Auditory model based subsetting of Head-Related Transfer Function datasets

Spagnol, Simone

Published in:
Proceedings of the 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)

DOI (link to publication from Publisher):
10.1109/icassp40776.2020.9053360

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

AUDITORY MODEL BASED SUBSETTING OF HEAD-RELATED TRANSFER FUNCTION DATASETS

Simone Spagnol

Aalborg University
Department of Architecture, Design & Media Technology
2450 Copenhagen, Denmark

ABSTRACT

The rising availability of public head-related transfer function (HRTF) data, measured on hundreds of different individuals, offers a user the possibility to select the best matching non-individual HRTF from a wide catalogue. To this end, reducing the number of alternatives to a small subset of candidate HRTFs is the first step towards an efficient selection process. In this article a novel HRTF subset selection algorithm based on auditory-model vertical localization predictions and a greedy heuristic is outlined, designed to identify a representative HRTF subset from a catalogue including the three biggest public datasets currently available (373 HRTFs overall). The so-resulting subset (6 HRTFs) is then evaluated on a fourth independent dataset. Auditory model predictions show that for over 95% of the subjects of this dataset there exists at least one HRTF out of the representative subset scoring minimal vertical localization error deviations compared to the best available non-individual HRTF out of the catalogue.

***Index Terms*—** Auditory model, binaural, HRTF selection, sound localization

1. INTRODUCTION

Head-related transfer functions (HRTFs) summarize the direction-dependent acoustic filtering that a free-field sound undergoes due to the head, torso, and pinna. While a pair of HRTFs, one for each ear, can be used to synthesize one or more virtual sounds coming from specific directions, its perception by a particular listener in a particular binaural sound reproduction setup can result in different levels of localization accuracy and spatial immersion. Beside the availability of technological supports such as − to name but a handful − individual headphone equalization, dynamic head tracking, and artificial reverberation [1], one desirable element in any binaural system is the application of individual HRTFs measured on the listener. It is known that generic non-individual HRTFs compared to individual ones are prone to increased localization error and front/back confusion [2].

While it is nowadays possible to calculate the individual HRTF by numerical simulation from a 3D geometry of the head [3], perceptual studies that validate numerically simulated HRTFs against measured ones are rare [4]. Acoustical measurements in controlled environments are therefore still required for obtaining ground-truth individual HRTFs, with high human and technical demands. Providentially, the rising availability of public HRTF data measured on several different human subjects makes it possible in theory for a listener to choose the best fitting HRTF out of hundreds of candidates. This solution, known as *HRTF selection*, can be carried out in its simplest form − wherein HRTFs are selected without any adaptation step − either automatically by means of anthropometry-based matching algorithms [5] or by direct intervention of the listener through perceptual quality evaluation procedures [6, 7]. In the latter case, given that the size of the starting dataset in terms of number of HRTF sets might be high, requiring extensive time for subjective evaluation, it is necessary to reduce it down to a few candidate HRTFs.

Previous work [7] reports reducing a 46-HRTF dataset down to a 7-HRTF subset based on qualitative ratings of median- and horizontal-plane non-individual HRTFs. That study required the time and effort of 45 individuals to determine the perceptually optimized subset. Instead, the present study proposes the use of systematic localization predictions by an auditory model combined with a greedy heuristic to identify a representative subset of candidate best-matching HRTFs from a wide catalogue. Indeed, recent efforts in auditory modeling allow to simulate the localization performance of a measured subject with virtual sounds filtered with individual or non-individual HRTFs [8]. This means that costly and lengthy localization tasks can be approximated using auditory models, making it possible to compute thousands of localization performances in negligible time. Similar to previous studies on HRTF individualization, this study focuses on error metrics for sources on the median plane, which is known to be the most critical region of error increase when moving from individual to non-individual HRTF rendering [2].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 797850.

