Obtaining an Optimal Set of Head-Related Transfer Functions with a Small Amount of Measurements

Mikko Parviainen
Laboratory of Signal Processing
Tampere University of Technology
Tampere, Finland
Email: mikko.p.parviainen@tut.fi

Pasi Pertilä
Laboratory of Signal Processing
Tampere University of Technology
Tampere, Finland
Email: pasi.pertila@tut.fi

***Abstract*—This article presents a method to obtain personalized Head-Related Transfer Functions (HRTFs) for creating virtual soundscapes based on small amount of measurements. The best matching set of HRTFs are selected among the entries from publicly available databases. The proposed method is evaluated using a listening test where subjects assess the audio samples created using the best matching set of HRTFs against a randomly chosen set of HRTFs from the same location. The listening test indicates that subjects prefer the proposed method over random set of HRTFs.**

***Index Terms*—acoustic signal processing, acoustic measurements, headphones**

I. Introduction

The reproduction of virtual acoustical space (VAS) is of interest in many applications. Recently virtual and augmented reality hardware has been introduced at prices that make them available for a vast group of users. To make the experience of such equipment realistic, individualized VAS reproduction is essential.

The perception of a sound event, i.e., its spatial properties such as direction is encoded into Head Related Transfer Functions (HRTFs). Therefore VAS can be reproduced using HRTFs. However, HRTFs for a given spatial location depend on anthropometric properties of each person. To create individualized VASs, HRTFs need to be obtained for each person from each spatial location with respect to the subject. The studies summarized in [1] show that humans are able to detect as small as the order of one degree changes in the sound source direction and therefore relatively high spatial resolution is required to obtain realistic soundscapes for each person. Using measurements to obtain such HRTFs is time-consuming and impractical for many consumer applications. In [2] a method is presented, which is relatively fast compared to traditional HRTF acquisition where subject's head is tracked and she/he is asked to rotate the head to different positions. While [2] may result in personalized HRTFs, acquisition of all positions of 3D space may be difficult for some subjects. Therefore, using the non-individualized HRTF may be better option in many cases, since the estimation of individualized HRTFs from non-individualized ones can produce quite realistic VAS (see e.g., in [3]). Over the years many efforts have been made to create personalized VASs without having to perform full scale HRTF and/or anthropometric measurements. Many proposed techniques utilize databases consisting of HRTFs of tens of subjects such as [4] and [5] or HRTFs estimated using a mannequin or dummy head [6]. The HRTF databases are created using dedicated setups in an anechoic chamber that allow movement of the loudspeaker to different positions around the subject who is wearing small microphones in the ears.

Approaches to obtain individualized HRTFs include subjective selection methods [7][3], scaling of amplitude response of HRTF at specific frequencies [8], clustering based method [9], a structural model technique [4], matching of anthropometric measurements [10], and a pinna response based tuning method [11]. A method proposed in [12] uses an artificial neural network and few measured HRTFs to learn estimate the personalized HRTFs. [13] uses partial depth images to extract anthropometric measurements in order to personalize Interaural Time Difference (ITD) information of generic HRTFs.

This paper proposes a database matching method that uses measurements conducted only in two sound source locations and based on them searches for the optimal database entry for the test subject. This enables a relatively fast way to create of VAS compared to full-scale HRTF measurements.

The rest of the paper is organized as follows. Section II describes the proposed method. Section III presents the listening experiment that is used to evaluate the method. Section IV presents the results and evaluates their statistical significance. Sections V and VI conclude the discussion.

II. Obtaining Optimized set of HRTFs

This section presents the method to obtain the best fitting set of HRTFs for a subject using the proposed method. The method is based on determining Interaural Level Difference (ILD) from a subject and comparing the estimated ILD to

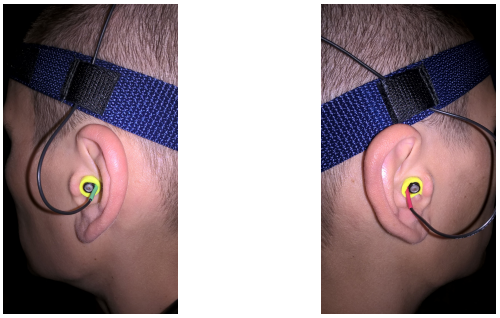

Fig. 1: Binaural microphone set worn by a test subject.

each subject in the HRTF database. This work utilizes, but is not limited to, LISTEN database [5].

