Florence, Italy 12-16 July 2015
ICSV22
The 22nd International Congress on Sound and Vibration

SUBJECTIVE EVALUATION OF A LOW-ORDER PARAMETRIC FILTER MODEL OF THE PINNA FOR BINAURAL SOUND RENDERING

Simone Spagnol, Sandro Scaiella, Michele Geronazzo and Federico Avanzini
University of Padova, Department of Information Engineering, via Gradenigo 6/B, I-35131 Padova, Italy
e-mail: spagnols@dei.unipd.it

The acoustic effects undergone by a sound wave on its way from the source to the listener's ears are summarized in a specific transfer function (Head-Related Transfer Function, HRTF), whose contribution given by the pinna alone is known as PRTF (Pinna-Related Transfer Function). The PRTF of a listener can be approximated with a specific synthetic model that simulates the main crests and troughs of its magnitude response through low-order peak and notch filters, the parameters of which are only partially related to the anthropometry of the listener himself. Starting from this model, the paper describes a psychoacoustic experiment designed with the aim of testing the accuracy of the model - in terms of vertical localization - for sound sources on the median plane and for different combinations of those parameters not related to anthropometry. Results on twelve experimental subjects suggest which combination of parameters offers a lower localization error with respect to the target elevation of the sound source, thus obtaining useful information for the design of structural models for binaural sound synthesis.

1. Introduction

The convincing illusion that a sound source is situated in a given virtual location represents the main objective of any 3D audio rendering system. The idea that lies behind such a system is to present two signals as close as possible to those that a real sound source positioned in that given spatial location would produce at the listener's eardrums. Among the diverse possibilities offered by 3D audio technologies, binaural (i.e. headphone-based) reproduction systems, if properly designed, allow tailoring immersive and realistic auditory scenes to any user without the need for loudspeaker-based systems.

Binaural audio rendering approaches are typically based on the concept of head-related transfer function, or HRTF. HRTFs capture the transformations undergone by a sound wave in its path from the source to the eardrum and in particular those caused by diffraction, reflection and resonance effects onto the torso, head, shoulders and pinnae of the listener. Such characterization allows virtual positioning of a number of sound sources in the surrounding space by filtering the corresponding signals through a pair of HRTFs, thus creating left and right ear signals to be delivered by headphones. In this way, three-dimensional sound fields can be simulated.

Non-individual HRTF sets, typically recorded by using dummy heads, are known to produce evident sound localization errors [1]. On the other hand, obtaining personal HRTF data for a vast

