# PREDICTING THE PRECISION OF ELEVATION LOCALIZATION BASED ON HEAD RELATED TRANSFER FUNCTIONS

*Sascha Dick, Jürgen Herre*

International Audio Laboratories Erlangen,
a joint institution of Universität Erlangen-Nürnberg and Fraunhofer IIS,
Am Wolfsmantel 33, Erlangen, Germany
sascha.dick@audiolabs-erlangen.de

## ABSTRACT

While the human hearing capability for horizontally localized sound sources is mostly based on binaural cues, the localization of elevation along the "cones of confusion" can only rely on spectral cues provided by the directionality of head related transfer functions (HRTF). This paper explores how the magnitude of spectral differences between HRTF affects the availability of spectral cues to perceive and detect different elevations and hence limits localization accuracy. A method to predict localization accuracy is presented, based on analysis of the gradient of HRTF log-spectral differences for elevation positional changes. To this end, the localization accuracy is estimated from an HRTF database with 45 subjects. Validation of the results shows good agreement with results from subjective localization experiments reported in literature.

***Index Terms***— Auditory perception, elevation perception accuracy, just noticeable differences, head related transfer functions

## 1. INTRODUCTION

Recent advancements in immersive audio and virtual reality introduce elevated sound sources and thus the dimension of height into consumer-grade technology (e.g. MPEG-H 3D Audio [1]). A good understanding of the perception of elevated sound sources is instrumental for perceptual audio coding, for the transmission of immersive audio and for efficient rendering algorithms e.g. on mobile devices with limited computational resources.

The human capability of localizing sounds is typically evaluated in listening test experiments that ask participants to determine the absolute direction of presented sound stimuli (e.g. [2–6]), or to distinguish the minimum audible angle for the variation of the position of a sound source (e.g. [7, 8]). However, these types of experiments are very time consuming. Thus, those experiments can only sample the localization rather sparsely, typically covering ca. 30-100 positions, either coarsely distributed over the full hemisphere, or limited to the horizontal and/or median plane in a finer grid.

On the other hand, the perceived sound localization is ultimately the result of the brain's interpretation of the signals arriving at the ears. The sound propagation from a source to the ears is described by head related transfer functions (HRTFs) [9, 10], which can be physically measured and thus be acquired much faster. Available databases sample the HRTF for multiple test subjects in a closely spaced grid ($>$1000 positions [11]). Alternatively, HRTFs can be modelled from the physical properties of the ears, head and torso [12].

Localization of sound sources in the horizontal plane is dominated by binaural cues, i.e. signal properties caused by the different propagation paths between the left and the right ear [13, 14]. The most prominent binaural cues are interaural level differences (ILD), interaural time differences (ITD) and interaural cross-correlation (ICC). Those cues can be approximated e.g. by a spherical head model [13, 15]. However, the perception of height or elevation cannot be explained by binaural cues. Sound sources distributed along the median plane or one of the so called "cones of confusion" share the same distance for both ears, resulting in identical time and level differences [16]. It has been shown that perception of elevation can, however, be attributed to the timbral colorations caused by position dependent filtering by the pinnae as well as the human head and upper body, as described by the individual HRTF [12, 17]. It should be noted, that asymmetries in the human HRTF can still contribute small binaural ILD cues, aiding the perception of elevation and resolving front-back-confusion [18]. Also, allowing head movement naturally introduces binaural cues, aiding with the perception of elevation [19]. However, the following investigation focuses on cases where no binaural cues are available for elevation perception.

In this paper, we present a method to predict the accuracy of sound elevation localization from HRTF data, assuming that the availability of spectral cues is the limiting factor for the perception of elevated sound sources. Consequently, if the differences between the HRTF of two positions are too small to be perceived, the brain will not be able to distinguish these positions, resulting in localization uncertainty.