A. Head Related Impulse Response (HRIR) and ILD Estimation

The HRIR is the time domain presentation of HRTF and HRIRs are estimated as the impulse response between binaural microphones and the emitted stimulus from a known location. Here, maximum length sequence (MLS) sequence of order 16 is used as a stimulus and the sequence is repeated eight times.

In order to obtain personalized set of HRTFs, ILD is estimated for the test subject and each subject in the database. The procedure is as follows. HRIRs are transformed into the frequency domain using discrete-time Fourier (DFT) transform and then the logarithm of the squared magnitude response is calculated:

$$H_{ch}(f) = 10\log_{10}\left|\mathcal{F}\left\{h_{ch}(n)\right\}\right|^2, \tag{1}$$

where $\mathcal{F}\{\cdot\}$ denotes DFT, n is time, $f = [0, \frac{\text{NDFT}}{2} - 1]$, and ch denotes the channel ($ch = 1$ is the microphone in the left ear and $ch = 2$ is the microphone in the right ear). NDFT is the length of DFT. Finally, ILD is estimated from the squared magnitude responses as

$$ILD(f) = H_1(f) - H_2(f). \tag{2}$$

B. Matching

In this work matching is based on measuring ILD from the locations that are parallel to the axis connecting the ears. These two source location are in spherical coordinates (AZ $= 90°, \text{EL} = 0°, \text{R} = 1.0$) and (AZ $= 270°, \text{EL} = 0°, \text{R} = 1.0$), where AZ $\in [0°, 360°)$ denotes the horizontal angle andAZ $\in [-90°, 90°]$ the vertical angle. Hereafter, the former location is denoted as AZ90EL0 and the latter as AZ270EL0 (see Figure 3 for the coordinate system). Subjects' measured ILDs are compared to each entry in LISTEN database using the log-spectral distance metric [14]

$$D_{LS} = \sqrt{\frac{1}{f_u - f_l + 1}\sum_{f=f_l}^{f_u} 10\log\left(\frac{ILD_s(f)}{ILD_{db}(f)}\right)^2}. \tag{3}$$

$ILD_s(f)$ and $ILD_{db}(f)$ are the ILDs of the subject s and the subject in the database db, respectively. f_l and f_u are lower and upper frequency bin indices, respectively. In this work the frequency range for the matching is 100 – 10000 Hz. This is done to avoid the influence of possible HRTF estimation errors in low and high frequency range.

Finally, the actual matching is performed. In brief, the database subjects are sorted according to the distance criterion D_{LS} and the optimal entry is chosen among the entries that rank the highest for all directions. The matching procudure in detail is presented in Figure 2.

In the description of the matching method we use the following notation.

db: indexing variable for database entries
db^{best}: the index of the best match
DB: the number of entries in the database
$\mathbf{s}$: source location
$\mathbf{D}^{\mathbf{s}}_{LS}$: a dictionary containing database index db and its calculated D_{LS} value using (3) for the source location $\mathbf{s}$
$cd^{\mathbf{s}}$: A set of candidates for the best match from source location $\mathbf{s}$
d: is the search depth (scalar)
$\mathbf{o}^{\mathbf{s}}$: a list of database indices db ordered in acsending order for source location $\mathbf{s}$

for db in $1, \ldots, DB$ **do**
  for $\mathbf{s}$ in {AZ90EL0, AZ270EL0} **do**
    calculate $D^{\mathbf{s}}_{LS}(db)$ using (3)
    $\mathbf{D}^{\mathbf{s}}_{LS} \leftarrow (db, D^{\mathbf{s}}_{LS}(db))$
  end for
end for
for $\mathbf{s}$ in {AZ90EL0, AZ270EL0} **do**
  $\mathbf{o}^{\mathbf{s}} = \text{argsort}(\mathbf{D}^{\mathbf{s}}_{LS})$ # sort dictionary and extract the keys, i.e. $db : s$, of the ordered dictionary
  $cd^{\mathbf{s}} = \varnothing$
end for
$d = 1$
while $cd^{\text{AZ90EL0}} \not\subset cd^{\text{AZ270EL0}}$ **or** $cd^{\text{AZ270EL0}} \not\subset cd^{\text{AZ90EL0}}$ **do**
  for $\mathbf{s}$ in {AZ90EL0, AZ270EL0} **do**
    $cd^{\mathbf{s}} =: cd^{\mathbf{s}} \cup \mathbf{o}^{\mathbf{s}}_{1:d}$
  end for
  $d = d + 1$
end while
Set $db^{best} = \{db :\in cd^{\text{AZ90EL0}}$ **and** $\in cd^{\text{AZ270EL0}}\}$

Fig. 2: The proposed matching algorithm.