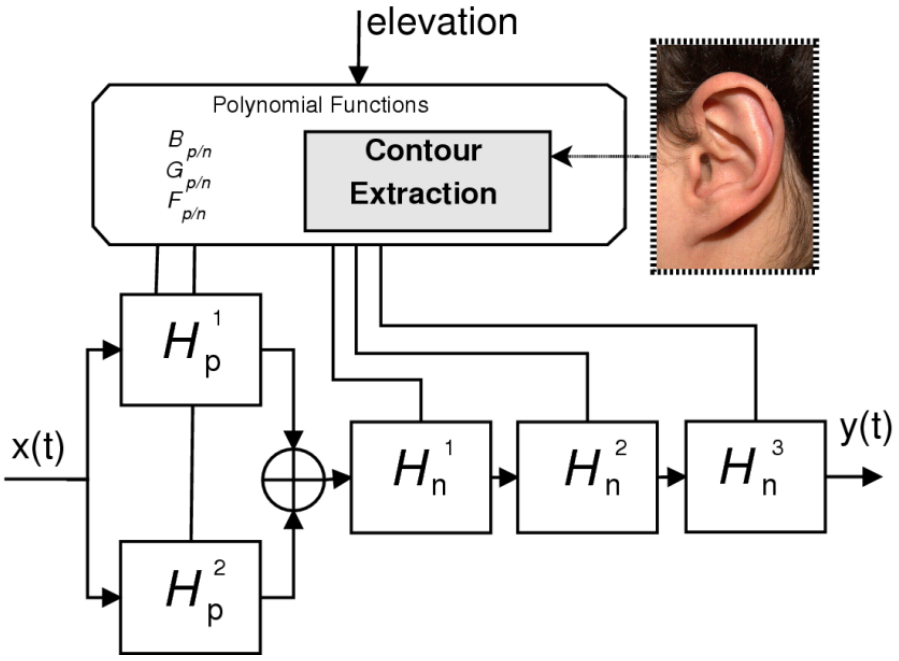

Figure 1: Schematic representation of the structural PRTF model.

number of users is simply impractical because specific hardware, anechoic spaces and long collection times are strictly required. Structural HRTF modeling [2] represents an attractive solution to all of these shortcomings. By isolating the effects of different components (head, pinnae, ear canals, shoulders, torso) and modeling each of them with a corresponding filtering element, the global HRTF is approximated through a proper combination of all the considered effects. Moreover, by relating the temporal/spectral features of each component to corresponding anthropometric quantities, one can in principle obtain an HRTF representation that is both computationally economical and customizable.

A complete structural filter model of the HRTF is currently being studied by the authors [3, 4]. In the model, special care is reserved for the contribution of the external ear to the HRTF (formally defined as *Pinna-Related Transfer Function*, PRTF): data and results collected to date allow in particular the development of a parametric PRTF model customizable according to individual anthropometric data, which in turn can be automatically estimated through straightforward image analysis [5, 6]. This means that by feeding such model with a suitable characterization of the listener's anthropometry and by rendering the resulting audio stream through motion-tracked headphones, low-cost custom binaural audio can be delivered in real time on any device.

In order to evaluate the parameters of the PRTF model that are not related to anthropometric features of the listener's ears [7], in this paper we present the design and implementation of a localization test whose goal is to determine which generic combination of the above parameters' values is able to render the desired sound source elevation as loyally as possible.

2. The structural PRTF model

Typically, the most relevant differences between the HRTFs of two distinct subjects are due to the different pinna features (shape, size, and orientation). The pinna has a fundamental role in shaping HRTFs thanks to two main acoustic phenomena, i.e., reflections and resonances. Consequently, the corresponding PRTF shows a sequence of peaks centered around the resonant frequencies and a sequence of notches located at all those frequencies where destructive interference between direct and reflected waves occurs. The spectral location of these peaks and notches represents a pivotal cue to the characterization of the sound source's spatial position, in particular of its elevation [8].

Previous literature suggests a number of solutions to synthetic PRTF modeling [9, 10]. However,

these models suffer from evident limits, e.g., the validity in an excessively restricted spatial region [9], and/or the absence of an explicit parametrization on the listener's anthropometry [10]. In a previous work [3] the authors proposed a structural PRTF model composed of two filter blocks. As Fig. 1 details, the first block (the *resonant* block) includes two second-order peak filters placed in parallel, while the second block (the *reflective* block) includes three second-order notch filters placed in series.

The authors also studied the relation between notch frequencies in PRTFs and pinna geometry. To this end, a *ray-tracing* procedure on pinna images was exploited to map reflection points at a given distance from the reference ear-canal point, each of which is directly derived from a single notch frequency. In particular, the authors conclude that the use of negative reflection coefficients is crucial in determining notch frequencies. Therefore, the relation between notch frequency and reflection point-ear canal distance can be approximated by the following simple equation,

$$d_i(\phi) = \frac{c}{2f_i(\phi)}, \tag{1}$$

where constant c is the speed of sound, ϕ is the elevation angle for the PRTF, f_i is the center frequency of the i-th notch, and d_i is the distance between the corresponding reflection point and the reference ear-canal point. Reflection points obtained from Eq. (1) were mapped on pinna images of a pool of experimental subjects, resulting in a close correspondence between reflection points and the three main pinna contours, i.e. helix border, antihelix/concha inner wall, and concha border.

