3D Virtual Audio with Headphones: A Literature Review of the Last Ten Years

Patrick Nowak [1,2], Véronique Zimpfer [1], and Udo Zölzer [2]

[1] *French-German Research Institute Saint-Louis, 68300 Saint-Louis, France*

[2] *Helmut Schmidt University, 22043 Hamburg, Germany, Email: patrick.nowak@hsu-hh.de*

Abstract

With the increasing number of applications for virtual reality also the research activity on 3D audio through headphones has risen. Thus, different approaches for improving the perception of the virtual experience have been developed. This paper summarizes and evaluates the work done on this topic during the last ten years. The investigations mainly address the individualization of the binaural technologies for improving the virtual source localization and externalization. These strategies can be basically divided into personalized headphone equalization and individualized head-related transfer functions (HRTFs). The former is responsible for getting rid of the colouration introduced by the headphone used during the playback, and the latter for introducing personal anthropometric characteristics into the utilized HRTFs.

Introduction

Human natural listening is based on the localization of sound sources using two ears [1]. The primary cues for horizontal localization are interaural cues, which represent differences between the left and the right ear. These differences appear as disparities in the time of arrival between the two ears (Interaural Time Difference, ITD) as well as level differences due to the head shadowing effect (Interaural Level Difference, ILD). Although these two binaural cues enable the horizontal localization, they do not provide information about the source elevation due to areas around the interaural axis where the interaural differences have identical values [2]. These so-called "cones of confusion" are the basis for front-back ambiguities and elevation errors. Therefore, monaural spectral filtering adds additional information to localize vertical sources [2]. For instance, high frequencies are stronger attenuated for rear sources than for frontal sources, due to the orientation and the shell-like structure of the pinna. Additionally, peaks and notches inside the spectrum are important features for perceiving elevated sources and solving front-back confusions [3]. All these aspects are summarized inside the HRTFs, which are highly individual due to differences in size and shape of the bodies between individuals. Due to these differences between individuals, when using non-individualized HRTFs during binaural synthesis through headphones, spatial and timbral distortions can occur [4,5]. These distortions lead to a higher rate of localization errors and front-back confusions of synthesised sources in comparison to real sources. Dynamic cues, which can be included through head tracking, can help to reduce these undesired effects [6]. These small movements lead to monaural and interaural cue changes that help the human brain to localize a sound source [7].

In 2007, Xu et al. [8] have summarized seven methods to individualize HRTFs. The methods, which are visited are the individualization by direct HRTF measurements, averaging or using typical HRTFs, subjective selection, scaling or grouping of non-individual HRTFs, theoretical computation, physical features, and tuning. In addition to this review, the present paper summarizes the newest research activities on 3D virtual audio with headphones in the last ten years, including approaches for personalized headphone equalization.

In the following, approaches for enhancing the 3D audio reproduction based on individualization methods are explained. These approaches can be basically grouped into two main topics. The first topic targets at the individualization of the used HRTF and the second at the personalization of the headphone equalization. Finally, an outlook of additional research topics in the field of 3D audio with headphones is given and the different methods are concluded.

HRTF Individualization

In order to measure individual HRTFs, a lot of requirements have to be fulfilled, e.g. the need of an anechoic measurement room and hundreds of different incident angles, resulting in a time-consuming and financially expensive measurement setup [8]. Thus, methods for HRTF individualization are in the main focus of 3D audio reconstruction via headphones and several different approaches have been developed. State of the art investigations include faster individual measurements [9,10], anthropometric matching or interpolation with databases [11–21], finite element simulations of the head [22] or the ear canal [23,24], and perceptually based selection [25].

Faster individual measurements target on shorten the measurement procedure of individual HRTFs in time, in order to offer a high spatial resolution in an acceptable period of time [9]. This acceleration is based on the usage of multiple exponential sweeps, which are signals consisting of several interleaved and overlapped sine sweeps [26]. Nevertheless, the measurement setup still consists of a rotating arc with numerous loudspeakers on it. In order to avoid this technical effort, an alternative approach has been taken by S. Li and J. Peissig [10], which uses an adaptive measurement procedure with arbitrary head movements, where the subject is sitting in front of a single speaker. During the measurement, a head tracker captures the current direction and a normalized least mean squares algorithm updates the corresponding HRTF.

