ICA 2019 AACHEN
PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

On the evaluation of head-related transfer functions with probabilistic auditory models of human sound localization

Michele GERONAZZO(1), Roberto BARUMERLI(2), and Federico AVANZINI(3)

(1)Dept. of Architecture, Design, and Media Technology, Aalborg University, Denmark, mge@create.aau.dk
(2)Dept. of Information Engineering, University of Padova, Italy, barumerli@dei.unipd.it
(3)Dept. of Computer Science, University of Milano, Italy, federico.avanzini@di.unimi.it

Abstract

Understanding spatial hearing leads to implement efficient and effective auralization rendering algorithms with headphones. Two important aspects contribute to sound localization: (i) acoustic filtering of listener body, and (ii) non-acoustic factors introduced by auditory periphery. Accordingly, head-related transfer functions (HRTFs) describe users acoustics in terms of their spatial filtering. Binaural synthesis through generic HRTFs (commonly a dummy head) is the most simple solution for an auralization framework. In this scenario, a high variability in localization tasks between subjects yields to an unreliable rendering. Listener's acoustic and perceptual characterization require HRTF modeling and auditory models predictions in order to provide an effective auralization on individual basis. Systemic comparisons of HRTF approximations and different user profiles can help to predict listener's performances. We consider a case study on both vertical and horizontal localization with different HRTFs and two probabilistic auditory models. In our analysis, spatial audio rendering with non-individual HRTFs has a special attention for its commercial relevance compared to unpractical and questionable use of individual HRTFs.

Keywords: head-related transfer function, localization, auditory model, non-individual HRTF

1 INTRODUCTION

According to the *spatial audio quality inventory* (SAQI) [16], localization accuracy is a relevant auditory quality for auralization in immersive virtual/augmented reality (VR/AR) or alternatively for *Virtual Auditory Displays* (VADs) and it is usually quantified via time-consuming psychoacustic experiments with human subjects. This paper deals with both elevation localization cues, i.e. perception of sound source position in the vertical spatial dimension, which is mainly provided by monaural spectral features of the *head-related Transfer Function* (HRTF, the Laplace transform of HRIR), and the complementary information provide by the binaural cues such as *interaural time difference* (ITD) and *interaural level difference* (ILD) for localization in azimuth, i.e. the horizontal dimension. Auditory attributes provided by HRTFs can be alternatively assessed through computational auditory models able to simulate the human auditory system. If the auditory model is well calibrated to the reality and properly validated, a perceptual metric can be developed to predict the perceptual performance of a VAD. Therefore in this work we use two such models to define perceptual metrics that allows to automatically assess the localization performance of non-individual HRTF sets. This is an attractive approach in that it allows to use quantitative predictions (from auditory models) rather than qualitative responses from listening tests, in order to identify relevant connections between listeners and HRTFs.

The main contributions of this paper are twofold. First, it promoted the approach based on systematic evaluation of HRTFs through auditory models allowing a rapid prototyping of HRTF models [9] and HRTF selection procedures [11, 14]. Second, it supports the vision of having several perceptual metrics for localization perception and spatial audio attributed that can be applied to the adaptive estimation of HRTF features and parameters for several listening conditions and tasks.

2 LOCALIZATION WITH HEAD-RELATED TRANSFER FUNCTIONS

The measurement of individual HRTFs usually requires special measuring apparatus and a time-consuming procedure with high uncertainty [5], leading to impractical solutions for real-world applications. Alternative methods for HRTF personalization are usually preferred, which look for a delicate trade-off between audio quality and handiness of the personalization procedure [14].

The most common approach for spatial audio rendering in VR/AR contexts makes use of dummy-head HRTFs, or a specific set of HRTFs for all listeners, without personalization. However, it is known that listening through dummy ears causes noticeable distortion in localization cues [12]. During the last decade, the increase of publicly available HRTF data has boosted research on novel approaches to the selection and modeling of non-individual HRTFs. [1] According to [14], HRTF selection problems can be characterized in terms of three issues:

- **metric domains**: acoustics, anthropometry, and psychoacoustics;
- **spatial ranges**: a subspace around the listener for whom the personalization process results in significant improvements for localization performances, e.g., horizontal or vertical plane only;
- **methods**: computational steps which allow to infer the most appropriate non-individual HRTF set for a listener; pre-processing actions such as *data unification*, *feature extraction* , *dataset reduction*, and *dimensionality reduction* can be performed prior the HRTF selection.