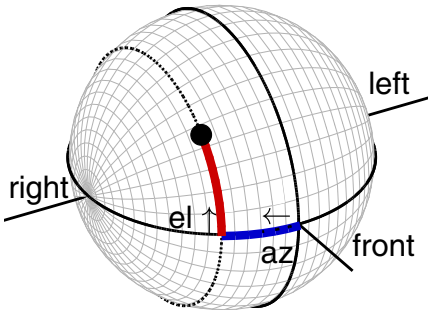


**Fig. 1**. Illustration of spherical interaural coordinate system

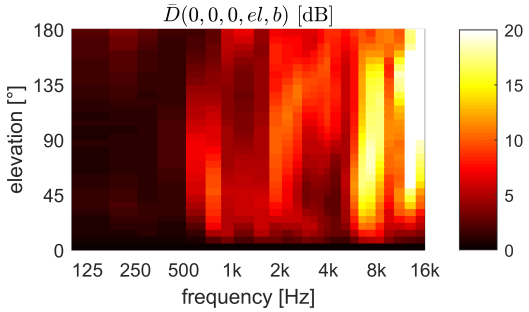


**Fig. 2**. Sum of log-spectral differences for HRTF for different elevations in median plane (vs. 0° frontal position)

## 2. HRTF BASED LOCALIZATION MODEL

### 2.1. Model Assumptions for Perception Accuracy

The absolute localization of a sound source in 3D-space is the result of complex neural processes in the human auditory system. While there are models aiming to predict the absolute localization including binaural cues (e.g. [20, 21]), this paper is focused on analyzing the perception accuracy of elevation.

Without available binaural cues, the difference between sound sources along cones of confusion is only present in the spectral cues due to different HRTFs. We assume that the threshold to detect spectral level differences is independent of their origin. Hence, the ability to distinguish between positions along the cones of confusion is limited by the ability to detect spectral level differences in the HRTF, assuming a sound source with sufficiently dense spectrum (e.g. pink noise). Consequently, the just noticeable difference (JND) for elevation can be derived from the JND for spectral level differences. The JND for level differences was found to be around in the range between 0.5dB to 2dB, dependent on frequency, signal pressure level, etc. [14].

### 2.2. HRTF Spectral Differences

For the analysis of HRTF data, we chose the CIPIC HRTF Database [11]. The database provides anechoic HRTFs for 45 test subjects recorded for 50 elevation times 25 azimuth positions. These positions are defined in the spherical interaural coordinate system as illustrated in Figure 1. Azimuth identifies the cone[1] of confusion in the range of ±90° (full left/right). Elevation indicates the position along the cone of confusion, i.e. the upper hemisphere corresponds to the range of 0° (front) to 180° (back). The database provides scripts for data access and FFT-based frequency response calculation, which we used to calculate the left and right frequency responses $H_{L/R}(\mathrm{az}, \mathrm{el}, k, s)$ of all individual HRTFs per subject $s$, for all available azimuth positions *az* and elevation positions *el* in 512 frequency bins $k$. To reflect the nonlinear frequency resolution of human hearing, we calculate the log-magnitude spectrogram of the energy within bands $b$ following the ERB-Scale [22, 23] (using 27 bands for range 20Hz-16kHz; assuming flat signal spectra within each band), with limits $k_b$ as

$$E_{L/R}(\mathrm{az}, \mathrm{el}, b, s) = 10 \cdot \log_{10} \sum_{k=k_b}^{k_{b+1}-1} \left| H_{L/R}(\mathrm{az}, \mathrm{el}, k, s) \right|^2 . \tag{1}$$