C. Personalization of ITD Information

Finding the optimal set of HRIRs is made in the proposed system based on matching ILD between the test subject and the database subjects. Even though the spectral contour has great similarity, the ITD can still be a mismatch, which affects the perceived direction of the sound. Therefore, it is necessary to correct ITD information of HRIRs fetched from the database. ITD is not changing with frequency (in contrast to ILD) and therefore it is relatively easy to personalize ITD of a non-personalized HRTF. The set of HRIRs with personalized ITD

TABLE I: Source locations. AZ and EL denote the azimuth and elevation angle in degrees, respectively. R is the distance from the subject's head center to the sources in meters.

Coordinates			
	AZ	EL	R
Source 1	270.0	0.0	1.0
Source 2	90.0	0.0	1.0
Source 3	330.0	30.0	1.0
Source 4	345.0	-30.0	1.0
Source 5	240.0	30.0	1.0
Source 6	225.0	-30.0	1.0

is obtained as follows. ITD of left/right HRIR pair of a given sound source location is obtained using cross-correlation:

$$\begin{aligned} C_{LR}(m) &= \begin{cases} \sum_{n=0}^{N-m-1} h_L(n+m)h_R(n), & m \geq 0 \\ C_{RL}(-m), & m < 0 \end{cases} \\ m_{ITD} &= \underset{m}{\text{argmax}}\, C_{LR}(m), \end{aligned} \tag{4}$$

where n is the time index, N is the window length, m is the lag index, $C_{LR}(m)$ is the cross-correlation function, and h_L and h_R are the real-valued measured HRIRs from the given location to left and right ear, respectively. m_{ITD} is the correlation lag index where the cross-correlation function has its maximum value and therefore is the estimate of ITD in samples for the given location. For unknown source locations, personalized ITD can be obtained by estimating the head radius using, e.g., the extended Woodworth/Schlosberg formula [15] and the measured ITD of a known source location.

The results presented e.g. in [16] [17] show that the HRTF phase spectrum can be approximated by the minimum phase representation of the transfer function augmented with ITD information. The HRTF is decomposed as follows.

$$H(j\omega) = H_{\text{min}}(j\omega)H_{\text{ap}}(j\omega), \tag{5}$$

where $H_{\text{min}}(j\omega)$ is the minimum phase component and $H_{\text{ap}}(j\omega)$ is the all-pass component. The left and right ear all-pass components encode the sound source position as ITD and $H_{\text{ap}}(j\omega)$ is of form $e^{j\omega\tau_{ap}}$. The left and right ear minimum phase components, in turn, encode the ILD and monaural source localization cues.

The HRTFs with corrected ITD are obtained as follows. First, the optimal set of HRTFs according to (3) is decomposed using (5). Then, the corrected HRTFs are obtained by delaying the phase of the lagging channel (left or right) with the estimated ITD:

$$\hat{H}^{L|R}(j\omega) = H_{min}^{L|R} \exp(-j\omega m_{ITD}/N), \tag{6}$$

where $H_{min}^{L|R}$ is the minimum phase HRTF of the left or the right channel, ω is the angular frequency, and N is the length of HRIR in samples.

III. Listening Experiment

The listening test consisted of multiple experiments and in each experiment subjects were asked to select from two audio samples the one that better matched the evaluation criteria. The two audio samples were created as follows. Sample 1: a stimulus was convolved with a pair of best matching HRIR for the subject using (3) and (6). Sample 2: the same stimulus was convolved with a pair of HRIRs of a random database entry from the same source location. The reference sample was also available for playback during the test and it was obtained by convolving the stimulus with each subject's measured HRIRs for the source location under test.