As a consequence, if we have an image of the pinna we can trace the above three contours, transform them into a sequence of polar coordinate pairs $(d_i(\phi), \phi)$ with respect to the ear-canal point, and derive from Eq. (1) notch frequencies for every desired elevation ϕ and for each of the three contours. The only independent parameter used in the model is indeed sound source elevation, which drives the evaluation of three polynomial functions (F_n^i, $i = 1, 2, 3$) that interpolate the obtained notch frequencies for a certain sampling step $\Delta\phi$.

For what concerns the bandwidth and gain of notches as parameters, no clear relation with the pinna shape was found. The authors previously approximated these parameters, as well as resonance parameters, using average values from a population of subjects [7]. Understanding what is the impact of the bandwidth and gain parameters in sound localization is thus the research question of the present study.

3. Experimental setup

In order to investigate individual elevation estimation performances in several instances of the structural PRTF model differing just in the notch bandwidth and gain parameters, we conducted a localization test on 12 subjects (3 female and 9 male, ages 23 to 41). Only 4 of these subjects had previous experience with localization tests and none of them underwent any training session. All subjects reported normal hearing defined as thresholds no greater than 25 dB HL in the range of 125 Hz to 8 kHz according to an audiometric screening based on an adaptive maximum likelihood procedure [11].

3.1 Experimental conditions

Let us consider the three parameters associated to a single PRTF notch, i.e., center frequency f_c; gain (or depth) G; and 3-dB bandwidth BW. As already mentioned, f_c is the only parameter associated to individual anthopometric measures. We consider three different values for notch bandwidth: $BW_1 = 0.15 f_c$, $BW_2 = 0.25 f_c$, and $BW_3 = 2$ kHz. BW_2 and BW_3 correspond to the minimum (relative and absolute, respectively) bandwidths thanks to which high-frequency notches are detectable [12, 13]. The choice of BW_1 is instead due to the observation that for such a bandwidth consecutive synthetic PRTF notches show a much lower amount of overlap.

Table 1: The twelve experimental conditions.

Condition	Gain	Bandwidth	#notches
C_1	-10 dB	$0.15 f_c$	3
C_2	-10 dB	$0.15 f_c$	2
C_3	-10 dB	$0.25 f_c$	3
C_4	-10 dB	$0.25 f_c$	2
C_5	-10 dB	2 kHz	3
C_6	-10 dB	2 kHz	2
C_7	-30 dB	$0.15 f_c$	3
C_8	-30 dB	$0.15 f_c$	2
C_9	-30 dB	$0.25 f_c$	3
C_{10}	-30 dB	$0.25 f_c$	2
C_{11}	-30 dB	2 kHz	3
C_{12}	-30 dB	2 kHz	2

For what concerns the notch depth parameter, we consider 2 different values: $G_1 = -10$ dB, i.e., the minimum threshold for which a high-frequency notch with bandwidth $BW_2 = 0.25 f_c$ is detectable [12], and $G_2 = -30$ dB.

Finally we consider the number of PRTF notches as a further variable, by including PRTFs composed of 2 peaks and 3 notches as well as PRTFs composed of 2 peaks and 2 notches, eliminating the middle one. Such a choice is inspired from the results obtained by Iida *et al.* [14], where the presence of two notches only was found to be sufficient for accurately localizing in the median plane, and is due again to avoid notch overlap. Table 1 lists and labels the 12 experimental conditions resulting from all possible combinations of the chosen bandwidth, depth, and number of notches' parameters.

3.2 Stimuli

The raw stimulus is a train of 3 uniformly distributed Hann-windowed white noise bursts lasting 300 ms each, presented at a sound level of 60 dB(A) measured at the ear canal entrance. Short 250-ms pauses separate every two consecutive bursts. The raw stimulus is processed through a headphone compensation filter [15] and then convolved with a pair of synthetic PRTFs chosen according to elevation and experimental condition. The resonant component of the structural model is fixed for all subjects and receives parameters averaged over a population of subjects, whose values vary with the elevation angle [7].