Secondary, a lot of strategies to individualize HRTFs by matching the anthropometric data of a new person to subjects included in a database are still prevailing [11]. In these approaches, the anthropometric data of a new person is often extracted from a picture or a 3D scan of the head. Afterwards, the data is compared to all the subjects inside a database and distance metrics are calculated to find the best match in case of anthropometric data. Finally, the HRTF of the best match is taken as the individualized HRTF of the new person. In this research topic, the main focus lays on identifying the relevant anthropometric parameters, which can be used to find the best match in the view of a subjective localization perception [27, 28]. In addition to the usage of distance metrics of the anthropometry, also measured ILDs [13] or the notch frequencies calculated from a photo of the ear [12] can be used for HRTF matching. In the former, the ILD is measured for a small set of directions and afterwards compared to the ILDs inside the database. Furthermore, the latter approach is based on the assumption that the reflections of the three pinna contours can lead to destructive interferences at the ear canal entrance [12]. Therefore, the distances of the three contours to the ear canal entrance are first calculated for different elevation angles of the incoming sound and then transformed to the corresponding frequencies. Finally, these frequencies are compared to the appearing notch frequencies in the database and the HRTF set with the smallest deviation is chosen as the individual one. Also the estimation of the peaks and notches in an HRTF from the pinna anthropometry [29] and the influence of them on the localization in the median plane [30] are still of main interest.

As an extension to this matching, Bilinski et al. [14] have proposed an approach for interpolating the individualized HRTFs for a new person from a small group of subjects in a database. In this approach, weighting factors are trained for the different subjects to recreate the anthropometric data of a new person. Then, the same factors are used to weight the corresponding HRTF magnitudes while interpolating them. In [15], this approach is extended for the phases of the HRTFs. Moreover, in [16] the different anthropometric parameters are considered as unequally relevant and are therefore weighted differently during the interpolation process. Additionally, coordinate transforms like the principal component analysis [17, 18] or spherical harmonics [19] are used to reduce the dimensions of the data. In fact, all these approaches are based on a training and testing phase, thus, neural networks can be used to perform the matching process. Z. Haraszy et al. [20] and C. J. Chun et al. [21] have presented matching processes using artificial neural networks or deep neural networks, respectively. Both networks make use of backpropagation in order to learn the allocation during the matching process.

In addition to direct measurements or individualization of HRTFs, also simulations to estimate the influence of the head [22] or the ear canal [23, 24] have been proposed. These simulations are based on boundary element methods, where models are created from the anthropometric data and afterwards, the acoustic properties can be simulated. These simulations rely on the principle of reciprocity, where a point source is located inside the ear and observation points surround the head. The output contains all the pressure values on the surface of the model and at the observation points. An important issue in the fast acquisition of personalized HRTFs is the complexity of generating good quality head models for the simulation. Therefore, T. Huttunen et al. [22] compared three different acquistion methods, starting from a system with 52 cameras and ending up in a single mobile phone camera. In [24], M. Hiipakka proposed a method, in which the sound pressure and velocity are measured at the entrance of the ear and then the pressure at the eardrum can be estimated through an ear canal model.

Individual Headphone Equalization

Although headphones easily separate the desired signals at the two ears, they introduce additional spectral colourations, which degrade the externalization and the determination of the elevation [31]. Nevertheless, if the sound pressure at the eardrum of a listener can be precisely duplicated during headphone playback, a 3D sound experience can be recreated [32]. Therefore, methods for the compensation of the headphone transfer function (HpTFs) between the loudspeaker and the eardrum are required. The measurement of the individual HpTFs can be done with individual off-line measurements [24, 33], perceptual adjustments [31] and on-line adaptive algorithms [34, 35].

The most direct way to achieve individual headphone equalization is to use probe microphones at the eardrum in order to measure the HpTF [36]. Because of the invasive nature of the measurement, which is dangerous and impractical for widespread use, the technique developed by M. Hiipakka [24], that was introduced above for individual HRTF measurements, can be used to measure the HpTF at the eardrum. The advantage over the traditional measurement method is, that the invasive and dangerous placement of the microphone close to the eardrum is avoided and the microphone can be placed at the entrance of the ear canal, while obtaining similar measurement results. Another problem of the direct measurement of HpTFs is the headphone repositioning. For addressing this matter, B. Masiero and J. Fels [33] proposed a method, where the upper limit of several measured HpTFs is calculated. This results in an equalization filter without strong peaks, which may have led to high amplifications when the corresponding notch does not occur.