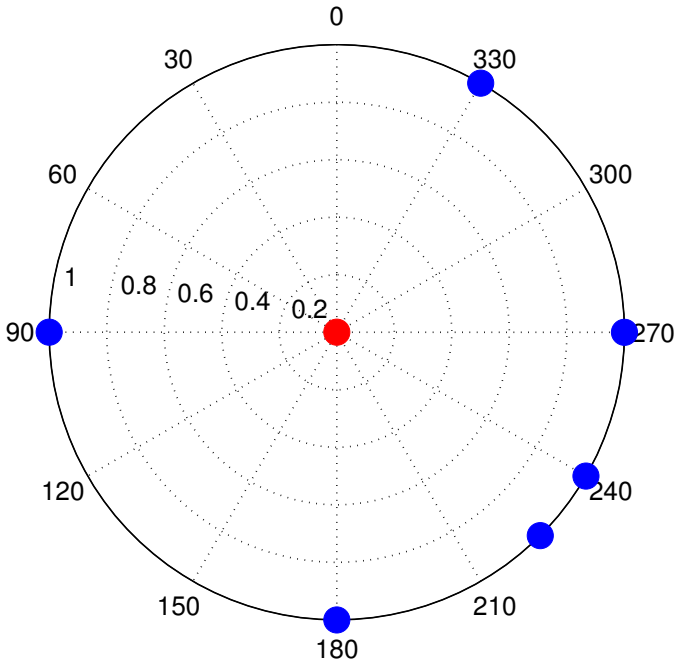

Fig. 3: The source locations (blue dots) and subject's head position (red dot) in the horizontal plane. The subject's orientation is AZ $= 0°$, EL $= 0°$.

The subjects evaluate the two samples by comparing them against the reference. They were instructed to choose the sample that is closer to the reference based on its spatial properties. The ground truth source location was shown to subjects by its spherical coordinates, see Figure 3. Subjects were personally guided before the test in terms of evaluation criteria, the coordinate system, and the software user interface. The subjects were instructed to listen to the samples several times before their final decision. Furthermore, the subjects were encouraged to playback each sample many times. Each subject was summoned to the listening test in an audito laboratory to guarantee an undisturbed test session. All subjects used Bose Quiet Comfort QC35 noise cancelling headphones with noise cancelling turned on.

Speech, pink noise bursts, and music were used as stimuli. Speech was an English male speaker. The pink noise bursts were 500 ms in duration followed by 200 ms pause. The onset of each burst was windowed by the exponential function that gradually increases reaching its maximum from 250 ms from the beginning. Similarly the end of the burst was windowed

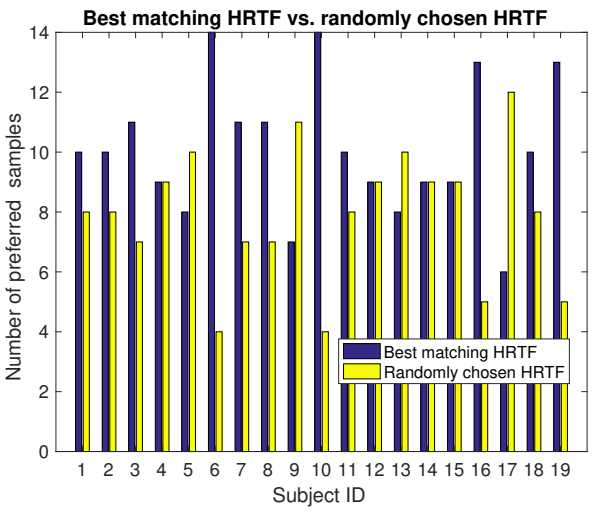

Fig. 4: HRTF preference results for each test subject. The bar height indicates how many of the presented sound samples were preferred by each subject that participated in the listening test.

