Audio Engineering Society

Convention Paper

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA

Eigen-Images of Head-Related Transfer Functions

Christoph Hold, Fabian Seipel, Fabian Brinkmann, Athanasios Lykartsis, and Stefan Weinzierl

Technical University Berlin, Audio Communication Group, Einsteinufer 17c, 10587 Berlin, Germany

Correspondence should be addressed to Christoph Hold (Christoph.Hold@alumni.tu-berlin.de)

ABSTRACT

The individualization of head-related transfer functions (HRTFs) leads to perceptually enhanced virtual environments. Particularly the peak-notch structure in HRTF spectra depending on the listener's specific head and pinna anthropometry contains crucial auditive cues, e.g. for the perception of sound source elevation. Inspired by the *eigen-faces* approach, we have decomposed image representations of individual full spherical HRTF data sets into linear combinations of orthogonal *eigen-images* by principle component analysis (PCA). Those eigen-images reveal regions of inter-subject variability across sets of HRTFs depending on direction and frequency. Results show common features as well as spectral variation within the individual HRTFs. Moreover, we can statistically de-noise the measured HRTFs using dimensionality reduction.

Introduction

Head related transfer functions (HRTFs) constitute the fundamental component of all binaural rendering techniques for the generation of virtual sound scenes. While generic HRTFs can provide acceptable localization performance for sound sources in the horizontal plane, they often fail to evoke a correct perception of elevation, for which spectral peaks and notches originating from the fine structure of the human outer ear (pinna) have been identified as an important cue [1, 2, 3, 4]. Consequently, individualized HRTFs have been found to improve localization and coloration so that individual binaural simulations enhance virtual acoustic realities [5].

Since measuring actual HRTFs requires a considerable effort, a vast amount of recent research has focused on developing methods to provide individual HRTFs based on the listener's anthropometry or perceptual feedback [6, 7]. This includes techniques such as selecting a best matching HRTF set from a large data base, individualizing generic HRTFs by means of frequency shifting and scaling [8, 9, 10], or constructing an individualized HRTF from a weighted superposition of orthogonal basis components [11, 12]. Lately, simulation approaches based on anthropometric features of the head and pinna have gained more interest as these features can be extracted semi-automatically from scale pictures [9], active shape models [13] or 3D head meshes [14] in order to directly relate them to individual spectral and temporal HRTF features or to synthesize HRTFs with acoustic modelling techniques [15, 16, 17, 18].

The above approaches are based on the finding that individual HRTFs share certain common spectral peak and notch structures for specific directions, since they

rely on relatively similar anthropometric pinna and head features [19].

Therefore, an insight into inter-subject variability in HRTFs and it's underlying anthropometric causes seems promising in order to further develop methods for HRTF individualization. To this goal we have applied an inter-subject principle component analysis (PCA) inspired by the eigen-faces approach [20] to a HRTF data base of 40 human subjects. This methodology reveals spectral target areas of high variance as well as shared components in the HRTF magnitude spectrum between subjects which will be discussed by the example of median and horizontal plane HRTFs.

Methods

The analyzed HRTFs were measured using a discrete spherical grid of five degree resolution in a distance of $r = 1{,}47$ m [21]. They are stored in a continuous spatial representation generated by spherical harmonic expansion which can be used to interpolate and eventually render the magnitude spectra on a pre-defined grid by using the inverse spherical Fourier transform [22, p. 17, p. 71].

For further processing we encode HRTFs both in the median and horizontal plane as images with each pixel's luminance value representing the energy of a unique frequency bin and angle of incidence. We restrict the analysis to left ears since we assume fairly symmetrical anthropometric features. Figure 1 shows an exemplary magnitude spectrum of a subject's left ear HRTF in the horizontal and median plane. Coordinate system conventions are adapted from Blauert [1, p. 14], where the zero degree point (ϕ and θ) is in front of the listener.

Similar to common picture detection and recognition tasks, every image is flattened to a single vector $\boldsymbol{x}$ containing $d = N_f \cdot N_\Phi$ luminance measurements corresponding to the pixels of HRTFs with N_Φ angles and N_f frequency bins. If we assume that all individual HRTFs from n subjects are realizations of the same random process, we can form the $n \times d$ matrix $\boldsymbol{X}$ as input to a singular value decomposition.