2. METHODS

2.1. The data

Several public HRTF datasets collecting acoustical measurements on tens of human listeners are available online. Most of these are stored in a common HRTF format known as Spatially Oriented Format for Acoustics (SOFA).[1] However, given the peculiarities of each HRTF dataset, attention must be devoted to merging them while minimizing the dataset-related bias. While some variables either are compensated for in the spectral analysis step of the auditory modeling process (e.g. filter length and sampling frequency, see next subsection) or have minimal impact on the HRTF measurement (e.g. source distance above 1m [9]), differences in the set of spatial measurement directions require special care. Even though spatial interpolation methods may be applied [10], they carry the drawback of introducing error. A simpler and more viable alternative lies in the choice of a subset of common median-plane measurement directions for different datasets.

Accordingly, this study is based on three HRTF datasets (ARI[2] [11], RIEC [12], and HUTUBS [13]) that share 23 common median-plane directions in the $[-30, 210]^\circ$ elevation range (according to the vertical polar coordinate system), with a constant step of 10° and excluding angles 90° and 110°. Following the exclusion of documented dummy-head HRTFs and of a couple of repeated measurements of the same individuals, a total of 373 unique human HRTFs are left,[3] forming the HRTF catalogue under study. A fourth dataset (ITA [14], 48 HRTFs) that also shares the same median-plane directions is considered for the final independent evaluation step.

2.2. The auditory model

The auditory model used herein is the sagittal-plane localization model by Baumgartner et al. [8] included in the Auditory Modeling Toolbox.[4] It is a probabilistic functional model that follows a template-based approach, i.e., assumes that human listeners create an internal *template* of their own HRTF as the result of a learning process, according to which spectral features are mapped to distinct polar angles for a given sagittal plane. When the listener receives an incoming broadband sound, its representation is compared to the internal template; the more similar, the larger the probability of perceiving the sound as coming from the corresponding direction. In the model, the incoming sound is obtained by convolution of a reference stimulus (here, an impulse) and a *target* HRTF.

The model receives its target and template inputs in the form of directional transfer functions (DTF), i.e., HRTFs with the common, direction-independent component across all directions removed as described by Majdak et al. [11]. Then, the target sound is created by convolving the target DTF with the reference stimulus. In order to approximate the spectral analysis carried out by the auditory periphery, the target sound and template are filtered using a gammatone filterbank. Relevant spectral cues are then extracted by positive spectral gradient extraction, resulting in target and template spectral gradient profiles. The comparison process between the two is carried out separately for the left and right channels along the polar response angle (an angle vector spanning the entire sagittal plane) by means of a L_1-norm distance metric, followed by a mapping to similarity indices accounting for listener-specific sensitivity. After combining the two monaural vectors of similarity indices by weighting them with an azimuth-dependent sigmoid function, the resulting vector undergoes a circular convolution with a circular normal distribution that scatters the similarity indices along the polar dimension, simulating the mapping to a motor response in the act of pointing to a target sound. Finally, the vector of similarity indices is scaled such that its sum across all response angles equals one, yielding a probability mass vector (PMV) representing the response probability along the polar dimension.

Focusing now on the median sagittal plane, the psychoacoustic performance metrics used to compare a template and a target HRTF are those known as *quadrant error rate* (QE) and *root mean square local polar error* (PE) [15]. For each target angle, QE is defined as the proportion of polar errors larger than 90° in absolute value, while PE is the root mean square average of polar errors that are less than 90°. The two errors are then averaged across all target angles to yield single-valued QE and PE metrics for every pair of template and target HRTFs. Formally, letting A_j be the set of angles corresponding to local response angles $\theta_i \in R$ to a target angle $\vartheta_j \in T$, $A_j = \{\theta_i \in R : |\theta_i - \vartheta_j| \bmod 180^\circ < 90^\circ\}$, these errors can be computed from PMVs $p_j[\theta_i]$ as follows:

$$QE = \frac{1}{|T|} \sum_{\vartheta_j \in T} \sum_{\theta_i \in R \setminus A_j} p_j[\theta_i], \tag{1}$$

$$PE = \frac{1}{|T|} \sum_{\vartheta_j \in T} \sqrt{\frac{\sum_{\theta_i \in A_j} (\theta_i - \vartheta_j)^2 p_j[\theta_i]}{\sum_{\theta_i \in A_j} p_j[\theta_i]}}. \tag{2}$$