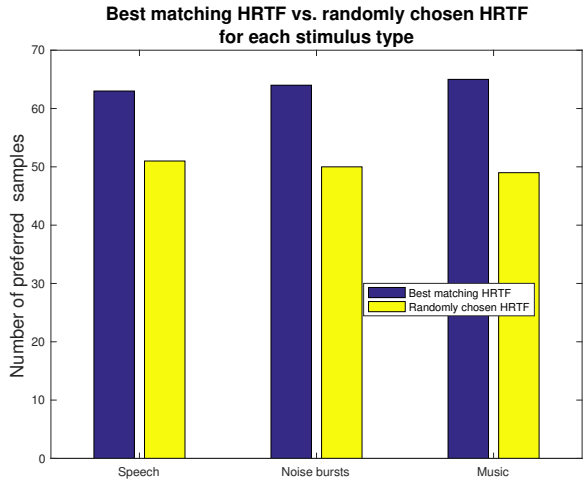

Fig. 5: Stimulus-wise preference.

by the time flipped exponential function that decreases from 250 ms to 500 ms. The music sample was an excerpt of a pop music song containing bass sounds and percussion instruments. Each sample was nine seconds in duration. Each content type was rendered to appear from the six locations indicated in Table I. In total, 18 samples were analyzed by each subject.

IV. Results

A. Raw Preference Test Results

Figure 5 presents the preference results for each stimulus type over all locations and Figure 6 the preference results for each source location over all stimulus types. Figure 4 contains the results over all the stimulus types and the source locations.

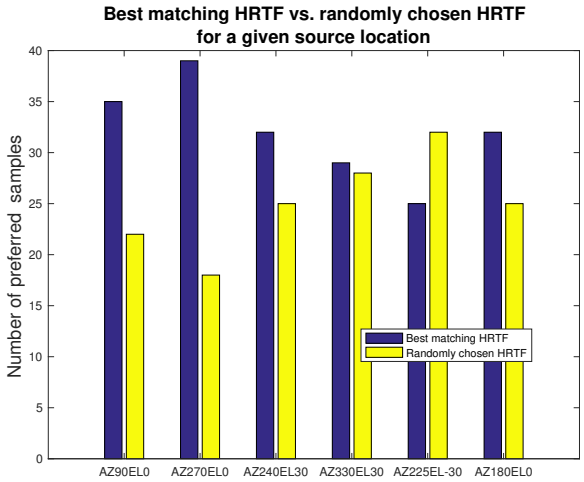

Fig. 6: Location-wise preference.

The bar heights indicate the number of preferred samples rendered using the personalized HRTFs and the randomly chosen HRTFs.

B. Statistical Significance

This section evaluates the HRTF preference results using statistical tests. To investigate whether subjects favor samples based on the proposed interaural level difference (ILD) matching or not, one way analysis of variance (ANOVA) was used to analyze the differences between the population groups. Here, there are two groups:

A: prefer the prosed method of HRTF selection and compensation for corrected source direction

B: prefer random HRTF entries with correct direction.

Performing ANOVA on Figure 4 results in $F(1, 36) = 9.32, p = 0.0043$. Figure 7 illustrates the analysis. The ANOVA results in a relatively low p-value which suggest that the proposed method is able to retrieve the best fitting HRTF for the test subjects.

V. Discussion

Even though the statistical test results suggest that the subjects tend to prefer the HRTF set obtained using the proposed method, many subjects reported difficulties to find differences between the sets. That is, for many subjects the randomly chosen HRTF set resulted in a satisfactory perception of sound source location. This can result either from the randomly chosen HRTF being close to the optimal one and/or the subject was not that critical about the sound source location. This is somewhat surprising since often non-invidualized HRTFs can result in inaccurate lateralization, poor vertical effects, and even front-back confusion of the source location [1][18].

Another discovery made during the design of the listening experiment was that subjects learn to hear the nuances better between the samples with more practice. A group of three subjects were used to choose an appropriate variant for the

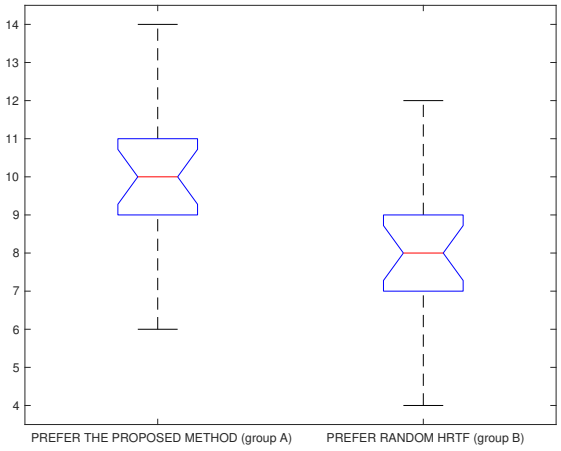

Fig. 7: Illustration of ANOVA of Figure 4.

listening test. This pilot group chose their best matching HRTF set more often than the subjects that did the test only once. This observation suggests that a localization preference task should consist of tens of samples. However, subjects may get tired if a listening session lasts long and therefore may be sloppy towards to the end. Therefore splitting a long listening session to few short session may result in more reliable evaluation. Another option is to repeat the test and remove the subjects with highly varying preference.