Several approaches can be applied and some examples are anthropometric database matching, linear regression models between acoustical and anthropometric features, subjective selection, or minimization of HRTF differences in the acoustic domain. Once one or a set of best HRTF candidates are identified, listener can also self-tune each HRTF set through spectral manipulations and enhancements, and weight adjustments (all these methods were briefly reviewed here [14]). Finally, in a phase of adaptation to non-individual HRTFs users can obtain multimodal feedback in localization/discrimination tasks and improve their performance [20].

3 HRTF ANALYSIS WITH PROBABILISTIC AUDITORY MODELS

With the increasing number of HRTF datasets publicly available worldwide, models of auditory perception are crucial in automating HRTF analysis and selection processes, that might have a reasonable statistical power for the entire human population. Relevant localization performances describing the perceptual dimensions affected by HRTF set variations should be taken into account in order to identify the required level of individualization of any system for auralization. In the following, we explain the idea beyond our methodology by means of two popular probabilistic auditory models for sound localization, both publicly available in the Auditory Modelling Toolbox [2].

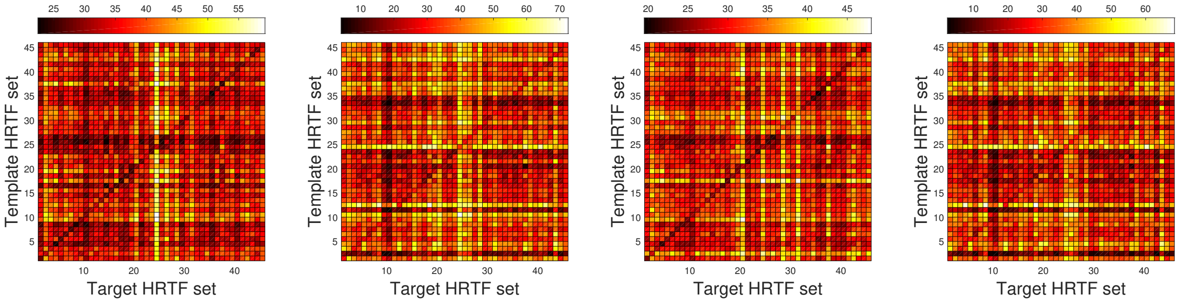
3.1 Vertical localization: Baumgartner 2013

In this paper, we adopt the Baumgartner *et al.* [6], where spectral features of sound events filtered with different HRTFs (target) correlate with the direction-of-arrival (DOA) of the HRTF template, leading to a spectro-to-spatial mapping. This approach is further supported by a recent study of Van Opstal *et al.* [21] where the authors estimated the listeners' spectral-shape cues for elevation perception from the distribution of localization responses.

The model is based on two different processing phases prior to the prediction of absolute elevation. During peripheral processing, an internal representation of the incoming sound is created. The *target* sound is converted into a DTF and filtered with a gammatone filterbank simulating the auditory processing of the inner ear. In the second phase, for each target/template angle and frequency band based on equivalent rectangular bandwidth (ERB), the algorithm computes the gain at the central frequency of each band and the target/template internal representations. The *inter-spectral difference* (ISD) for each band is extracted from the differences in dB between each target angle and all template angles; for each target angle, the *spectral standard deviation* (SSD) of the

[1] See, for instance, the official website of the Spatially Oriented Format for Acoustics (SOFA) project, http://sofaconventions.org

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



(a) (b) (c) (d)

Figure 1. Localization predictions resulting from auditory model simulations on 45 CIPIC subjects in the median plane. *"All-against-all"* matrices for (a) polar error (PE) [deg], (b) quadrant error rate (QE) [%], (c) global polar error (GPE) [deg], and (d) front-back confusion rate (FB) [%].