In our simulations all model parameters are set to their default values, including a fixed sensitivity value of 0.7. This value coincides with the mean sensitivity of the virtual listener pool used in Baumgartner et al. [8] to minimize the error between actual and predicted localization performances. Performance metrics are calculated for every possible pair of $N = 373$ template and target HRTFs from our catalogue, yielding the two $N \times N$ (for a total of 139129 comparisons) matrices $\boldsymbol{Q}$ (for QE) and $\boldsymbol{P}$ (for PE) where rows represent template HRTFs, columns represent target HRTFs, and individual HRTF predictions appear on the diagonal.

[1] `http://www.sofaconventions.org`

[2] ARI data includes all the `hrtf b/c` files available as of May 2019.

[3] This study relies on the strong assumption that no two HRTFs from different datasets belong to the same individual, which is reasonable considering that the datasets were collected in different cities.

[4] `http://amtoolbox.sourceforge.net`

2.3. The subsetting algorithm

The goal is to identify a small subset of N HRTFs that fit the large majority of the human sample represented in the catalogue. Given the absence of an absolute criterion for determining whether an HRTF fits a listener based on QE and PE alone, we use a threshold based on a maximum error tolerance heuristic. In particular, a target HRTF fits a virtual listener (i.e., a template HRTF) only when both QE and PE are no greater than the minimum available non-individual QE/PE for the virtual listener times the constant tolerance

$$t_{\boldsymbol{E}} = \frac{1}{N-1} \sum_{i} \frac{P_{10}^{j}(\boldsymbol{E}_{ij})}{\min\limits_{j} \boldsymbol{E}_{ij}}, i \neq j, \tag{3}$$

where $\boldsymbol{E} \in \{\boldsymbol{P}, \boldsymbol{Q}\}$, P_{10}^{j} represents the 10th percentile of the error along the target dimension, and the -1 in the denominator accounts for the individual HRTF entry. By conservatively selecting the average ratio of the 10th percentile to the minimum value of the error, roughly just 10% of the non-individual error values are considered acceptable. After setting the matrices diagonals to infinity (in order not to consider individual HRTFs), the *fitness matrix* $\boldsymbol{F}$ is computed as

$$\boldsymbol{F}_{ij} = \bigwedge_{\boldsymbol{E} \in \{\boldsymbol{P}, \boldsymbol{Q}\}} \boldsymbol{E}_{ij} \leq t_{\boldsymbol{E}} * \min_{k} \boldsymbol{E}_{ik}. \tag{4}$$

In other words, the j-th target HRTF fits the i-th template HRTF (and $\boldsymbol{F}_{ij} = 1$) if and only if both QE and PE fall within the relative tolerances $t_{\boldsymbol{Q}}$ and $t_{\boldsymbol{P}}$; $\boldsymbol{F}_{ij} = 0$ otherwise.

Our question now is how to select a minimum subset of columns of $\boldsymbol{F}$ such that all rows are covered by at least one positive entry. This is an alternative formulation of the *set cover* problem, a very well-known question in operations research that looks for the smallest collection of subsets of a universe whose union covers the universe itself. It is a NP-complete problem, implying that there exists no deterministic polynomial-time solution. However, several polynomial-time approximation algorithms are available, including a classical greedy heuristic that iteratively picks the subset covering the largest portion of still uncovered universe items [16]. Here, the universe is $U = \{1, \ldots, N\}$ and the collection of subsets is $C = \{C_1, \ldots, C_N\}$ with $C_j = \arg_i \boldsymbol{F}_{ij} = 1$.