VI. Conclusions

This paper presented a method to obtain personalized set of HRTF. The method relies on measuring Interaural Level Difference (ILD) from the locations that are parallel to the axis connecting the ears. The personalized HRTFs are obtained from a catalogue of HRTF from which the optimal entry is selected using a similarity measure between ILDs of the target subject and the subjects in the database. The method was tested with a group of 19 subjects. The performance of the prososed method was measured using a listening test. The statistical testing of the preference results show that subject tend to prefer the best matching set obtained using the proposed method over a random set of HRTF from the same location.

Acknowledgment

This research has been supported by Nokia Technologies.

References

[1] J. Blauert, *Spatial Hearing: The psychophysics of human sound localization*. Cambridge, MA, USA: The MIT Press, 1999.

[2] R. Ranjan, J. He, and W.-S. Gan, "Fast Continuous Acquisition of HRTF for Human Subjects with Unconstrained Random Head Movements in Azimuth and Elevation," in *Audio Engineering Society Conference: 2016 AES International Conference on Headphone Technology*, August 2016. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18351

[3] A. Andreopoulou and A. Roginska, "Evaluating HRTF similarity through subjective assessments: Factors that can affect judgment," in *Proceedings of the 40th ICMC 11th SMC Conference, Athens, Greece*, 2014.

[4] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, 2001, pp. 99–102. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=969552

[5] O. Warusfel, "LISTEN HRTF database," Room Acoustics Team, IRCAM, 2003, retrieved on April 2, 2017. [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/

[6] B. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone," *MIT Media Lab Perceptual Computing - Technical Report #280*, May 1994.

[7] B. U. Seeber and H. Fastl, "Subjective Selection of Non-Individual Head-Related Transfer Functions," *Proceedings of the 2003 International Conference on Auditory Display, Boston*, July 2003.

[8] C.-J. Tan and W.-S. Gan, "User-defined spectral manipulation of hrtf for improved localisation in 3d sound systems," *Electronics Letters*, vol. 34, no. 25, pp. 2387–2389, December 1998.

[9] S. Shimada, N. Hayashi, and S. Hayashi, "A clustering method for sound localization transfer functions," *J. Audio Eng. Soc*, vol. 42, no. 7/8, pp. 577–584, 1994. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=6935

[10] D. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.* IEEE, 2003, pp. 157–160.

[11] 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 Transactions on Fundamentals of Electronics Communications and Computer Sciences*, vol. 91, pp. 345–356, 2008.

[12] V. Lemaire, F. Clerot, S. Busson, R. Nicol, and V. Choqueuse, "Individualized HRTFs from few measurements: a statistical learning approach," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 4, July 2005, pp. 2041–2046 vol. 4.

[13] H. Gamper, D. Johnston, and I. J. Tashev, "Interaural time delay personalisation using incomplete head scans," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2017. [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2017/03/ITD_personalisation_ICASSP2017.pdf

[14] M. Deza and E. Deza, *Encyclopedia of Distances*. Springer Berlin Heidelberg, 2016. [Online]. Available: https://books.google.fi/books?id=KQHdDAAAQBAJ

[15] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc*, vol. 47, no. 9, pp. 675–705, 1999. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=12095

[16] A. Kulkarni, S. Isabelle, and H. Colburn, "On the minimum-phase approximation of head-related transfer functions," in *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*. IEEE, 1995, pp. 84–87.

[17] J. Nam, M. A. Kolar, and J. S. Abel, "On the minimum-phase nature of head-related transfer functions," in *Audio Engineering Society Convention 125*, October 2008. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14698

[18] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *Acoustical Society of America Journal*, vol. 94, pp. 111–123, 1993.