowed

Fig. 2: Averaged values for the four proposed error metrics with individual HRTFs vs. "best available" non-individual HRTFs: (a) complete DTFs, and (b) DTFs with torso removed. Asterisks and bars indicate, where present, a significant difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$ according to the statistical test).

sets). To assess whether the best-available non-individual HRTF set was comparable with individual one, a statistical test was performed: PE and QE were evaluated with a *t-test*, while a *Wilcoxon matched-pair signed rank test* was used for GPE and FB, because they did not exhibit a normal distribution [12].

In the end this analysis provided, for each subject, two best available non-individual HRTF sets (each predicted by one of the two auditory models). The correlation between the predictions of the two models for each subject (matrix rows) were computed in order to quantitatively assess the extent to which they can be equivalently employed in the context of a HRTF selection procedure.

IV. Results and Discussion

The results returned from the virtual experiments are reported graphically in Fig. 2; table I reports row-wise correlations that were averaged across all subjects for each metric.

At a first inspection, the predicted localization errors showed similar results, in magnitude, when compared to human prediction errors in elevation [6], [13], both using the individual and non-individual HRTF sets. This suggested a good agreement with the reality. On the other hand, statistical tests exhibited significant differences between individual and non-individual listening conditions for the 2014 model. Such differences were absent or weak in the 2013 model. This discrepancy in the latter model could be attributed to the systemic greater perceived error when the best non-individual HRTF set is imposed.

Metric	Torso		Windowing	
	avg	std	avg	std
PE	0.995	0.001	0.996	0.001
QE	0.978	0.012	0.981	0.013
GPE	0.996	0.001	0.997	0.001
FB	0.969	0.020	0.976	0.019

TABLE I: Model correlations using the predicted metrics. Column *Torso* reports results using full DTFs , while column *Windowing* reports results for the DTFs filtered with a Hanning window to remove torso reflections. Values were averaged on all subjects.

Further insight can be gained by looking at the results when acoustic contribution of the torso was removed through DTF windowing. It can be noticed that the best non-individual HRTF sets were more comparable, from a statistical point of view, with the individual HRTF set, at least for the `baumgartner2013` model. For the 2014 model, this improved similarities were probably related to an increased magnitude of perception errors for both considered HRTF sets.

Despite differences in the average values of the error metrics (the 2014 model systematically produces larger errors), and the different outcomes from statistical tests in terms of significance, the models produced highly correlated results for each subject (see Table I).This outcome was also confirmed by the similar trend that the two models showed on ordered non-individual HRTF (see Fig. 3). Moreover, from Fig. 3, it can be noticed that the 2014 model returned lower variance than the 2013 model. The regression lines were computed, reporting no noticeable difference in slopes: for full DTFs (Fig. 3.a) values were 0.46, 0.54, whereas for DTFs with torso removed (Fig. 3.b) values were 0.44, 0.42, for the 2013 and 2014 models respectively. These similar trends are interesting, since the two models rely on very different assumptions to imitate the same perceptual process.

Despite these high correlations and similar trends, we are inclined to consider that the 2013 model's outcomes could be more fitted to reality. This consideration is supported by the validation of the `baumgartner2013` model through a comparison with human subjects [6], while no similar assessment has been performed for the latter model at the time of writing.

It has to be stressed that 2014 model tended to select the individual set as the best for the subjects in terms of performances, while the 2013 model was inclined to select a non-individual HRTF set that has similar performances to the individual with both full DTFs and torso acoustics removed. From the literature, Asano *et al.* [17] reported

ISBN 978-90-827970-1-5 © EURASIP 2018

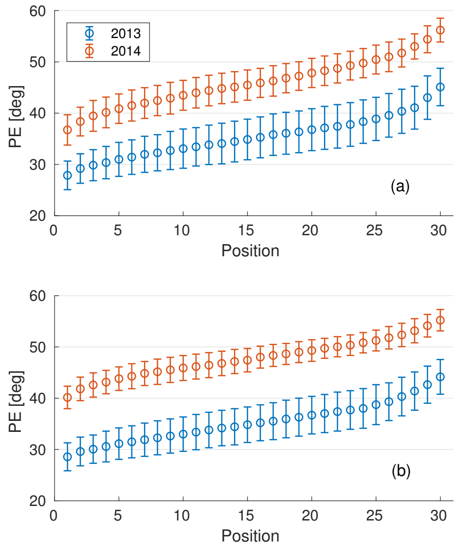


Fig. 3: Averaged polar errors with non-individual HRTF sets based on rank position. The computation was performed on sorted PE prediction in both models. (a) Complete DTFs, and (b) DTFs with torso removed.

vertical localization performances of two expert listeners with non-individual HRTFs as accurate as with individual sets. On the other hand, the best performance in a localization task is typically achieved when subjects use their individual HRTF sets [13]: this behavior was probably empathized by the introduction of the positive spectral gradients in the 2014 model although this conclusion was slightly supported by the performed statistical tests.