Algorithm Greedy subset selection

Require: $U = \bigcup\limits_{C_j \in C} C_j$
  $S \leftarrow \emptyset, V \leftarrow U$
  while $|V| > 0.1\,|U|$ **do**
    Choose $C_j \in C \setminus S$ that maximizes $|C_j \cap V|$
    $V \leftarrow V \setminus C_j$
    $S \leftarrow S \cup \{C_j\}$
  end while
  return S

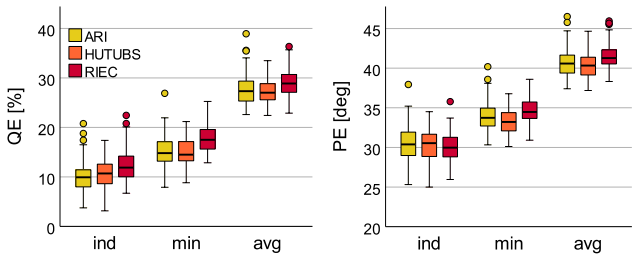

Fig. 1. Polar error metrics (left: QE, right: PE) of individual (ind), best non-individual (min), and average non-individual (avg) target HRTFs for virtual listeners in the catalogue grouped by dataset.

The algorithm is outlined above and differs from its original formulation in that the loop terminates as soon as the subset selection covers 90% of the universe (not the whole of it). This choice is due to the fact that, during its last iterations, the algorithm might select subsets that cover just a few rows and therefore be little representative of the universe.

2.4. Validation metrics

The L selected target HRTFs, whose corresponding indices $j_1, \ldots, j_L \in U$ appear in the selected subset of columns $S = \{C_{j_1}, \ldots, C_{j_L}\}$, are finally evaluated on the independent set of $M = 48$ template HRTFs included in the ITA dataset by comparison with four alternative subset selection methods based on choosing (1) the *most fitting* target HRTFs, corresponding to the L columns of $\boldsymbol{F}$ with highest sum; (2) the *lowest QE* target HRTFs, i.e. the L columns of $\boldsymbol{Q}$ with lowest average value; (3) the *lowest PE* target HRTFs, i.e. the L columns of $\boldsymbol{P}$ with lowest average value; (4) a random subset of L target HRTFs. Performance metrics are calculated as in Section 2.2 against all N target HRTFs from the catalogue, yielding the two $M \times N$ (for a total of 17904 comparisons) matrices $\boldsymbol{Q}^{val}$ (for QE) and $\boldsymbol{P}^{val}$ (for PE). The fitness matrix $\boldsymbol{F}^{val}$ is then computed from $\boldsymbol{Q}^{val}$ and $\boldsymbol{P}^{val}$ as in Eq. (4). Given one of the five selection methods to be compared (with selected HRTF indices $j_1, \ldots, j_L$), the proportion of covered virtual listeners from the ITA dataset serves as the final score

$$S = \frac{\sum_i \bigvee_{j \in \{j_1, \ldots, j_L\}} \boldsymbol{F}_{ij}^{val}}{M} * 100\% \tag{5}$$

3. RESULTS

Figure 1 reports the distribution of QE and PE for individual and non-individual HRTF conditions (minimum and average error by template HRTF) broken down by catalogue dataset. Visual inspection suffices for detecting the expected advantage in using individual HRTFs; indeed, in all cases but a few the best non-individual HRTF (i.e., the one giving

Table 1. HRTF selection by the greedy algorithm and four other methods. Legend: A = ARI, H = HUTUBS, R = RIEC.