In order to avoid direct measurements, D. Griesinger proposed an application for individual headphone equalization through equal loudness matching [31]. The procedure is split into two stages. In a first step, the personal equal loudness curve for a loudspeaker at a distance of one meter in front of the user has to be found by adjusting the loudness of the third octave frequency bands to a given reference signal. Afterwards, the same procedure is repeated for headphone playback with an additional adjustment of the level between the two ears to perceive the

sound centred. Finally, the headphone equalization for a frontal speaker is calculated in frequency-domain by subtracting the dB values of the headphone equal loudness curve from the one of the loudspeaker.

In contrast to this off-line measurement procedures, an on-line adaptive equalization approach is proposed by R. Ranjan and W.-S. Gan in [34], in order to adapt to headphone repositioning. The principle of this approach relies on the basics of adaptive feedforward active noise control techniques, which are based on a filtered-x least mean squares algorithm. The resulting FIR filter coefficients, that normally conduct the active noise control, will now perform the equalization of the HpTF. Additionally, a microphone-to-eardrum-reference-point response from an ear canal model can be used to yield the adaptation for a virtual microphone at the eardrum [35].

Additional Reserach Topics

Besides the individualization, also other topics are inside the focus of current research in 3D audio with headphones. These topics include the improvement of localization performance with non-individual HRTFs through training [37], the influence of head tracking on the localization and externalization [38,39], and the combination of virtual sources with the real environment [34].

Conclusion

In this literature review a broad spectrum of different approaches for individualization of binaural synthesis have been presented. These strategies are mainly split into the individualization of the used HRTFs and the personalization of the headphone equalization. In both topics a lot of different approaches have been developed or improved during the last ten years. Additionally, an outlook of research topics not based on individualization is given.

Literatur

[1] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 33–42, Jan 2011.

[2] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT Press, 1997.

[3] J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," *Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2607–2624, Nov 1992.

[4] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, Jul 1993.

[5] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *Journal of the Audio Engineering Society*, vol. 44, no. 6, pp. 451–496, Jun 1996.

[6] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2841–2853, Jun 1999.

[7] G. Theile, "Equalization of studio monitor headphones," in *Audio Engineering Society Conference: 2016 AES International Conference on Headphone Technology*, Aug 2016.

[8] S. Xu, Z. Li, and G. Salvendy, "Individualization of head-related transfer function for three-dimensional virtual auditory display: A review," in *Virtual Reality.* Springer Berlin Heidelberg, 2007, pp. 397–407.

[9] J.-G. Richter, G. Behler, and J. Fels, "Evaluation of a fast HRTF measurement system," in *Audio Engineering Society Convention 140*, May 2016.

[10] S. Li and J. Peissig, "Fast estimation of 2D individual HRTFs with arbitrary head movements," in *22nd International Conference on Digital Signal Processing (DSP)*, Aug 2017, pp. 1–5.

[11] X.-Y. Zeng, S.-G. Wang, and L.-P. Gao, "A hybrid algorithm for selecting head-related transfer function based on similarity of anthropometric structures," *Journal of Sound and Vibration*, vol. 329, no. 19, pp. 4093 – 4106, 2010.

[12] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini, "Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4463–4467.

[13] M. Parviainen and P. Pertilä, "Obtaining an optimal set of head-related transfer functions with a small amount of measurements," in *2017 IEEE International Workshop on Signal Processing Systems (SiPS)*, Oct 2017, pp. 1–5.

[14] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4468–4472.

[15] I. Tashev, "HRTF phase synthesis via sparse representation of anthropometric features," in *2014 Information Theory and Applications Workshop (ITA)*, Feb 2014, pp. 1–5.

[16] M. Zhu, M. Shahnawaz, S. Tubaro, and A. Sarti, "HRTF personalization based on weighted sparse representation of anthropometric features," in *2017 International Conference on 3D Immersion (IC3D)*, Dec 2017, pp. 1–7.

