ENHANCING VERTICAL LOCALIZATION WITH IMAGE-GUIDED SELECTION OF NON-INDIVIDUAL HEAD-RELATED TRANSFER FUNCTIONS

Michele Geronazzo, Simone Spagnol, Alberto Bedin and Federico Avanzini

Department of Information Engineering
University of Padova, Italy
{geronazzo,spagnols,bedinalb,avanzini}@dei.unipd.it

ABSTRACT

A novel approach to the selection of generic head-related transfer functions (HRTFs) for binaural audio rendering through headphones is formalized and described in this paper. A reflection model applied to the user's ear picture facilitates extraction of the relevant anthropometric cues that are used for selecting two HRTF sets in a database fitting that user, whose localization performances are evaluated in a complete psychoacoustic experiment. The proposed selection increases the average elevation performances of 17% (with a peak of 34%) with respect to generic HRTFs from an anthropomorphic mannequin. It also significantly enhances externalization and reduces the number of up/down reversals.

***Index Terms*—** spatial hearing, binaural audio, HRTF

1. INTRODUCTION

One of the main limitations of binaural audio through headphones that cause its exclusion from commercial applications in virtual and augmented reality lies in the lack of individualization of the entire rendering process. Since recording individual head-related transfer functions (HRTFs, i.e. the frequency- and location-dependent acoustic transfer functions between the sound source and the eardrum of a listener) is both time- and resource-expensive, obtaining reliable HRTFs for a particular subject in different and more convenient ways is desirable. A common practice employs the trivial selection of an unique HRTF set for all listeners (i.e. recorded on a dummy head built according to mean anthropometric data, such as the *KEMAR* mannequin [1]). However, anthropometric features of the human body have a key role in HRTF shaping: several studies have attested how listening to non-individual binaural sounds results in evident front-back confusion, lack of externalization and localization errors [2].

Computational models generate synthetic HRTFs from a physical [3] or structural interpretation of the acoustic contribution of head, pinna, shoulders and torso. These models have different degrees of simplification, going from basic geometries [4, 5] to more accurate descriptions capable to reproduce the peaks and notches of the HRTF [6] . HRTF spectral details also emerge exploiting principal component analysis (PCA) [7] allowing to further tune the HRTF to a specific listener.

In this work we investigate the alternative approach of selecting non-individual HRTF sets from an existing database, according to two criteria extrapolated from a pinna reflection model [8]. The idea is that the two chosen HRTFs should render better spatial sounds than a generic one (KEMAR) thanks to the closer relation between pinna geometry and localization cues, especially in the vertical dimension.

2. HRTF SELECTION

2.1. Previous works

The last decade registered a notable increase of the number of psychoacoustic tests related to HRTF selection techniques. The most common approach, which we also adopt in this paper, is to use a specific criterion in order to choose the best HRTF set for a particular user from a database. Seeber and Fastl [9] proposed a procedure according to which one HRTF set was selected among 12 based on multiple criteria such as spatial perception, directional impression and externalization. Even though their selection minimized both localization error variance and inside-the-head localization, it was only tested on the frontal horizontal plane. Zotkin et al. [10] selected the HRTF set that best matched an anthropometric data vector of the pinnae (7 parameters), testing the $[-45°, +45°]$ elevation range in the front hemisphere in dynamic conditions. Results showed a general yet not universal decrease of the average elevation error.

Similarly, selection can be targeted at detecting a subset of HRTFs in a database that fit the majority of a pool of listeners. Such an approach was pursued e.g. by So et al. [11] through cluster analysis and by Katz and Parseihian [12] through subjective ratings. The choice of the personal best HRTF among this reduced set is, however, left to the listener.

A different selection approach was undertaken by Hwang et al. [13] and Shin and Park [14]. They modeled HRIRs on the median plane as linear combinations of basis functions whose weights were then interactively self-tuned by the lis-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

teners themselves. Results of the respective tests on a few experimental subjects, although giving mixed results, showed how this method generally reduces the localization error with respect to generic HRTFs, as well as the number of front/back reversals.

2.2. Selection criteria

Thanks to the physical connection between the uniqueness of the listener's pinna shape and elevation cues in sound localization, this work exploits the use of a revised pinna reflection model [15] on a 2-D image as a selection mechanism for HRTFs. According to a ray-tracing method,[1] the three main frequency notches of a specific median-plane HRTF can be extracted with reasonable accuracy by calculating the distance between a point lying approximately at the ear canal entrance (which we refer to as the *focus* point) and each point lying on the three pinna contours thought to be responsible for pinna reflections, i.e. the helix border (C_1 in Fig. 1), the antihelix and concha inner wall (C_2), and the concha outer border (C_3).