It is worthwhile to clarify two possibly confounding design choices. First, notice that the presented stimulus is filtered through a PRTF, which is not a complete HRTF. However, the only substantial difference between a median-plane PRTF and a median-plane HRTF lies in the presence of the contribution of torso and shoulders in the latter. Since this contribution represents a weak low-frequency elevation cue only, we can assume $PRTF(\phi) \approx HRTF(\phi)$ in the median plane. Second, although both parameters associated to resonances and notch frequencies vary with elevation, notch bandwidth and gain remain fixed along the whole elevation range within each experimental condition. However, following the results of a previous study on a HRTF database, as elevation increases notch depth decreases and bandwidth increases on average [7]. Still, the above experimental choice was due to the difficulty in detecting a common trend of these two parameters within the analyzed HRTF set.

3.3 Protocol

Acquisition of a profile picture of the subject is the first step performed in order to have a representation of his left pinna. In a second phase, the picture is first rotated in order to horizontally align the tragus with the nose tip; then, the maximum protuberance of the tragus is chosen as the ear-canal

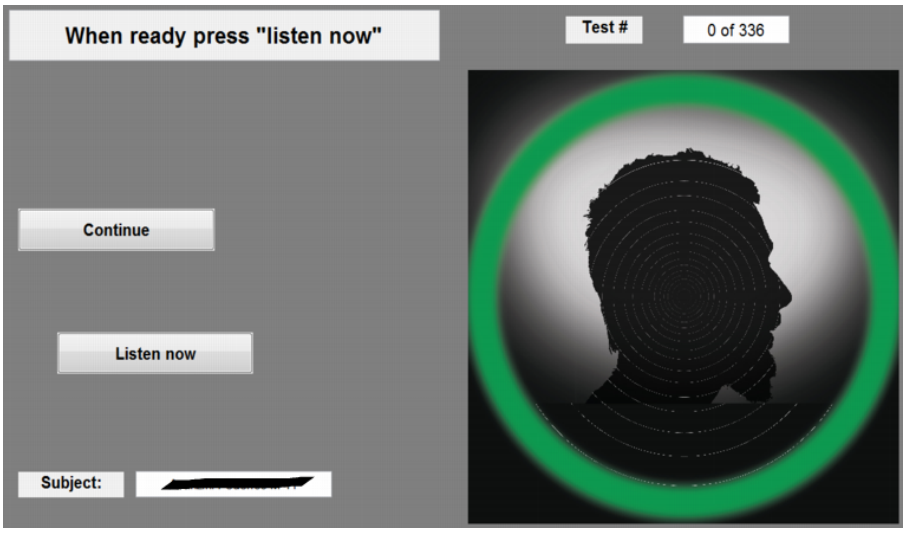

Figure 2: The GUI used in the localization test.

entrance point. The three main contours of the pinna are manually traced and then used to calculate notch frequencies as previously described in Section 2. The subject then enters a Sound Station Pro 45 silent booth and wears a pair of Sennheiser HDA 200 headphones plugged to a Roland Edirol AudioCapture UA-101 external audio card working at 44.1 kHz sampling rate.

The tested elevations are $\phi \in [-45°, -30°, -15°, 0°, 15°, 30°, 45°]$, all in the frontal half of the median plane, where a $0°$ elevation corresponds to a source directly in front. Elevations higher than $45°$ were discarded because of the general lack of spectral notches in the corresponding HRTFs [16], whereas for elevations lower than $-45°$ posture issues would have complicated elevation perception.

Each stimulus related to a single elevation and condition is diotically presented 4 different times. Considering then 7 elevations, 12 conditions and 4 repetitions, we obtain a total of $7 \times 12 \times 4 = 336$ stimuli separated into 4 blocks of 84 trials, each one associated to a single repetition. The ordering of stimuli within each block is pseudorandomly computed.