The Karhunen-Loève transform is considered as an optimal orthogonal transform with respect to mean-square error [23]. Applied to a zero-mean empirical sample, it is identical to the principal component analysis [23] leading to a compact representation of the input $\boldsymbol{X}$,

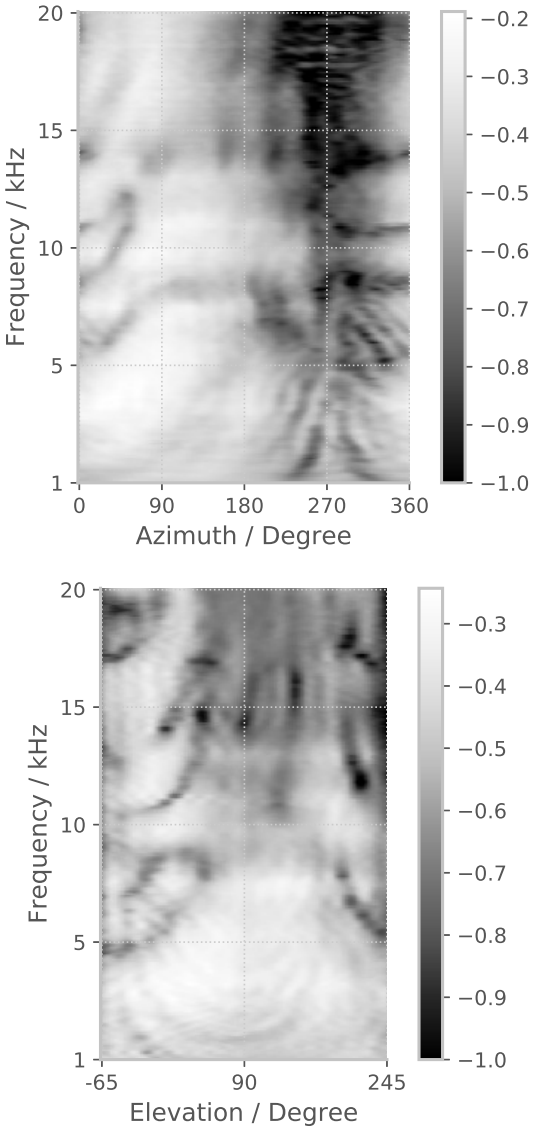

Fig. 1: Grey-scale image representations of left ear HRTFs form the input data. Here, horizontal plane (top) and median plane (bottom) of subject 1.

where all columns of the output $\boldsymbol{Y}$ are uncorrelated. The principal components transform is defined as

$$\boldsymbol{Y} = \boldsymbol{B}^t(\boldsymbol{X} - E(\boldsymbol{X})) \,, \tag{1}$$

with

$$\boldsymbol{C}_x = \boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^t \,, \tag{2}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues λ_1 to λ_n of the covariance matrix $\boldsymbol{C}_x$ [23]. The eigenvalues denote how much variance of the input each corresponding eigenvector explains. The columns of $\boldsymbol{B}$ contain the eigenvectors $\boldsymbol{b}_1$ to $\boldsymbol{b}_n$ which can be interpreted as basis vectors or *eigen-images* in our scenario. They are typically listed by their eigenvalue λ in descending order.

The reconstruction for subject m is then formed by a linear combination of the basis vectors $\boldsymbol{b_i}$ and their associated scalar weight entries $Y_{m,i}$ as

$$\tilde{\boldsymbol{X}}_m(\tilde{n}) = \sum_{i=1}^{\tilde{n}} Y_{m,i} \cdot \boldsymbol{b}_i \,, \tag{3}$$

where $\tilde{n}$ specifies the number of (orthogonal) *eigen-images* used for reconstruction. It is often referred to as the number of *dimensions*. Note that $\tilde{\boldsymbol{X}}$ is identical to $\boldsymbol{X}$ if $\tilde{n} = n$. Truncating the sum to $\tilde{n} < n$ principal components results in an incomplete reconstruction $\tilde{\boldsymbol{X}}$ consisting of a linear combination of only those $\tilde{n}$ major dimensions.
This can be written in matrix notation with truncated $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{B}}$ as

$$\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{B}} \,. \tag{4}$$

We further assume additive white noise that is uncorrelated with the signal. Since the orthogonal transform results in decorrelated basis vectors, we expect to be able to separate signal and noise, because the energy of the relevant signal should concentrate in the first components whereas noise will be evenly distributed across dimensions. If the eigenvalues do not decrease with further dimensions, the additional basis vectors primarily explain this noisefloor. Discarding these last components while reconstructing $\tilde{\boldsymbol{X}}$ can hence be interpreted as statistical de-noising, as commonly used in image processing [24].