Specifically, given the i-th contour C_i, an elevation φ and assuming each reflection to be negative and responsible for a single notch, we calculate the frequency where destructive interference between the direct sound and the sound reflected by the pinna contour occurs as

$$f_0^i(\varphi) = \frac{1}{t_i(\varphi)} = \frac{c}{2d_i(\varphi)} \tag{1}$$

where c is the speed of the sound, $t_i(\varphi)$ the temporal delay between the direct and reflected rays, and $d_i(\varphi)$ the distance between the pinna reflection point and the focus point.

These frequencies were found to closely approximate notch frequencies appearing in the corresponding measured HRTFs of a number of subjects [15]. Given a subject whose personal HRTFs are not available, it is consequently possible for him to select the HRTF set in a database that has the minimum mismatch between the f_0^i frequencies extracted from his own pinna contours and the F_0^i notch frequencies of the available median-plane HRTFs, extracted through a *structural decomposition algorithm* [16]. More formally, the above mismatch is defined as

$$m = \frac{1}{n}\sum_{i=1}^{n}\frac{w_i}{|\varphi|}\sum_{\varphi}\frac{|f_0^i(\varphi) - F_0^i(\varphi)|}{F_0^i(\varphi)}, \tag{2}$$

where n is the maximum number of notches in the available HRTFs in the $4 - 16$ kHz frequency range (typically 3), w_i, $i = 1 \ldots n$ is a convex combination of weights and φ spans all the $[-45°, 45°]$ frontal elevation angles for which the i-th notch is present in the corresponding HRTF.

The relative importance of the pinna contours can be determined by tuning the w_i's. Once fixed, the HRTF set in the database whose mismatch is the lowest is selected.

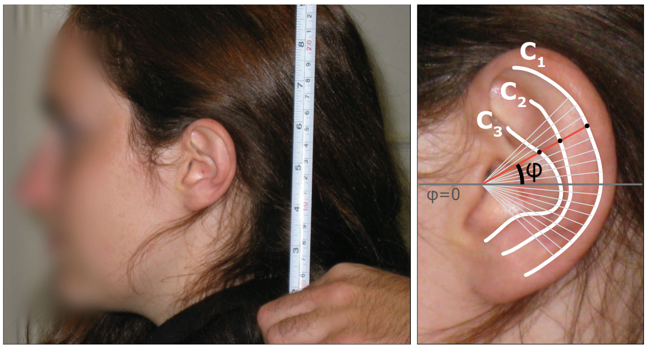

Fig. 1. Side-face picture and pinna contours of one subject.

3. LOCALIZATION TASK

Eight subjects (6 males and 2 females) whose age varied from 22 to 40 (mean 27.4, SD 6.1), took part to the localization task. All subjects reported normal hearing according to the adaptive maximum likelihood procedure proposed in [17].

3.1. Apparatus

The listening tests were performed in a Sound Station Pro 45 silent booth. Sennheiser HDA 200[2] headphones were plugged to a Roland Edirol AudioCapture UA-101 external audio card working at 44.1 kHz sampling rate.

Subjects entered localization judgments in a GUI designed in MATLAB. In the GUI three different frames required judgments of elevation angle, azimuth angle, and externalization. Perceived elevation[3] was entered by manipulating a vertical slider spanning all elevations from $-90°$ to $90°$ which interactively controlled a blue marker moving onto an arc-shaped profile, very similarly to the input interface described in [13]. Perceived azimuth was selected by placing a point in a circular ring surrounding a top view of a stylized human head, inspired by the GUI described in [18]. The externalization judgment simply required the subject to select one of two answers to the question "where did you hear the sound?", i.e. "inside the head" or "outside the head". More details on the software environment can be found in [19].

3.2. Stimuli

All stimuli used as sound source signal a train of three 40-ms gaussian noise bursts with 30 ms of silence between each burst, repeated three times. This type of sound has already been proved to be more effective than a basic white noise burst [12]. The average measured amplitude of the raw stimulus at the entrance of the ear canal was 60 dB(A).

[1] This is possible because in the frequency band where notches appear the wavelength is small enough compared to the dimensions of the pinna.

[2] These dynamic closed circumaural headphones offer an effective passive ambient noise attenuation and high-definition reproduction of high frequencies.