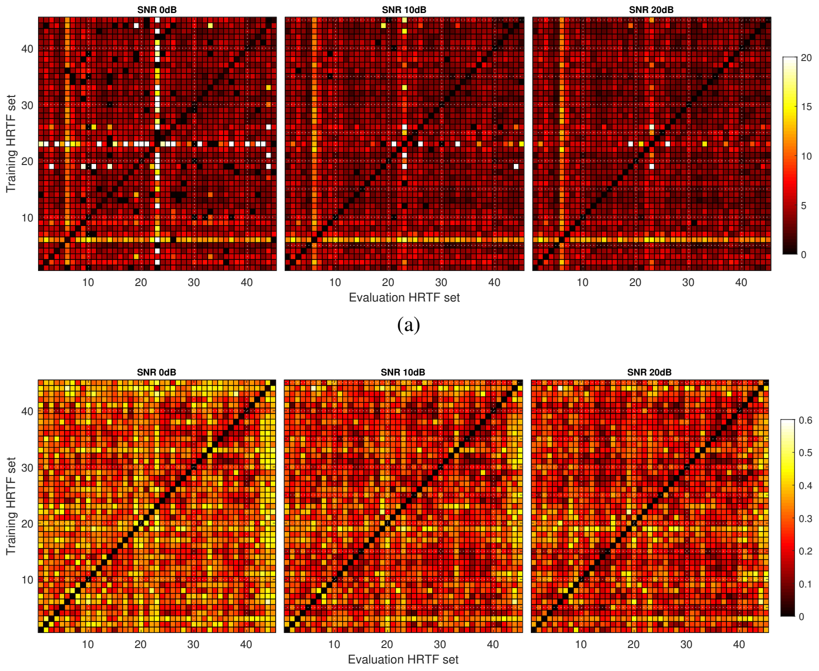
ISD is computed across all template angles. The probability that a virtual listener points to a specific response angle defines the *similarity index* (SI) which receives as input the template-dependent SSD for the argument of a Gaussian distribution with zero mean and standard deviation called *uncertainty*, U. The lower the U, the higher the sensitivity of the listener in discriminating different spectral profiles resulting in a measure of probability. Simulation data are stored in probability mass vectors, where each response angle has the probability that the virtual listener points at it.

3.2 Horizontal localization: May 2011

The auditory model implemented by May et al. in [19] relies on the features space composed by ITD and ILD for the DOA estimation constrained to the frontal horizontal plane at zero elevation. The model adopts the common used place theory developed by Jeffress [15] and it estimates the DOA relying on a machine learning model. The auditory front-end consists of a gammatone filterbank followed by inner hair cell modelling. Each channel of the binaural sound is decomposed with a 4th-order Gammatone filterbank of 32 elements. The center frequencies of the filter-bank are equally distributed between 80 Hz and 5 kHz by means of ERBs. Phase compensation for each channel was required to synchronize the binaural cues over the same time window. Moreover, the gain of each filter is adjusted to follow to the middle-ear transfer function. Later on, an half-wave rectification with square-root compression is used to simulate neural transduction. Binaural cues are calculated using a 20 ms window, overlap of 10 ms, with a sampling frequency $f_s = 44.1$ kHz. Each auditory channel returns the ILD as the ratio of the energy integrated in the time window between left and right ears, and the ITD is extracted for every frame as $(\tau + \delta)/f_s$ where τ is the time lag which maximizes the normalized cross-correlation and δ is the peak position relative to τ. To improve the ITD estimation an exponential integration was applied around τ. Hence, for each window the feature space is represented in Equation 1 where index i represents the gammatone filter and T the number of observations.

$$X_i = \{x_{i,1}, \ldots, x_{i,T}\} = \{(itd_i(1), ild_i(1)), \ldots, (itd_i(T), ild_i(T))\} \tag{1}$$

Finally, azimuth estimation is achieved through a machine learning approach based on the *gaussian mixture model* (GMM). For each gammatone filter a GMM was trained to derive the azimuth α_i from the integration of the features composed of ITD and ILD pairs computed for that specific channel. The training of each GMM was perfomed by means of the iterative *expectation-maximization* (EM) algorithm by considering different source-receiver directions on different positions into a simulated reverberated room.



(a)

(b)

Figure 2. Speech-in-noise localization predictions resulting from auditory model simulations on 45 CIPIC subjects in the horizontal plane with three different SNRs. *"All-against-all"* matrices for (a) horizontal error [deg], and (b) front-back confusion rate [%]. White cells in the matrix (when present) denote values higher than the full-scale.

4 SIMULATIONS

Simulations were initially run on the median plane only, where acoustic properties of the external ear provide vertical localization cues [2] with minimum interference from other localization cues; simulations accounted for the CIPIC database for whom individual HRTF measurements of 45 virtual subjects are available [1]: 2500 HRIRs each, given by the combination of 25 azimuths × 50 elevations × 2 ears, measured at sampling rate $f_s = 44.1$ kHz (200 samples). Elevation ϕ is uniformly sampled on the range $-45°$ to $+230.625°$ in $5.625°$ steps. For each virtual subject, we set an uncertainty value $U = 2$, which reasonably approximates the uncertainty of a real listener in localization tasks [18].