Selection method	Selected HRTFs	Score
Greedy algorithm	A62, A129, A789, H23, H26, R8	95.8%
Most fitting	A32, A129, A137, A828, R32, R55	79.2%
Lowest QE	A66, A129, A137, A229, A251, A828	70.8%
Lowest PE	A52, A129, A789, R32, R69, R76	81.2%
Random (average)	(10^8 random subsets)	36.8%

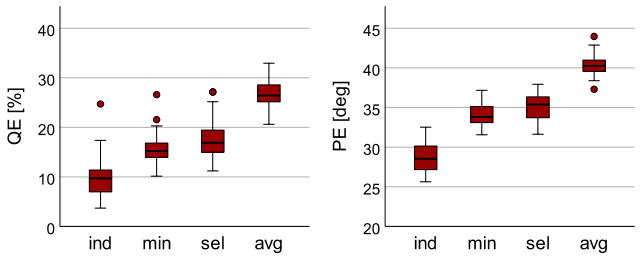

Fig. 2. Polar error metrics (left: QE, right: PE) of individual (ind), best non-individual (min), best selected (sel), and average non-individual (avg) target HRTFs for ITA listeners.

minimum error) scores higher error values than the individual one. On the other hand, one-way analyses of variance[5] with dataset as between-subjects factor and each of the error conditions as dependent variable reveal that, at the $\alpha = .05$ level, the dataset effect is statistically significant ($F(2, 372) \in [10.92, 35.61]$, $p < .001$) except for the individual PE condition ($F(2, 372) = 1.43$, $p < .24$). This effect might be related to different demographics between the populations represented in the datasets.

The proposed algorithm selects $L = 6$ target HRTFs, covering 92% of the virtual listeners in the HRTF catalogue. These HRTFs are listed in Table 1 alongside with those chosen by the four alternative methods. The subset selected by our algorithm achieves the highest score on the validation dataset among all, $S = 95.8\%$ – meaning that only 2 virtual ITA listeners out of 48 do not have a low-error HRTF in the subset. Interestingly, none of the 10^8 random subset generations (out of a total of $\binom{N}{L} \approx 3.6 * 10^{12}$ possible subsets) could achieve a higher score. Another interesting result is that for each of the six selected target HRTFs there exists at least one virtual listener in the ITA dataset covered by that HRTF only: this suggests the high degree of orthogonality of the HRTF subset.

Figure 2 reports the distribution of QE and PE for individual, best non-individual, best selected, and average non-individual target HRTFs for ITA virtual listeners. Notice that the best among the six selected HRTFs scores errors that are extremely close to the overall best non-individual HRTF, with an average difference of 1.95% in QE and $1.06°$ in PE.

It has to be acknowledged that the assumptions underlying

[5]Homogeneity of variance was verified using Levene's test.

Table 2. Most fitting target HRTFs among 46 LISTEN subjects, according to the subjective test in [7] and to the proposed fitness metric.

Subjective test

HRTF	BE	BF	BQ	AZ	BN	AR	BL	AH	AV
No. of subjects	16	16	16	15	14	13	13	12	12

Fitness metric

HRTF	AR	BQ	AZ	BF	BR	BT	BC	BE	BN
No. of subjects	17	14	10	8	8	8	7	7	7

the used ad-hoc metrics do not guarantee that the hypothetical listeners in the ITA dataset would be perceptually satisfied with the best among the six selected HRTFs. In order to partially address this limitation, the $\boldsymbol{F}$ matrix for the 46 LISTEN dataset [17] HRTFs used in [7] was computed (considering all available median-plane angles, as in the previous study).[6] Interestingly, as reported in Table 2, six out of the top nine most fitting target HRTFs according to our metric (i.e. the 9 columns of $\boldsymbol{F}$ with highest sum) coincide with six out of the top nine most selected (i.e. rated as "excellent") HRTFs in the subjective test. Considering that the subjective test differed in that it also included horizontal localization predictions, this result suggests the reliability of the used auditory model in giving localization predictions as well as the relevance of vertical localization in spatial quality perception. A truly subjective test with the proposed subsetting algorithm is planned as future work.

4. CONCLUSIONS

This study suggest that a large HRTF catalogue can be efficiently reduced by two orders of magnitude while preserving at least one HRTF fitting the very large majority of a pool of listeners in terms of localization error in the median plane. Auditory models can act as efficient tools for HRTF evaluation, allowing large-scale localization data analyses with little computational resources. To the best of the author's knowledge, no other study to date has offered evaluations of HRTFs measured on such a high number of different individuals.