Results

The input variance explained by each principal component is given by its eigenvalue. To get a first insight, figure 2 shows the determined eigenvalues in descending order. The eigenvalues drop more than two thirds within the first five dimensions. It is common practice to reduce the number of dimensions to the order where the first *knee* is apparent in the eigenvalues, or to choose the first n components which explain a specific quantity of variance. We hereby chose ten dimensions to be the inflection point – explaining 66% of the variance – assuming that smaller eigenvalues are presumably evoked by noise.

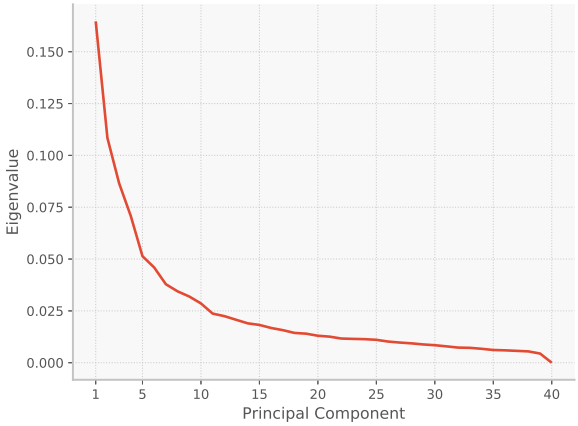

Fig. 2: Normalized eigenvalues of corresponding *eigen-images* in descending order, derived by inter-subject principal component analysis.

Derived eigenvectors can again be illustrated as images, also called *eigen-images*. Figure 3 shows the first three *eigen-images*, where the color corresponds to the absolute values. Colored regions highlight variance of the input data explained by each basis vector. The corresponding colorbars denote the value range, whereas the latter reduces with each additional orthogonal dimension, as expected. The *eigen-images* form the basis from which the HRTFs can be reconstructed by a linear combination (eq. 3).

Superposition of the above shown *eigen-images* yields a cumulative representation which results in highlighted areas where frequency and angle dependent variance can be found within the first ten dimensions in the horizontal (fig. 4) and median plane (fig. 6), respectively. In both cases, most inter-subject variance

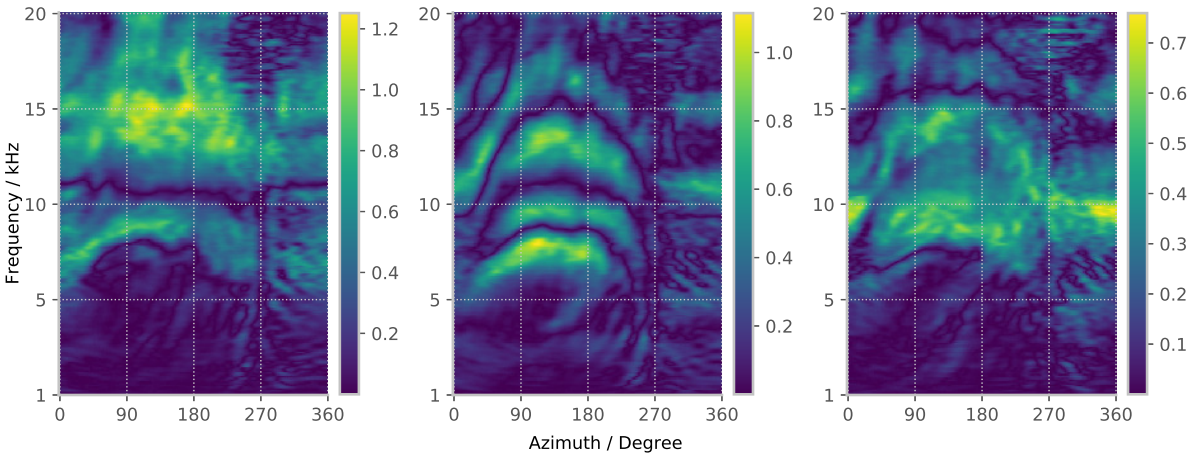

Fig. 3: First three *eigen-images* of the horizontal plane HRTFs (left to right). Colors indicate the amount of inter-subject variance per pixel according to right-hand colorbars.