To analyze spectral differences between spatial positions, the sum of spectral differences (SSD) between positions $(\mathrm{az}_1, \mathrm{el}_1)$ and $(\mathrm{az}_2, \mathrm{el}_2)$ is calculated as the sum of the absolute left and right log-spectral differences as

$$\begin{aligned} &D(\mathrm{az}_1, \mathrm{el}_1, \mathrm{az}_2, \mathrm{el}_2, b, s) = \\ &= \quad \left| E_L(\mathrm{az}_1, \mathrm{el}_1, b, s) - E_L(\mathrm{az}_2, \mathrm{el}_2, b, s) \right| + \\ &\quad + \left| E_R(\mathrm{az}_1, \mathrm{el}_1, b, s) - E_R(\mathrm{az}_2, \mathrm{el}_2, b, s) \right| , \end{aligned} \tag{2}$$

which is averaged over all $n_s = 45$ subjects, resulting in

$$\bar{D}(\mathrm{az}_1, \mathrm{el}_1, \mathrm{az}_2, \mathrm{el}_2, b) = \frac{1}{n_s} \sum_s D(\mathrm{az}_1, \mathrm{el}_1, \mathrm{az}_2, \mathrm{el}_2, b, s). \tag{3}$$

We determine each subject's SSD first and then average across subjects, to capture the influence of individual comb-filter effects in the HRTF. Conversely, calculating the average of all subjects' HRTFs would smooth out comb filter effects, and thus potentially result in an underestimation of the individually perceived spectral changes.

Figure 2 illustrates the average spectral differences for elevation along the median plane, compared to the frontal position. The results show prominent differences of up to 18dB for upper positions in the region of 8kHz. This is in accordance with the frequency regions of the "directional bands" found by Blauert [2] to be dominant for the localization of elevated sound sources. Also, differences for rear positions are found (ca. 1kHz: ≤ 7dB; 2-6kHz ≤ 10dB) corresponding to the bands for front vs. back localization ( [2]: 1kHz: perceived in front; 2-6kHz: perceived in back). This finding reinforces our assumption that properties of elevation perception can be derived from the HRTF's physical properties.

[1] We use the common nomenclature "cone of confusion" here. However, it should be noted that no distance dependence is considered, hence the *cones* are represented by *circles* on a sphere.

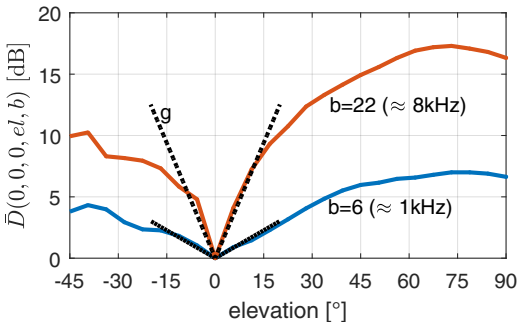


**Fig. 3**. Example of gradient approximation from SSD

### 2.3. Gradient of Spectral Differences

To determine the elevation localization accuracy, we investigate the rate of how fast the received spectral energy changes when diverging from a given position. This rate corresponds to the gradient of the SSD with respect to a given position. Figure 3 illustrates the SSD in the median plane compared against the frontal position for selected bands. The approximate gradient for an elevation $\mathrm{el}_0$ is calculated as average over neighboring measurement points with elevation $\mathrm{el}_i$ that span the range of $\pm 15°$ elevation difference, i.e. $\Delta\mathrm{el}_i = |\mathrm{el}_0 - \mathrm{el}_i| \leq 15°$. A linear, symmetrical approximation of the relation between elevation difference and SSD is defined as function

$$f_i = g \cdot \Delta\mathrm{el}_i. \tag{4}$$

The approximate gradient $g$ is is determined via least-squares fit, i.e. minimizing the sum of squared errors (SSE) for $\bar{D}_i = \bar{D}(\mathrm{az}_1, \mathrm{el}_0, \mathrm{az}_1, \mathrm{el}_i, b)$:

$$\mathrm{SSE} = \sum_i \left(\bar{D}_i - f_i\right)^2 \tag{5}$$

The linear fit is illustrated in Figure 3 with dotted black lines. For instance, the fitted gradient for frequencies around 1kHz at 0° is 0.15 dB/°. Vice versa, this means that a spectral difference of 1dB requires the elevation to change by ≈ 7°.