[3] Azimuth and elevation are defined according to the vertical polar coordinate system.

Table 1. Global mean results of the localization task.

	S_1 (KEMAR)	S_2 ($w_1 = w_2 = w_3 = \frac{1}{3}$)	S_3 ($w_1 = 1, w_2 = w_3 = 0$)
Azimuth error (mean/SD)	20.0°±3.0°	21.7°±5.3°	21.3°±4.5°
Elevation error (mean/SD)	31.6°±4.4°	29.9°±5.1°	26.2°±4.7°
Linear fit slope (elevation)	0.20	0.30	0.40
r^2 goodness-of-fit (elevation)	0.10	0.17	0.31
Front/back reversal rate	36.6%	32.9%	34.3%
Up/down reversal rate	18.3%	14.7%	9.0%
Externalization rate	62.2%	64.7%	69.7%

Experimental stimuli were then created by filtering the sound source signal through different HRTF sets and a headphone compensation filter obtained with the algorithm presented in [20] applied to measured responses of a KEMAR mannequin without pinnae. It has to be highlighted that compensation was not individual; however, such kind of processing offers an effective equalization of the headphone up to $8 - 10$ kHz on average and simulates a realistic application scenario where it is not feasible to design personal compensation filters. The HRTF sets were selected among the 45 subjects of the CIPIC database [21].

3.3. Procedure

Acquisition of pinna images was the first step performed in order to compute the mismatch defined in Sec. 2.2. We created an ad-hoc capture environment in order to acquire left side-face pictures of the experimental subjects (see Fig. 1). In a second phase, pictures were first rotated in order to horizontally align the tragus with the nose tip; then, the maximum protuberance of the tragus was chosen as the focus point. Contours C_1, C_2 and C_3 were manually traced and then used to calculate scaled distances from the focus point and consequently the f_0 frequencies as previously described.

For each subject, a fixed HRTF set corresponding to the KEMAR subject with large pinnae (CIPIC ID *21*) was included as control condition. Moreover, two different selection criteria were considered, corresponding to two different convex combinations of the weights in Eq. (2). In summary, for each subject three HRTF sets were selected based on the following criteria:

- criterion S_1: KEMAR subject;
- criterion S_2: minimum m, with $w_1 = w_2 = w_3 = \frac{1}{3}$;
- criterion S_3: minimum m, with $w_1 = 1$, $w_2 = w_3 = 0$.

We verified that for each of the tested subjects S_2 and S_3 select different HRTF sets, denoting an adequate pool of subjects in the database and a reasonable differentiation between the two criteria. We also excluded subject *21* from the candidate selected HRTF sets of S_2 and S_3.

Eighty-five stimuli per HRTF set, each repeated twice, were presented to each experimental subject, for a total of $85 \times 3 \times 2 = 510$ trials. These were generated considering all of the possible combinations of 10 azimuth values (from $-180°$ to $180°$ in $30°$-steps, excluding $\pm 90°$) and 8 elevation values (from $-45°$ to $60°$ in $15°$-steps), plus 5 presentations of the $90°$-elevation point in order to balance the number of stimuli per elevation. Subjects were instructed to enter the elevation, azimuth, and externalization judgments in this specific order for each trial. Each presentation of the 85 positions within a fixed HRTF set, proposed in random order, made up one block of trials, implying that each subject performed a total of 6 blocks. The sequence of presentation of the blocks followed a latin-square design. In order to reduce fatigue of the subject, we added a 3-minute pause between blocks.

4. RESULTS AND DISCUSSION

Localization errors in azimuth and elevation were analyzed separately with front/back confusions on perceived azimuth resolved (with the exception of a $30°$ cone of confusion around $\pm 90°$). Furthermore, linear fitting was performed on the front/back-corrected polar-angle evaluations. One subject who performed elevation judgments at chance performance, corresponding to guessing the direction of the sound (mean elevation error $\approx 45°$), for all three HRTF sets was treated as an outlier and discarded from the analysis.

The mean and SD of localization errors for the three different selections, along with mean linear fit details, front/back and up/down confusion rates,[4] and perceived externalization, are shown in Table 1. Note that the adopted criteria have little effect on azimuth localization; this is reasonable as long as the selection is performed on pinna features only and not on the optimization of interaural differences. Similarly, the mean front/back reversal rate is not greatly affected by the HRTF choice, probably because of the number of dominant factors that contribute to its resolution such as dynamic localization cues. However, S_3 remarkably succeeds in significantly improving both the mean externalization and up/down reversal rates – up/down reversals are more than halved with respect to S_1. We now concentrate on a more detailed analysis of the elevation results.