A screen placed in front of the subject shows the graphical interface reported in Fig. 2. At each sound stimulus presentation the subject uses a common mouse to indicate a point inside the green ring (representing the median plane) corresponding to the perceived sound direction. A few tenths of seconds after the click a new stimulus is presented. At the end of each block of trials a countdown frame appears, during which the subject is allowed to take a 3-minute break. The whole localization test duration is 45 minutes on average, breaks included.

4. Results

For each of the 12 subjects and each of the 12 experimental conditions we first calculated the average localization error, defined as the absolute difference between target and perceived elevation having accounted for front/back reversals,[1] and the front/back reversal rate over all stimuli. Since no significant differences among conditions were noticed in the front/back results, we omit reporting them in this paper. Localization error values are instead reported in Table 2.

First of all, we notice a clear difference in terms of absolute localization error between the 4 subjects with previous experience in localization tests (S_3, S_6, S_8, S_{12}) and the remaining subjects. In order to prevent the results of inexperienced subjects from penalizing the global statistics, we decided not to further consider the results of those subjects whose performances are bad (i.e., comparable

[1] Front/back correction is carried out by symmetrically mapping (with respect to the frontal plane) all values included in the back hemisphere into the front hemisphere. Such a practice is common in localization tests [17] and is due to the high reversal rate observed with virtual auditory displays.

Table 2: Average localization error divided per subject and experimental condition. Conditions whose performances are comparable to a random performance are reported in red. Subjects considered for the following analysis along with their best experimental conditions are reported in green.

Subject	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}
S_1	31.4°	30.7°	32.1°	30.5°	33.6°	26.1°	37.2°	28.7°	37.3°	39.6°	40.8°	29.5°
S_2	32.9°	41.3°	40.9°	39.8°	46.2°	41.2°	38.0°	45.0°	50.0°	42.4°	48.3°	35.2°
S_3	20.0°	19.6°	12.3°	17.9°	16.6°	20.1°	13.3°	16.4°	14.4°	16.7°	11.5°	17.8°
S_4	40.6°	47.5°	44.5°	44.0°	39.1°	43.0°	47.5°	36.3°	47.3°	33.4°	52.6°	45.2°
S_5	24.1°	24.8°	24.3°	29.1°	22.7°	28.1°	24.7°	25.5°	25.6°	26.3°	23.1°	24.7°
S_6	26.9°	20.3°	22.8°	24.8°	27.0°	17.9°	20.5°	16.8°	26.6°	17.0°	19.4°	23.6°
S_7	35.2°	42.4°	47.6°	38.4°	47.2°	40.4°	36.3°	40.7°	43.4°	45.7°	45.5°	46.6°
S_8	19.9°	21.2°	20.4°	18.9°	17.5°	23.4°	15.9°	19.3°	19.3°	16.7°	16.6°	17.9°
S_9	30.4°	41.0°	36.3°	26.6°	28.6°	25.4°	37.8°	26.1°	35.1°	36.2°	32.4°	39.4°
S_{10}	39.6°	41.3°	46.6°	42.1°	32.1°	40.8°	45.5°	38.4°	36.3°	48.6°	45.8°	53.2°
S_{11}	35.3°	47.8°	44.0°	49.0°	42.4°	46.1°	39.9°	43.6°	26.6°	46.4°	28.6°	39.4°
S_{12}	33.6°	38.7°	31.4°	36.5°	29.0°	45.4°	35.2°	33.7°	34.3°	42.6°	28.7°	41.5°

Table 3: Results averaged on the eight good localizers. The three best conditions per error metric are reported in red.