This is repeated for all positions and bands in the HRTF database, which contains values in the range of $-80° \leq \mathrm{az} \leq 80°$, resulting in an approximation of the gradient of (HRTF induced log-) spectral differences along elevation (GSDE). Figure 4 illustrates the GSDE, averaged over frequency bands, in the the upper hemisphere as viewed from the top (equal-area homolographic projection [24]). As expected from literature, the results show the largest gradients for positions in the front and back, and smaller gradients above the listener (see Sect. 3). Additionally, the gradient decreases for lateral positions towards the interaural axis, as the radius of the circles along the cones of confusion decreases.

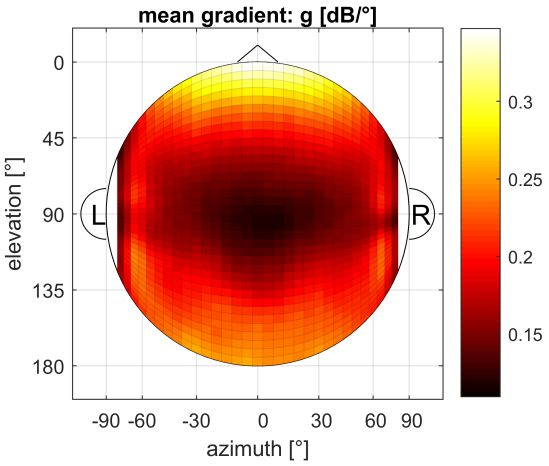


**Fig. 4**. Top view of mean gradient of HRTF log-spectral differences along elevation

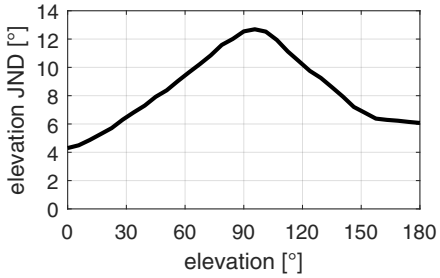


**Fig. 5**. Estimated localization accuracy in median plane

### 2.4. Elevation JND

Assuming a given JND to detect spectral level differences $\mathrm{JND}_{\mathrm{LD}}$, the JND for the perception of elevation differences $\mathrm{JND}_{\mathrm{el}}$ can be computed from the GSDE as

$$\mathrm{JND}_{\mathrm{el}} = \frac{\mathrm{JND}_{\mathrm{LD}}}{g}. \tag{6}$$

Considering a range of 0.5dB to 2dB for $\mathrm{JND}_{\mathrm{LD}}$ [14] and a minimum $\mathrm{JND}_{\mathrm{el}}$ of 4° (white noise in frontal position [13]), we chose $\mathrm{JND}_{\mathrm{LD}} = 1.5$dB in the following. Thus, in the median plane, illustrated in Figure 5, the resulting JND is 4° JND in front, 13° above and 6° in the back of the listener. Figure 6 illustrates the resulting elevation JND for the full upper hemisphere, based on the mean gradient in Figure 4.

## 3. EVALUATION

To evaluate our method, we compare the elevation JND we derived from HRTF analysis against results in literature from subjective localization experiments [3–6], and minimum audible angle experiments [8], focusing on the median plane, where the largest number of data points from different experiments was available. Localization listening test experiments