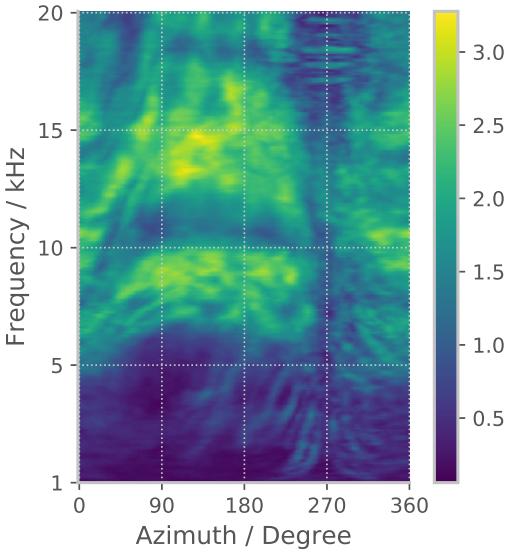

Fig. 4: Image representation of first ten superimposed *eigen-images* in the horizontal plane. Colors indicate inter-subject variance.

is visible above 5 kHz where pinna related cues dominate the HRTF. In the horizontal plane two areas stand out. A lower one between $6-10$ kHz and a slightly higher one located between $12-15$ kHz. Less variance is seen at the contralateral side around 270° azimuth. In this region, the HRTF is dominated by high frequency damping caused by the acoustic shadow of the head. More areas of high variance are found in the median plane. They roughly form an upside-down V-shapes that follow the change in HRTF-Notch frequencies across source elevation. Below 5 kHz, U-shaped arches emerge that originate from variance in the comb-filter induced by the shoulder reflection [25, 26].

Figure 5 illustrates the statistical de-noising effect. A reconstruction with the maximum order of $\tilde{n} = 40$ perfectly represents the unprocessed zero-mean data. Truncating the sum to $\tilde{n} < 40$ results in de-noising. With a truncation to $\tilde{n} = 20$, the fine structure is still preserved in great detail. One must bear in mind that a super-fine structure can also be evoked by measurement noise, so gentle smoothing methods could increase the generalization of data.

We assume that exaggerated de-noising is apparent in the most truncated version shown. The statistical smoothing has evened out the data. For the minimum of $\tilde{n} = 10$ the peaking areas become larger and also smoother, whereas some details already vanish. The compromise of $\tilde{n} = 20$ according to visual inspection preserves the fine structure and still incorporates the de-noising as well as reduction to half of the dimensionality.

Discussion

The decomposition into orthogonal *eigen-images* reveals which frequencies and which angles of

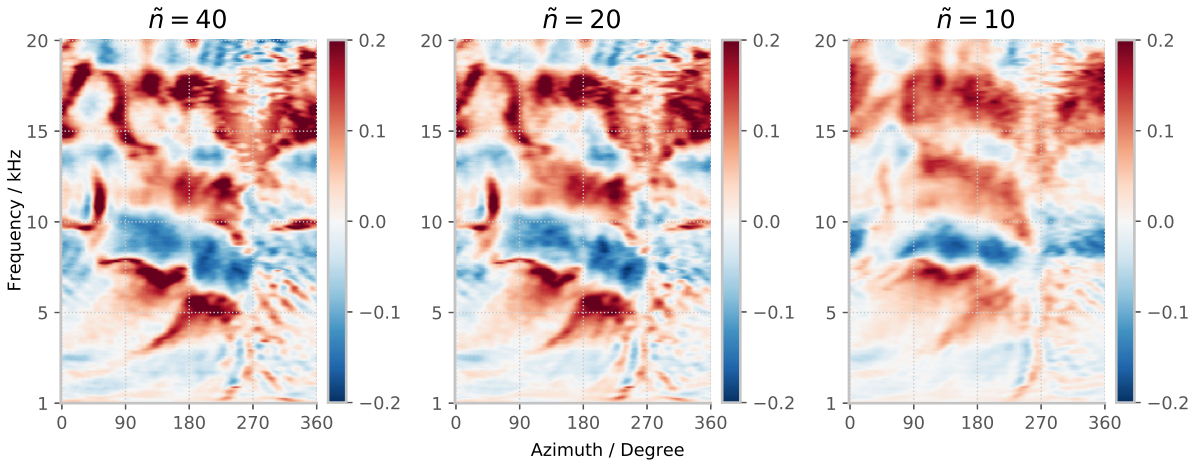

Fig. 5: De-noising process by reducing the number of principle components to $\tilde{n}$ = 40, 20 and 10 dimensions. Reconstruction of the (zero-mean) left ear horizontal plane HRTF of subject 18. Colors clip at $\pm$0.2.