[4] The up/down confusion rate is calculated with a tolerance of $30°$ in elevation angle around the horizontal plane, and averaged over all target elevations except $\varphi = 0°$.

Table 2. Elevation results divided per subject.

ID	Criterion	S_1	S_2	S_3
SA	Mean elev. error	34.7°	37°	26.7°
	Slope	0.094	0.016	0.281
	r^2	0.023	0.001	0.231
SB	Mean elev. error	25.4°	20.4°	21°
	Slope	0.444	0.670	0.606
	r^2	0.303	0.534	0.534
SC	Mean elev. error	34.9°	31.8°	30.4°
	Slope	0.162	0.231	0.252
	r^2	0.184	0.335	0.341
SD	Mean elev. error	27.1°	29.6°	18°
	Slope	0.286	0.231	0.677
	r^2	0.223	0.143	0.627
SE	Mean elev. error	32.5°	30.5°	29.3°
	Slope	0.077	0.115	0.159
	r^2	0.074	0.073	0.196
SF	Mean elev. error	29.3°	27.6°	29°
	Slope	0.309	0.355	0.317
	r^2	0.192	0.249	0.200
SG	Mean elev. error	37.4°	32.4°	28.3°
	Slope	0.026	0.477	0.500
	r^2	0.002	0.208	0.301

Table 2 illustrates the elevation-related scores of every subject, i.e. mean elevation error, slope of the linear fit, and r^2 goodness-of-fit. Note that S_1 has the average worst performance, while S_3 always scores better results. S_3 gives an average improvement of 17.4% in elevation error with a peak of 33.6% compared to S_1, suggesting that the most external contour, C_1, has high significance for elevation cues. Conversely, S_2 is unreliable as its performance is sometimes the best and sometimes the worst among the three criteria. This could be related to the non-individual headphone compensation that introduces spectral distortion starting from around $8-10$ kHz, where the spectral notches due to the two inner pinna contours generally lie. Consequently, weights assigned to the two inner contours should be differentiated with respect to that of C_1.

More evidence of the benefits brought by S_3 can be appreciated in Fig. 2, which reports elevation scatterplots of subject SG. Note the progressive improvement of the elevation judgments along with the three criteria, witnessed by the rise of both the linear fit slope (red line) and the goodness of fit.

As a separate note, a deeper analysis of the results highlighted that the best elevation performances of S_3 are achieved for sound sources coming from the back (with a mean improvement of the elevation error of 28% compared to S_1). This finding highlights that the HRTF selection criterion, even though developed in the front median plane, is robust and positively affects perception in the posterior listening space too. Finally, since selection was based on a picture of the left pinna, we compared the results for sources in the left and right hemispheres. No significant differences were found, allowing to conclude that for the tested subjects the chosen ear did not influence elevation judgments.

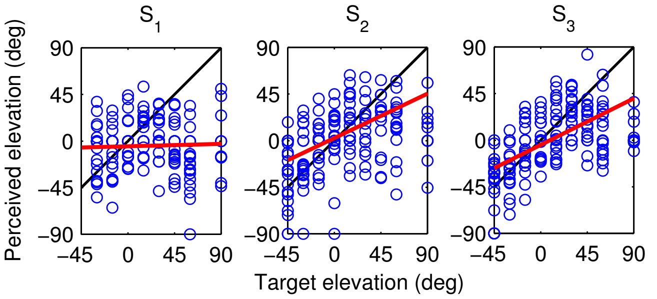

Fig. 2. Elevation scatterplots of subject SG. Black line: ideal response curve. Red line: linear fit of data.

5. CONCLUSIONS

To sum up, the exploitation of the pinna reflection model for HRTF selection is promising and the reported experiment confirms these expectations. Compared to the use of a generic HRTF with average anthropometric data, the pinna reflection approach increases the average elevation performances of 17%, significantly enhancing both the externalization and the up/down confusion rates. The average improvement can be compared to the results found by Zotkin et al. in [10], where the increase of the elevation performance between a generic HRTF and a HRTF selected on anthropometric parameters was reported to be around 20-30% for 4 subjects out of 6. However, a more careful calculation of the average performance on all six subjects shows that the average elevation error decrease is about 6.5%. Still, our results are not directly comparable to theirs because of the different experimental conditions (e.g. presence of head tracking, use of a hand pointer for localization, different elevation range, small number of stimuli).