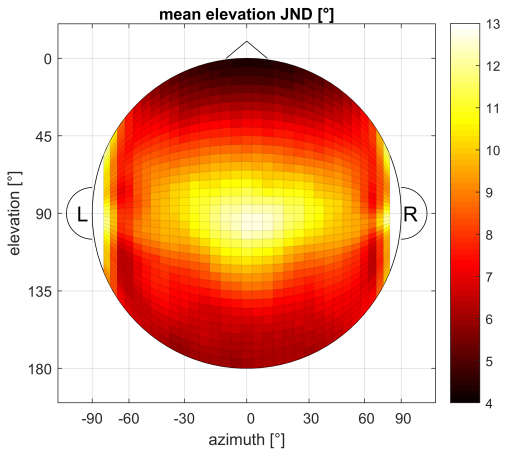


**Fig. 6**. Top view of estimated JND for elevation perception

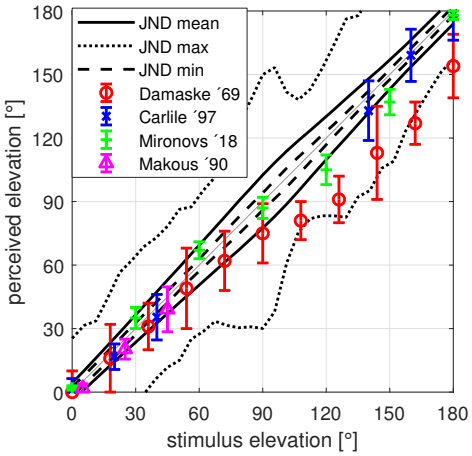


**Fig. 7**. Comparison of HRTF based elevation JND vs. localization listening test results from literature (Damaske '69 [3], Makous '90 [4], Carlile '97 [5], Mironovs '18 [6])

typically yield two types of localization errors: The localization accuracy or consistency for a given position (modelled by our elevation JND) and an absolute bias between the signal's physical and perceived location. Since most of these experiments were not directly focused to the concept of JNDs, different statistical parameters for localization accuracy were reported (e.g. confidence intervals or quartiles).

Thus, we compare the predictions of our JND model to the absolute localization results from the experiments. Figure 7 shows the direction of the presented stimuli versus the perceived direction in the listening tests. The black lines indicate the corridor of localization precision as predicted by our JND model from the mean GSDE. To consider different spectral characteristics of test stimuli we also determined the JND predicted by the bands with the lowest GSDE, yielding the maximum JND (JND max), and respectively the highest GSDE (JND min), as plotted as dotted and dashed black lines.

For elevations up to 90° (front), the mean localization positions from the literature results fall within the corridor of the JND predicted by the mean gradient. For elevations beyond 90° (back) there is an apparent localization bias towards smaller elevation found in literature. Our HRTF based JND model does not aim to explain absolute localization. However, the biased positions are approximated by the range of the maximum predicted JND. This suggests the hypothesis, that without sufficient binaural or spectral cues for accurate localization, the human auitory system tends to localize the position with the smallest elevation that can be plausibly interpreted from the available cues.

Figure 8 shows the relative localization accuracy, i.e. the characteristics of how the accuracy changes for different positions. Here, we normalize the range of localization accuracy within each experiment by mapping the minimum to 0 and the maximum 1 (neglecting localization bias). The modelled JND fits to the overall characteristics, though there is a high variance between the individual experiments.

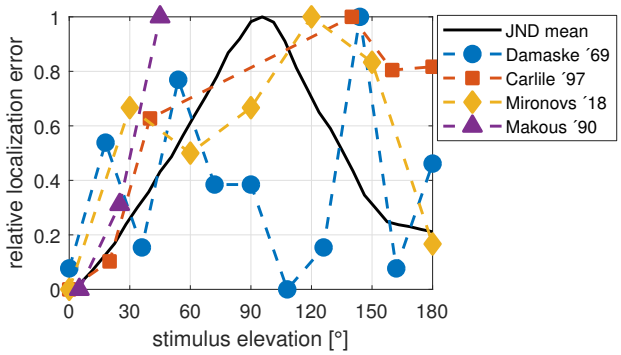


**Fig. 8**. Comparison of relative HRTF based elevation JND vs. listening test experiments from literature (Damaske '69 [3], Makous '90 [4], Carlile '97 [5], Mironovs '18 [6])

## 4. CONCLUSION

In this paper, we showed that spectral properties of the HRTF can be used to explain the JND for perception of elevation along the cones of confusion. Based on the HRTF measurements provided by the CIPIC Database, the gradient of the log-spectral energy differences was approximated. From this, elevation JNDs were then derived, assuming that the ability to detect spectral level differences in the HRTF limits the localization precision. The derived JNDs were validated against listening subjective localization listening test experiments in literature, showing good agreement.