V. Conclusions

This work explores differences in human perception of sound source location by imposing non-individual HRTF listening for a localization task in the mid-sagittal plane. The evaluation was performed on a subset of the CIPIC HRTF database using two different auditory models. These models mimic human hearing processing from a functional point of view. This methodology is motivated by the opportunity of using virtual experiments instead of time-demanding psychoacoustic tests, which require expensive and technologically advanced setups with potentially unreliable results [3], [12].

Four different perceptual metrics were implemented in order to identify relevant differences between auditory models in the HRTF selection problem. Even though the outcomes from our analysis reported a high correlation between models, the `baumgartner2014` exhibited a clearer distinction between individual and non-individual sets compared to the `baumgartner2013` that gave more relevance to macroscopic patterns rather than local details [17].

Finally by removing the torso reflections it was noticed that predictions with best non-individual HRTF sets resulted similar to individual ones.

The proposed methodology for the HRTF selection task can be extended by introducing different lateral angles for the localization evaluation. Further improvements can be achieved through a different tuning of the individual parameters with screening tests on spectral profiling, and by introducing supplementary metrics based on the anthropometry of the subjects, such as external ear shape [12].

Acknowledgment

The authors would like to thank Leonardo Chiarello for his work in the early phases of this project. This study was supported by the 2016-2021 strategic program “Knowledge for the World” awarded by Aalborg University to MG.

References

[1] C. I. Cheng and G. H. Wakefield, “Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space.” Proc. 107th Conv. Audio Eng. Society, Sep. 1999.

[2] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, “Localization using nonindividualized headrelated transfer functions,” *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123, Jul. 1993.

[3] R. Barumerli, M. Geronazzo, and F. Avanzini, “Round robin comparison of inter-laboratory hrtf measurements – assessment with an auditory model for elevation.” *Proc. of IEEE 4th VR Workshop on Sonic Interactions for Virtual Environments (SIVE18)*, pp. 1–5, Mar. 2018.

[4] B. F. G. Katz and G. Parseihian, “Perceptually based head-related transfer function database optimization,” *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. EL99–EL105, Jan. 2012.

[5] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini, “Improving elevation perception with a tool for image-guided head-related transfer function selection,” in *Proc. of the 20th Int. Conference on Digital Audio Effects (DAFx-17)*, Edinburgh, UK, Sep. 2017, pp. 397–404.

[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, 2014.

[7] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The cipic hrtf database,” in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, 2001, pp. 99–102.

[8] R. Baumgartner, P. Majdak, and B. Laback, “Assessment of Sagittal-Plane Sound Localization Performance in Spatial-Audio Applications,” in *The Technology of Binaural Listening*. Springer, Berlin, 2013.

[9] B. R., P. Majdak, and B. Laback, “Modeling sound-source localization in sagittal planes for human listeners,” *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802, 2014.

[10] E. H. A. Langendijk and A. W. Bronkhorst, “Contribution of spectral cues to human sound localization,” *J. Acoust. Soc. Am.*, vol. 112, no. 4, pp. 1583–1596, Sep. 2002.

[11] M. Geronazzo, S. Spagnol, and F. Avanzini, “Mixed Structural Modeling of Head-Related Transfer Functions for Customized Binaural Audio Delivery,” in *Proc. 18th Int. Conf. Digital Signal Process. (DSP 2013)*, Jul. 2013, pp. 1–8.

[12] ——, “Do we need individual head-related transfer functions for vertical localization? The case study of a spectral notch distance metric,” *IEEE/ACM Trans. Speech Audio Process.*, vol. 26, no. 7, pp. 1243 – 1256, Jul. 2018.

[13] J. C. Middlebrooks, “Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency,” *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1493–1510, Aug. 1999.

[14] L. A. J. Reiss, “Spectral Edge Sensitivity in Neural Circuits of the Dorsal Cochlear Nucleus,” *Journal of Neuroscience*, vol. 25, no. 14, pp. 3680–3691, Apr. 2005.

[15] M. M. Van Wanrooij and A. J. Van Opstal, “Relearning Sound Localization with a New Ear,” *Journal of Neuroscience*, vol. 25, no. 22, pp. 5413–5424, 2005.

[16] M. Morimoto, “The contribution of two ears to the perception of vertical angle in sagittal planes,” *J. Acoust. Soc. Am.*, vol. 109, no. 4, pp. 1596–1603, Mar. 2001.

[17] F. Asano, Y. Suzuki, and T. Sone, “Role of spectral cues in median plane localization,” *J. Acoust. Soc. Am.*, vol. 88, no. 1, pp. 159–168, Jul. 1990.