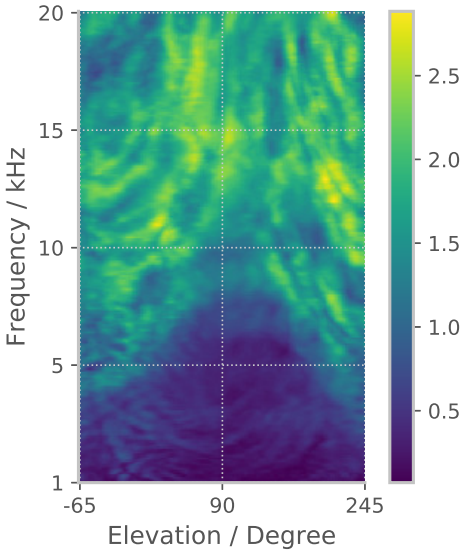

Fig. 6: Image representation of first ten superimposed *eigen-images* in the median plane. Colors indicate inter-subject variance.

incidence provoke differences in individual HRTFs among subjects. Hereby, not every region shows a considerable amount of variance, i.e. parts of the measured HRTFs are not significantly different between subjects, e.g. the frequency region below 5 kHz. This corresponds to the obvious fact, that anthropometrics hardly affect the level in the lower frequency region, because the wavelengths are large compared to the size of head and pinna. Therefore the variance is generally increasing towards smaller wavelengths/higher frequencies. Beyond 12 kHz even subtle differences in the outer ear anthropometry already drastically change the HRTF.

Regarding the dimensionality reduction, visual inspection suggests that the first ten dimensions might be sufficient for an appropriate representation of the examined database. This would result in a compression to only a quarter of the original data. Truncating the dimensionality constitutes a variance filtering operation which also statistically de-noises the HRTFs. This results in a loss of amplitude during reconstruction, which can be compensated by re-scaling in post-processing. The described procedure constitutes, however, an unsupervised statistical analysis. Thus, dismissed and filtered structures are not necessarily perceptually irrelevant and further listening tests are required to determine the effect on perception. Moreover, we have only analyzed single ear HRTF spectra, ignoring interaural time-differences in this study.

Summary

With the current contribution, we have presented a methodology to decompose individual HRTFs from

a database of individual subjects into a subset of orthogonal basis vectors, interpreted as *eigen-images* by applying principal component analysis.

Each individual HRTF set can thus be reconstructed as a linear combination of those *eigen-images*. Truncating the reconstruction by discarding dimensions with small eigenvalues results in statistical de-noising and data reduction. A superposition of *eigen-images* shows the amount of inter-subject variability in different regions and allows to analyze the actual individuality of HRTFs depending on frequency and direction.

Based upon the results above, the presented methodology can foster several applications for future research. Reducing the dimensionality of HRTF sets to only a few inter-subject principle components requires perceptual evaluation by listening tests and has the potential of storing individual HRTF databases in a compressed form of only a few components. A comparison of such principle components between different data sets can further gain insights to HRTF structures in terms of variance distribution. Finally, the PCA approach presented in this work marks a first step towards an individualization of HRTFs by machine learning techniques: Dimensionality reduction and de-noising procedures enables clustering and classification to provide better results in lower-dimensional data structures, especially when the amount of samples is limited. This is essential for explaining the individual structure of the HRTF based on the inter-subject variability in anthropometric features.

Code Repository

The code producing the results of this paper is available online.

`github.com/chris-hld/HRTF-EigenImages`

References

[1] Blauert, J., *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.

[2] Gardner, M. B. and Gardner, R. S., "Problem of localization in the median plane: effect of pinnae cavity occlusion," *The Journal of the Acoustical Society of America*, 53(2), pp. 400–408, 1973.

[3] Wright, D., Hebrank, J. H., and Wilson, B., "Pinna reflections as cues for localization," *The Journal of the Acoustical Society of America*, 56(3), pp. 957–962, 1974.