Condition	Error	U/D reversals	slope+r^2
C_1	27.7°	17.0%	0.52
C_2	30.5°	16.5%	0.47
C_3	27.9°	17.0%	0.65
C_4	29.1°	16.1%	0.51
C_5	27.2°	16.1%	0.65
C_6	29.1°	16.1%	0.48
C_7	28.1°	19.6%	0.59
C_8	26.3°	12.5%	0.69
C_9	27.4°	12.5%	0.70
C_{10}	30.2°	17.9%	0.48
C_{11}	25.2°	12.9%	0.82
C_{12}	29.2°	16.5%	0.42

to a random performance) in 8 conditions out of 12 or more. Thus, we calculated the localization error associated to a generator of pseudorandom numbers uniformly distribuited in the $[-90°, 90°]$ range, finding that the average error converges to a value of $49°$ and its fifth percentile to a value of approximately $40°$. The latter value was considered as the average localization error threshold above which the performance associated to a single experimental condition is labeled as bad. As a consequence, subjects S_2, S_4, S_7 and S_{10} were discarded from the following analysis.

The following Table 3 reports results per condition averaged onto the 8 remaining subjects, i.e., the good localizers. In particular, this table reports

- the average localization error, including all values above threshold (error $> 40°$);
- the *up/down* reversal rate, calculated with a tolerance of $15°$ around the horizontal plane for all target elevations except $\phi = 0°$;
- the sum of the slope and r^2 (coefficient of determination) parameters of the regression line computed between target and perceived elevation.

If we compare the three data columns we can easily conclude that condition C_{11} stands out as

being the most effective. Looking back to Table 2 we notice that C_{11} yields, for 5 among the 8 good localizers, the first or the second lowest localization error, and that only Subject S_1 fails in effectively localizing the virtual sound source. We can also notice that the odd-numbered conditions (i.e., HRTFs including 3 notches) yield lower errors with respect to the associated even-numbered conditions (i.e., HRTFs including 2 notches) with the exception of C_7 and C_8, and that the globally best conditions (C_8, C_9, C_{11}) have the greater notch depth ($G_2 = -30$ dB) in common.

At a first glance, notch bandwidth does not seem to have an impact on the localization scores. Nevertheless, if we consider the localization error results of each of the 8 good localizers, we notice that all those subjects whose best performance is associated to a condition with notch gain $G_1 = -10$ dB (S_1, S_5, S_9) best perform with bandwidth $BW_3 = 2$ kHz. By contrast, all of the three bandwidth values appear in the best-scoring conditions having notch gain $G_2 = -30$ dB (subjects $S_3, S_6, S_8, S_{11}, S_{12}$). An inspection of the magnitude responses of filters associated to conditions C_7, C_9, C_{11} (not reported in this paper) reveals how these are very similar, regardless of the specified bandwidth.

To sum up, results of the above localization test suggest that, in order to maximize the chances of correctly perceiving the elevation of a virtual source, the following notch parameters are required:

– a notch depth greater than 10 dB, in accordance with Moore [12];
– the presence of the second notch in the $7 - 10$ kHz range, in accordance with Iida *et al.* [14];
– a sufficiently large bandwidth, i.e. fixed at 2 kHz, in accordance with Alves-Pinto *et al.* [13].

5. Conclusions and future work

The aim of the localization test described in this paper was to analyze the influence of parameters such as notch depth, notch bandwidth and the number of notches included in a synthetic PRTF model on elevation perception of median-plane virtual sound sources. Results of the localization test substantially confirm the indications found in previous literature concerning those parameters. Furthermore, these offer an encouraging acknowledgment to the effectiveness of the structural model in rendering elevation of a virtual sound source for those subjects that had previous experience with localization tests.

Nevertheless, further localization tests comparing customized synthetic HRTFs and individually measured HRTFs are needed in order to have a strong validation of the model itself. Regarding PRTF parameters, further issues worth examining are customization of the resonant component of the PRTF based on the concha shape [18], as well as the use of elevation-dependent notch depth and bandwidth parameters.

The structural PRTF model as it currently is represents a notable extension of other pinna models available in the literature as it is easily customizable and includes a large portion of the frontal hemisphere, thus resulting suitable for real-time control of virtual sources in a number of applications involving frontal auditory displays. Further extensions of the model may include source projection behind, above, and below the listener.