We found that the selection criterion assigning the whole weight to contour C_1 gives the best results. Indeed, pinna contours may have different weights and could play different roles in the selection. As future work, we are planning to exploit the three contours in a tuning process: while C_1 will be used to pick out the candidate HRTF sets, the other contours will select the "best" HRTF set among the remaining.

It is worthwhile to mention that the experiment was performed in non-optimal experimental conditions (e.g. no individual HRTFs for comparison, non-individual headphone compensation); still, the listening setup comes closely to a feasible scenario for practical applications. In light of this, we are currently developing a tool that automatically extracts pinna contours from a set of 2D images [22]. An extension of the reflection model to three dimensions would greatly improve the accuracy of the extraction and selection processes.

6. REFERENCES

[1] W. G. Gardner and K. D. Martin, “HRTF measurements of a KEMAR,” *J. Acoust. Soc. Am.*, vol. 97, no. 6, pp. 3907–3908, June 1995.

[2] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, “Binaural technique: Do we need individual recordings?,” *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, June 1996.

[3] E. A. Lopez-Poveda and R. Meddis, “A physical model of sound diffraction and reflections in the human concha,” *J. Acoust. Soc. Am.*, vol. 100, no. 5, pp. 3248–3259, November 1996.

[4] R. Teranishi and E. A. G. Shaw, “External-ear acoustic models with simple geometry,” *J. Acoust. Soc. Am.*, vol. 44, no. 1, pp. 257–263, 1968.

[5] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, “Approximating the head-related transfer function using simple geometric models of the head and torso,” *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2053–2064, November 2002.

[6] B. F. G. Katz, “Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation,” *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2440–2448, November 2001.

[7] D. J. Kistler and F. L. Wightman, “A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction,” *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1637–1647, March 1992.

[8] J. Hebrank and D. Wright, “Spectral cues used in the localization of sound sources on the median plane,” *J. Acoust. Soc. Am.*, vol. 56, no. 6, pp. 1829–1834, December 1974.

[9] B. U. Seeber and H. Fastl, “Subjective selection of non-individual head-related transfer functions,” in *Proc. 2003 Int. Conf. Auditory Display (ICAD03)*, Boston, MA, USA, July 2003, pp. 259–262.

[10] D. N. Zotkin, R. Duraiswami, and L. S. Davis, “Rendering localized spatial audio in a virtual auditory space,” *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, August 2004.

[11] R. H. Y. So, B. Ngan, A. Horner, J. Braasch, J. Blauert, and K. L. Leung, “Toward orthogonal non-individualised head-related transfer functions for forward and backward directional sound: Cluster analysis and an experimental study,” *Ergonomics*, vol. 53, no. 6, pp. 767–781, June 2010.

[12] 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, February 2012.

[13] S. Hwang, Y. Park, and Y. Park, “Modeling and customization of head-related impulse responses based on general basis functions in time domain,” *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 965–980, November 2008.

[14] K. H. Shin and Y. Park, “Enhanced vertical perception through head-related impulse response customization based on pinna response tuning in the median plane,” *IEICE Trans. Fundamentals*, vol. E91-A, no. 1, pp. 345–356, January 2008.

[15] 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.*, vol. 21, no. 3, pp. 508–519, March 2013.

[16] M. Geronazzo, S. Spagnol, and F. Avanzini, “Estimation and modeling of pinna-related transfer functions,” in *Proc. 13th Int. Conf. Digital Audio Effects (DAFx-10)*, Graz, Austria, September 2010, pp. 431–438.

[17] D. M. Green, “A maximum-likelihood method for estimating thresholds in a yes-no task,” *J. Acoust. Soc. Am.*, vol. 93, no. 4, pp. 2096–2105, April 1993.

[18] D. R. Begault, E. M. Wenzel, and M. R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source,” *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916, October 2001.

[19] M. Geronazzo, S. Spagnol, and F. Avanzini, “A modular framework for the analysis and synthesis of head-related transfer functions,” in *Proc. 134th Conv. Audio Eng. Soc.*, Rome, Italy, May 2013, number 8882.

[20] A. Lindau and F. Brinkmann, “Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings,” *J. Audio Eng. Soc.*, vol. 60, no. 1/2, pp. 54–62, January 2012.

[21] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The CIPIC HRTF database,” in *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, New Paltz, New York, USA, October 2001, pp. 1–4.

[22] 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)*, Pula, Italy, September-October 2013, pp. 301–306.