[4] Hebrank, J. and Wright, D., "Spectral cues used in the localization of sound sources on the median plane," *The Journal of the Acoustical Society of America*, 56(6), pp. 1829–1834, 1974.

[5] Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D., "Binaural technique: Do we need individual recordings?" *Journal of the Audio Engineering Society*, 44(6), pp. 451–469, 1996.

[6] Nicol, R., "Binaural technology," Audio Engineering Society, 2010.

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

[8] Schonstein, D. and Katz, B. F., "HRTF selection for binaural synthesis from a database using morphological parameters," in *International Congress on Acoustics (ICA)*, 2010.

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

[10] Middlebrooks, J. C., "Individual differences in external-ear transfer functions reduced by scaling in frequency," *The Journal of the Acoustical Society of America*, 106(3), pp. 1480–1492, 1999.

[11] Hölzl, J., "An initial Investigation into HRTF Adaptation using PCA," *IEM Project Thesis, Institut für elektronische musik und akustik. Graz, Austria*, 2012.

[12] Hoene, C., Patino Mejia, I. C., and Cacerovschi, A., "MySofa—Design Your Personal HRTF," in *Audio Engineering Society Convention 142*, Audio Engineering Society, 2017.

[13] Torres-Gallegos, E. A., Orduña-Bustamante, F., and Arámbula-Cosío, F., "Personalization of head-related transfer functions (HRTF) based on automatic photo-anthropometry and inference from a database," *Applied Acoustics*, 97, pp. 84–95, 2015.

[14] Dinakaran, M., Grosche, P., Brinkmann, F., and Weinzierl, S., "Extraction of Anthropometric Measures from 3D-Meshes for the Individualization of Head-Related Transfer Functions," in *Audio Engineering Society Convention 140*, Audio Engineering Society, 2016.

[15] Katz, B. F., "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation," *The Journal of the Acoustical Society of America*, 110(5), pp. 2440–2448, 2001.

[16] Ziegelwanger, H., Kreuzer, W., and Majdak, P., "Mesh2HRTF: open-source software package for the numerical calculation of head-related transfer functions," in *22st International Congress on Sound and Vibration*, 2015.

[17] Kreuzer, W., Majdak, P., and Chen, Z., "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range," *The Journal of the Acoustical Society of America*, 126(3), pp. 1280–1290, 2009.

[18] Kahana, Y., *Numerical modelling of the head-related transfer function*, Ph.D. thesis, University of Southampton, 2000.

[19] Takemoto, H., Mokhtari, P., Kato, H., Nishimura, R., and Iida, K., "Mechanism for generating peaks and notches of head-related transfer functions in the median plane," *The Journal of the Acoustical Society of America*, 132(6), pp. 3832–3841, 2012.

[20] Turk, M. and Pentland, A., "Eigenfaces for recognition," *Journal of cognitive neuroscience*, 3(1), pp. 71–86, 1991.

[21] Fuß, A., Brinkmann, F., Jürgensohn, T., and Weinzierl, S., "Ein vollsphärisches Multikanalmesssystem zur schnellen Erfassung räumlich hochaufgelöster, individueller kopfbezogener Ubertragungsfunktionen," *Fortschritte der Akustik–DAGA Nürnberg*, pp. 1114–1117, 2015.

[22] Rafaely, B., *Fundamentals of spherical array processing*, Springer, 2015.

[23] Gerbrands, J. J., "On the relationships between SVD, KLT and PCA," *Pattern Recognition*, 14, 1981.

[24] Zhang, L., Dong, W., Zhang, D., and Shi, G., "Two-stage image denoising by principal component analysis with local pixel grouping," *Pattern Recognition*, 43(4), pp. 1531–1549, 2010.

[25] Algazi, V. R., Avendano, C., and Duda, R. O., "Elevation localization and head-related transfer function analysis at low frequencies," *J. Acoust. Soc. Am.*, 109(3), pp. 1110–1122, 2001, doi:https://doi.org/10.1121/1.1349185.

[26] Brinkmann, F., Roden, R., Lindau, A., and Weinzierl, S., "Audibility and interpolation of head-above-torso orientation in binaural technology," *IEEE J. Sel. Topics Signal Process.*, 9(5), pp. 931–942, 2015, doi:https://doi.org/10.1109/jstsp.2015.2414905.