[17] R. Bomhardt, H. Braren, and J. Fels, "Individualization of head-related transfer functions using principal component analysis and anthropometric dimen-

sions," *Proceedings of Meetings on Acoustics*, vol. 29, no. 1, 2016.

[18] C. S. Reddy and R. M. Hegde, "A joint sparsity and linear regression based method for customization of median plane HRIR," in *2015 49th Asilomar Conference on Signals, Systems and Computers*, Nov 2015, pp. 785–789.

[19] R. Sridhar and E. Choueiri, "A method for efficiently calculating head-related transfer functions directly from head scan point clouds," in *Audio Engineering Society Convention 143*, Oct 2017.

[20] Z. Haraszy, D.-G. Cristea, V. Tiponut, and T. Slavici, "Improved head related transfer function generation and testing for acoustic virtual reality development," in *World Scientific and Engineering Academy and Society Circuits, Systems, Communications and Computers (WSEAS CSCC) Multiconference - WSEAS International Conference on Systems (ICS)*, July 2010.

[21] C. J. Chun, J. M. Moon, G. W. Lee, N. K. Kim, and H. K. Kim, "Deep neural network based HRTF personalization using anthropometric measurements," in *Audio Engineering Society Convention 143*, Oct 2017.

[22] T. Huttunen, A. Vanne, S. Harder, R. R. Paulsen, S. King, L. Perry-Smith, and L. Kärkkäinen, "Rapid generation of personalized HRTFs," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, Aug 2014.

[23] S. Schmidt and H. Hudde, "Accuracy of acoustic ear canal impedances: Finite element simulation of measurement methods using a coupling tube," *Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 3819–3827, Jun 2009.

[24] M. Hiipakka, "Estimating pressure at the eardrum for binaural reproduction," PhD dissertation, Aalto University, 2012.

[25] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. EL99–EL105, Feb 2012.

[26] P. Majdak, P. Balasz, and B. Laback, "Multiple exponential sine sweep method for fast measurement of head-related transfer functions," *Journal of the Audio Engineering Society*, vol. 55, no. 7/8, pp. 623–637, Jul/Aug 2007.

[27] J. Fels and M. Vorländer, "Anthropometric parameters influencing head-related transfer functions," *Acta Acustica United with Acustica*, vol. 95, pp. 331–342, 2009.

[28] S. Ghorbal, T. Auclair, C. Soladieé, and R. Séguier, "Pinna morphological parameters influencing HRTF sets," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*, Sept 2017.

[29] 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," *The Journal of the Acoustical Society of America*, vol. 137, no. 2, pp. 690–701, 2015.

[30] K. Iida and Y. Ishii, "Effects of adding a spectral peak generated by the second pinna resonance to a parametric model of head-related transfer functions on upper median plane sound localization," *Applied Acoustics*, vol. 129, pp. 239 – 247, 2018.

[31] D. Griesinger, "Playback of non-individual binaural recordings without head tracking, and its potential for archiving and analyzing concert hall acoustics," in *Proceedings of the 22nd International Congress on Acoustics (ICA)*, Sept 2016.

[32] M. R. Schroeder, D. Gottlob, and K. F. Siebrasse, "Comparative study of european concert halls: Correlation of subjective preference with geometric and acoustic parameters," *Journal of the Acoustical Society of America*, vol. 56, no. 4, pp. 1195–1201, 1974.

[33] B. Masiero and J. Fels, "Perceptually robust headphone equalization for binaural reproduction," in *Audio Engineering Society Convention 130*, May 2011.

[34] R. Ranjan and W. S. Gan, "Natural listening over headphones in augmented reality using adaptive filtering techniques," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1988–2002, Nov 2015.

[35] J. Liski, V. Välimäki, S. Vesa, and R. Väänänen, "Real-time adaptive equalization for headphone listening," in *25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 638–642.

[36] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3, pp. 171 – 218, 1992.

[37] C. Mendonça, G. Campos, P. Dias, J. Vieira, J. P. Ferreira, and J. A. Santos, "On the improvement of localization accuracy with non-individualized HRTF-based sounds," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 821–830, 2012.

[38] G. D. Romigh, D. S. Brungart, and B. D. Simpson, "Free-field localization performance with a head-tracked virtual auditory display," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 943–954, Aug 2015.

[39] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. Katz, and C. de Boishéraud, "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2011–2023, 2017.