Acknowledgment

This work was supported by the research project Personal Auditory Displays for Virtual Acoustics, University of Padova, under grant no. CPDA135702.

REFERENCES

1. H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi. Binaural technique: Do we need individual recordings? *J. Audio Eng. Soc.*, 44(6):451–469, June 1996.

2. C. P. Brown and R. O. Duda. A structural model for binaural sound synthesis. *IEEE Trans. Speech Audio Process.*, 6(5):476–488, September 1998.

3. S. Spagnol, M. Geronazzo, and F. Avanzini. On the relation between pinna reflection patterns and head-related transfer function features. *IEEE Trans. Audio, Speech, Lang. Process.*, 21(3):508–519, March 2013.

4. S. Spagnol and F. Avanzini. Distance rendering and perception of nearby virtual sound sources with a near-field filter model. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2015. Submitted for publication (Jan. 2015).

5. S. Spagnol, M. Geronazzo, D. Rocchesso, and F. Avanzini. Synthetic individual binaural audio delivery by pinna image processing. *Int. J. Pervasive Comput. Comm.*, 10(3):239–254, July 2014.

6. S. Spagnol, D. Rocchesso, M. Geronazzo, and F. Avanzini. Automatic extraction of pinna edges for binaural audio customization. In *Proc. IEEE Int. Work. Multi. Signal Process. (MMSP 2013)*, pages 301–306, Pula, Italy, September-October 2013.

7. M. Geronazzo, S. Spagnol, and F. Avanzini. A head-related transfer function model for real-time customized 3-D sound rendering. In *Proc. INTERPRET Work., SITIS 2011 Conf.*, pages 174–179, Dijon, France, November-December 2011.

8. S. K. Roffler and R. A. Butler. Factors that influence the localization of sound in the vertical plane. *J. Acoust. Soc. Am.*, 43(6):1255–1259, June 1968.

9. P. Satarzadeh, R. V. Algazi, and R. O. Duda. Physical and filter pinna models based on anthropometry. In *Proc. 122nd Conv. Audio Eng. Soc.*, pages 718–737, Vienna, Austria, May 2007.

10. K. J. Faller II, A. Barreto, and M. Adjouadi. Augmented Hankel total least-squares decomposition of head-related transfer functions. *J. Audio Eng. Soc.*, 58(1/2):3–21, January/February 2010.

11. D. M. Green. A maximum-likelihood method for estimating thresholds in a yes-no task. *J. Acoust. Soc. Am.*, 93(4):2096–2105, April 1993.

12. B. C. J. Moore, S. R. Oldfield, and G. J. Dooley. Detection and discrimination of spectral peaks and notches at 1 and 8 kHz. *J. Acoust. Soc. Am.*, 85(2):820–836, February 1989.

13. A. Alves-Pinto and E. A. Lopez-Poveda. Detection of high-frequency spectral notches as a function of level. *J. Acoust. Soc. Am.*, 118(4):2458–2469, October 2005.

14. K. Iida, M. Itoh, A. Itagaki, and M. Morimoto. Median plane localization using a parametric model of the head-related transfer function based on spectral cues. *Appl. Acoust.*, 68(8):835–850, August 2007.

15. A. Lindau and F. Brinkmann. Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings. *J. Audio Eng. Soc.*, 60(1/2):54–62, January 2012.

16. S. Spagnol, M. Hiipakka, and V. Pulkki. A single-azimuth pinna-related transfer function database. In *Proc. 14th Int. Conf. Digital Audio Effects (DAFx-11)*, pages 209–212, Paris, France, September 2011.

17. F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening.II: Psychophysical validation. *J. Acoust. Soc. Am.*, 85(2):868–878, February 1989.

18. P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Frequency and amplitude estimation of the first peak of head-related transfer functions from individual pinna anthropometry. *J. Acoust. Soc. Am.*, 137(2):690–701, February 2015.