For the second part of this work we extended the models to estimate the DOA for the full horizontal plane. The training method was also based on the Gaussian Mixture Models but we adopted the Multi Conditional Training proposed in the article of Ma *et al.* [17]. The training set consisted of synthetic binaural samples generated by the anechoic MIT HRTF dataset limited to the horizontal plane with zero elevation [7] convolved with the TIMIT speech corpus [8] and combined with diffuse noise at three *Signal-to-Noise Ratios* (SNRs) $0, 10, 20$ dB. The diffuse noise used the sum of 72 uncorrelated, white Gaussian noise sources, each one assigned to all HRTF directions. A set of 20 sentences was randomly selected for each azimuth location. Finally, the anechoic signal was corrupted with diffuse noise, and ITD/ILD were computed. For more details on this implementation, please refer to [3].

5 RESULTS

Vertical localization metrics were defined by the simulations performed with Baumgartner's model. All median plane template angles for each HRTF set were considered in the computation of the following psychoacoustic performance metrics for vertical localization, according to [14]:

- *local polar RMS error, PE*: quantifies the average "local" localization error;
- *quadrant error rate, QE*: quantifies the localization confusion related to the rate of "non-local" responses, i.e. where the absolute polar error exceeds ±90°.
- *global polar error, GPE*: quantifies the absolute polar localization error, with front-back confusions "resolved";
- *front-back confusion rate, FB*: quantifies the localization confusion by measuring the rate of frontal responses where the target position is on the back region and *vice versa*, excluding elevation angles above the listener. A ±30° area is considered in this definition.

Horizontal localization metrics were defined by the simulations performed with May's model. The *global lateralization error* (GLE) was computed by averaging of the estimated errors among all available HRTF grid points in the horizontal plane and the twenty speech samples. This aggregation required the absolute distance between the real and the estimated DOA that was corrected in presence of the front-back confusion. Front-back confusion in the horizontal plane (FBH) was considered within a tolerance of ±10° around $\alpha_{real} \in \{-90°, 90°\}$, while the correction was performed by mirroring estimations across the inter-aural axis. This last metric should be considered in complement to *FB*.

The *"All-against-all"* matrices of each of the aforementioned metrics can be seen in Fig. 1 and Fig. 2. Predicted performances in the diagonal represent the simulated listening condition with individual HRTFs. It is worthwhile to notice that, the 21st and 45th columns simulate localization performances with generic HRTFs, in particular KEMAR with big and small pinnae, respectively. From a qualitative point of view, both models support the well-grounded idea of the existence of several non-individual HRTF sets that provide better localization performance on individual basis compared to KEMAR. Moreover, one can compute the "best available" non-individual HRTF set by considering for i-th row of the matrix, i.e. for the i-th subject, the j-th column, i.e. non-individual HRTF, providing the minimum error value (excluding the diagonal). Accordingly, the angular error between individual and best available HRTFs further supports the idea that there exists a non-individual HRTF set allowing localization close to individual HRTFs [14].

6 DISCUSSION AND CONCLUSIONS

It is worthwhile to notice that these results are strongly dependent on the choice of a specific auditory model which serves as a ground truth in our research framework. From a methodological point of view, nothing prevents to replicate our study employing different auditory models which might be able to better characterize human spatial hearing [4]. A tight connection between real listening evaluation and auditory model calibration is the key element for obtaining reliable and meaningful results.

Other perceptual attributes such as tonal quality discrepancy, externalization, immersion, to name but a few [22] will be investigated in order to provide a complete set of metrics able to describe the listener in terms of acoustic (i.e. HRTFs) and non-acoustic (i.e. perceptual and cognitive) factors. The *"All-against-all"* comparisons based on all available metrics and attributes lends themselves well to machine learning algorithms, especially increasing the availability of (non-)acoustic data. In principle, the proposed methodology will allow to quantitatively compare performances of any HRTF selection approach. More in general, a fast prototyping of HRTF models and selection procedures might be easily tested against the optimal solutions resulted from that specific set of metrics.

The proposed approach paves the way for a complete listener characterization beyond the acoustic phenomena, with a special emphasis on the listener-to-contexts/scenarios/task relationship which is inherently a multi-modal and multi-domain process (i.e., acoustics, psycho-physics. cognition, and emotion). The VR/AR technologies of the future will be able to take advantage of such a detailed description, classifying task, contexts and listeners while adjusting rendering complexity in real-time [10, 13, 12].