In conclusion, we presented a method to predict elevation precision from HRTF data without requiring subjective localization experiments. As HRTF data can be measured in a fine grid or even be model based, this allows for faster and more precise prediction of localization accuracy. It also shows that, regardless of the brain's interpretation, localization accuracy is limited by the physical properties of sound propagation.

## 5. REFERENCES


[1] ISO/IEC, "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio," International Standard 23008-3, 2015.

[2] Jens Blauert, "Sound localization in the median plane," *Acta Acustica united with Acustica*, vol. 22, no. 4, pp. 205–213, 1969.

[3] P. von Damaske and B. Wagner, "Richtungshörversuche über einen nachgebildeten Kopf," *Acustica*, vol. 21, no. 1, pp. 30–35, 1969.

[4] James C. Makous and John C. Middlebrooks, "Two-dimensional sound localization by human listeners," *The journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, 1990.

[5] Simon Carlile, Philip Leong, and Stephanie Hyams, "The nature and distribution of errors in sound localization by human listeners," *Hearing research*, vol. 114, no. 1-2, pp. 179–196, 1997.

[6] Maksims Mironovs and Hyunkook Lee, "On the accuracy and consistency of sound localization at various azimuth and elevation angles," in *Audio Engineering Society Convention 144*, May 2018.

[7] A. W. Mills, "On the minimum audible angle," *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, 1958.

[8] David R. Perrott and Kourosh Saberi, "Minimum audible angle thresholds for sources varying in both elevation and azimuth," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1728–1731, 1990.

[9] Brian C.J. Moore, *An Introduction to the Psychology of Hearing*, pp. 249–253, Academic Press, San Diego, 5 edition, 2003.

[10] Henrik Møller, Michael Friis Sørensen, Dorte Hammershøi, and Clemen Boje Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc*, vol. 43, no. 5, pp. 300–321, 1995.

[11] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, Oct 2001, pp. 99–102.

[12] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson, "Structural composition and decomposition of HRTFs," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, Oct 2001, pp. 103–106.

[13] Jens Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT Press, Cambridge, Massachusetts, 1997.

[14] Hugo Fastl and Eberhard Zwicker, *Psychoacoustics: Facts and Models*, Springer-Verlag, Heidelberg, 3 edition, 2007.

[15] Neil L. Aaronson and William M. Hartmann, "Testing, correcting, and extending the woodworth model for interaural time difference," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 817–823, 2014.

[16] C. L. Searle, L. D. Braida, M. F. Davis, and H. S. Colburn, "Model for auditory localization," *The Journal of the Acoustical Society of America*, vol. 60, no. 5, pp. 1164–1175, 1976.

[17] Richard O. Duda, V. Ralph Algazi, and Dennis M. Thompson, "The use of head-and-torso models for improved spatial sound synthesis," in *Audio Engineering Society Convention 113*, Oct 2002.

[18] Ramona Bomhardt and Janina Fels, "The influence of symmetrical human ears on the front-back confusion," in *Audio Engineering Society Convention 142*, May 2017.

[19] Stephen Perrett and William Noble, "The effect of head rotations on vertical plane sound localization," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2325–2332, 1997.

[20] Benjamin Hammond and Philip J. B. Jackson, "Full-sphere binaural sound source localization by maximum-likelihood estimation of interaural parameters," in *Audio Engineering Society Convention 142*, May 2017.

[21] J. Woodruff and D. Wang, "Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 806–815, April 2013.

[22] Brian C.J. Moore and Brian R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[23] Jeroen Breebaart, Steven van de Par, and Armin Kohlrausch, "Binaural processing model based on contralateral inhibition. I. model structure," *The Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1074–1088, 2001.

[24] John P. Snyder, *Map projections: A working manual*, pp. 249–252, U.S. Geological Survey Professional Paper 1395, Washington, 1987.