This study focused on vertical localization accuracy. While horizontal localization accuracy – that mainly relies on interaural time differences (ITDs) – is of equal importance, when presenting binaural sounds it is good practice not to directly use non-individual ITDs yet couple minimum-phase non-individual HRTFs with an individual anthropometric ITD model [19]. Still, future work in HRTF subsetting/selection might consider the inclusion of models for horizontal localization [20], sound externalization [21], distance perception [22], and ultimately other key perceptual attributes that go beyond the basic issue of localization such as coloration, immersion, and realism [23].

[6]Correspondence between LISTEN IDs and publication IDs was determined thanks to cross-referencing available in [18].

5. REFERENCES

[1] 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, 2001.

[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, 1996.

[3] H. Ziegelwanger, P. Majdak, and W. Kreuzer, "Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization," *J. Acoust. Soc. Am.*, vol. 138, no. 1, pp. 208–222, 2015.

[4] C. Guezenoc and R. Seguier, "HRTF individualization: A survey," in *Proc. 145th Conv. Audio Eng. Soc.*, New York, NY, USA, 2018.

[5] M. Geronazzo, S. Spagnol, and F. Avanzini, "Do we need individual head-related transfer functions for vertical localization? The case study of a spectral notch distance metric," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1243–1256, 2018.

[6] 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, 2003, pp. 259–262.

[7] 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, 2012.

[8] R. Baumgartner, 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.

[9] S. Spagnol, "On distance dependence of pinna spectral patterns in head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 137, no. 1, pp. EL58–EL64, 2015.

[10] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient real spherical harmonic representation of head-related transfer functions," *IEEE J. Select. Topics Signal Process.*, vol. 9, no. 5, pp. 921–930, 2015.

[11] P. Majdak, M. J. Goupell, and B. Laback, "3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training," *Atten. Percept. Psychophys.*, vol. 72, no. 2, pp. 454–469, 2010.

[12] K. Watanabe, Y. Iwaya, Y. Suzuki, S. Takane, and S. Sato, "Dataset of head-related transfer functions measured with a circular loudspeaker array," *Acoust. Sci. & Tech.*, vol. 35, no. 3, pp. 159–165, 2014.

[13] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718, 2019.

[14] R. Bomhardt, M. de la Fuente Klein, and J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model database," in *Proc. 172nd Meet. Acoust. Soc. Am.*, Honolulu, HI, USA, 2016.

[15] 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, 1999.

[16] D. S. Johnson, "Approximation algorithms for combinatorial problems," *J. Comput. Syst. Sci.*, vol. 9, no. 3, pp. 256–278, 1974.

[17] O. Warusfel, "LISTEN HRTF database," 2003, http://recherche.ircam.fr/equipes/salles/listen/.

[18] A. Andreopoulou and B. F. G. Katz, "Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees," *J. Multimodal User In.*, vol. 10, no. 3, pp. 259–271, 2016.

[19] R. Bomhardt, M. Lins, and J. Fels, "Analytical ellipsoidal model of interaural time differences for the individualization of head-related impulse responses," *J. Audio Eng. Soc.*, vol. 64, no. 11, pp. 882–894, 2016.

[20] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605, 2011.

[21] R. Baumgartner, P. Majdak, H. S. Colburn, and B. G. Shinn-Cunningham, "Modeling sound externalization based on listener-specific spectral cues," *J. Acoust. Soc. Am.*, vol. 141, no. 5, pp. 3630, 2017.

[22] S. Spagnol, E. Tavazzi, and F. Avanzini, "Distance rendering and perception of nearby virtual sound sources with a near-field filter model," *Appl. Acoust.*, vol. 115, pp. 61–73, 2017.

[23] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, "Perceptual attributes for the comparison of head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3623–3632, 2016.