ACKNOWLEDGEMENTS

This study was supported by the internationalization grant of the 2016-2021 strategic program “Knowledge for the World" awarded by Aalborg University to Michele Geronazzo.

REFERENCES

[1] 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.*, pages 1–4, New Paltz, New York, USA, Oct. 2001.

[2] F. Asano, Y. Suzuki, and T. Sone. Role of spectral cues in median plane localization. *The Journal of the Acoustical Society of America*, 88(1):159–168, 1990.

[3] R. Barumerli, A. Almenari, M. Geronazzo, G. M. Di Nunzio, and F. Avanzini. Auditory models comparison for horizontal localization of concurrent speakers in adverse acoustic scenarios. In *In Proc. 23rd International Congress on Acoustics*, pages 1–8, Aachen, DE, Sept. 2019.

[4] R. Barumerli, M. Geronazzo, and F. Avanzini. Localization in Elevation with Non-Individual Head-related Transfer Functions: Comparing Predictions of Two Auditory Models. In *Proc. 26th European Sig. Proc. Conf. (EUSIPCO 2018) - accepted for publication*, Rome, Italy, Sept. 2018. IEEE Signal Processing Society.

[5] R. Barumerli, M. Geronazzo, and F. Avanzini. Round Robin Comparison of Inter-Laboratory HRTF Measurements – Assessment with an Auditory Model for Elevation. In *Proc. of IEEE 4th VR Workshop on Sonic Interactions for Virtual Environments (SIVE18)*, Reutlingen, Germany, Mar. 2018. IEEE Computer Society.

[6] R. Baumgartner, P. Majdak, and B. Laback. Modeling sound-source localization in sagittal planes for human listeners. *The Journal of the Acoustical Society of America*, 136(2):791–802, 2014.

[7] W. G. Gardner and K. D. Martin. HRTF Measurements of a KEMAR. *J. of the Acoustical Society of America*, 97(6):3907–3908, June 1995.

[8] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

[9] M. Geronazzo, A. Carraro, and F. Avanzini. Evaluating vertical localization performance of 3d sound rendering models with a perceptual metric. In *2015 IEEE 2nd VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*, pages 1–5, Arles, France, Mar. 2015. IEEE Computer Society.

[10] M. Geronazzo, J. Fantin, G. Sorato, G. Baldovino, and F. Avanzini. Acoustic Selfies for Extraction of External Ear Features in Mobile Audio Augmented Reality. In *Proc. 22nd ACM Symposium on Virtual Reality Software and Technology (VRST 2016)*, pages 23–26, Munich, Germany, Nov. 2016. ACM.

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

[12] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini. Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization. *J. Audio Eng. Soc.,*, 67(6):1–15, June 2019.

[13] M. Geronazzo, E. Sikström, J. Kleimola, F. Avanzini, A. De Götzen, and S. Serafin. The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios. In *Proc. 17th IEEE/ACM Int. Symposium on Mixed and Augmented Reality (ISMAR)*, pages 90–97, Munich, Germany, Oct. 2018. IEEE Computer Society.

[14] 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 Transactions on Audio, Speech, and Language Processing*, 26(7):1243–1256, July 2018.

[15] L. A. Jeffress. A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41(1):35–39, 1948.

[16] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl. A Spatial Audio Quality Inventory (SAQI). *Acta Acustica united with Acustica*, 100(5):984–994, Sept. 2014.

[17] N. Ma, T. May, and G. J. Brown. Exploiting Deep Neural Networks and Head Movements for Robust Binaural Localization of Multiple Sources in Reverberant Environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2444–2453, Dec. 2017.

[18] P. Majdak, R. Baumgartner, and B. Laback. Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. *Front Psychol*, 5:1–10, Apr. 2014.

[19] T. May, S. v. d. Par, and A. Kohlrausch. A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):1–13, Jan. 2011.

[20] C. Mendonça, G. Campos, P. Dias, and J. A. Santos. Learning Auditory Space: Generalization and Long-Term Effects. *PLoS ONE*, 8(10):e77900, Oct. 2013.

[21] A. J. V. Opstal, J. Vliegen, and T. V. Esch. Reconstructing spectral cues for sound localization from responses to rippled noise stimuli. *PLOS ONE*, 12(3):e0174185, Mar. 2017.

[22] L. S. R. Simon, N. Zacharov, and B. F. G. Katz. Perceptual attributes for the comparison of head-related transfer functions. *The Journal of the Acoustical Society of America*, 140(5):3623–3632, Nov. 2016.