Assessment of Head-Related Transfer Function Time-Alignment Preprocessing Through Spatial Principal Component Analysis

Payman Azaripasand, Bernhard U. Seeber

Audio Information Processing, Technical University of Munich, 80333 Munich, p.azaripasand@tum.de

Introduction and Summary

Due to the symmetrical shape of the head and ears, the Head-Related Transfer Functions (HRTFs) exhibit common attributes across spatial directions. The Head Related Impulse Response (HRIR) signals experience temporal shifts and filtering effects based on the direction of sound arrival at the eardrums. On the other hand, the individual's specific anthropometric measures significantly influence the spatial filtering properties of HRTFs. One effect is the time it takes for the sound to reach the ear canal, and the difference between the time-of-arrivals at each ear is known as Interaural Time Differences (ITDs), which play a fundamental role by providing temporal cues for localization. Methods have been proposed to encode the HRTF dataset into lower dimensional space [1], [2], [3]. However, the encoding methods are prone to losing encoding efficiency and distorted signal reconstruction when treating the HRIR signals containing the time shifts and misaligning signal features such as peaks, notches, and rising edges. Preprocessing methods have also been proposed [3], [4] to streamline the efficient encoding of the HRTF signals. Besides, for effective representation learning by machine learning approaches, target alignment has been suggested to enhance the training of the algorithms by providing faster convergence [5] and higher correlated signal reconstruction [6]. Here, we are interested in finding the signal alignment methods that can perform efficiently on HRIR signals.

In this study, we evaluate different time-alignment preprocessing methods based on Interaural Time Difference calculation (ITD) techniques. It involves aligning all HRIRs of a measurement set by eliminating their relative time delays. The time delays between HRIRs are calculated based on time delay estimation methods previously reported to be efficient on HRIRs [7] and [8]. We then evaluate the effectiveness of the preprocessing methods by monitoring the encoding efficiency of spatial Principal Component Analysis (sPCA). Previously, PCA (Principal Component Analysis) was used to encode the HRTF data in the time or frequency domain. Kistler et al. in [1] Log-magnitude domain HRTFs were decomposed into Directional Transfer Functions (DTFs), and the contribution of the first 5 Principal Components (PCs) was evaluated across different directions. Chen et al. in [9] showed that 12 eigen functions could reconstruct 99% of the variance for KEMAR artificial head log-magnitude HRTFs. Xie et al. [10], [11] also showed that 35 PCs could represent 98 percent of the variance for HRTFs of multiple subjects. In [12] and [2], the KEMAR artificial head HRTFs were encoded in time and complex frequency domain representations, and the results reveal that the 20 spatial PCs can span 99 percent of the variance in both cases. We also propose using sPCA, a linear decomposition method, as it is computationally efficient and stable. Results indicate that threshold-based onset detection yields the most efficient preprocessing by requiring the smallest number of spatial PCs to explain a certain level of variance in the dataset.

Materials and Methods

Dataset

The dataset used in this study is the HUTUBS database [13]. It contains HRTFs sets of 92 unique subjects, each with 440 directional HRTF measurements with azimuth resolution of $\theta \approx 10^{\circ}$, elevation resolution of $\varphi = 10^{\circ}$, and 256 time samples at a sampling rate of 44.1 kHz.

Time-alignment preprocessing

The position-dependent temporal shifts in the HRIRs are responsible for misaligned peaks and notches in the HRIR that cause their encoding to require more eigenfunctions. To overcome these misalignments, the HRIRs in each measurement set for each subject are, by convention, aligned to the ipsilateral HRIR of the left ear at $\theta = 90^{\circ}$ azimuth and $\varphi = 0^{\circ}$ elevation in the same coordinate system as in [13]. The intuition behind this convention is that the signal on the ipsilateral ear has the least onset delay. The time-delay calculation methods we chose were based on [7] and [8]. The delay calculation methods are onset detection based on the threshold level of -10 dB relative to the maximum peak in the signal, maximum cross-correlation of the raw signal, cross-correlation of the minimum-phase components of the signals, centroid of the squared envelope of the signals, centroid of the interaural cross-correlation of the raw signal, and its squared values, and group-delay calculated from the full-pass raw signals. The threshold value of -10 dB was chosen based on its performance compared to other threshold values of -3 dB, -7 dB, and -20 dB. The HRIR signals were first ten times upsampled, the relative delay calculated, the delay subtracted from the signal, band-pass filtered (100 Hz-20 kHz) to remove the possible frequency components above the Nyquist frequency of the original sampling rate, and finally, the signal downsampled to the original sampling rate.

The sPCA method and HRTF decomposition

Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction and data analysis. At its core, PCA aims to transform a dataset into a new coordinate system where the variables are uncorrelated and ordered by their variance. Spatial PCA is a variant of PCA specifically tailored for analyzing spatial data, such as Head-Related Impulse Responses (HRIRs) in auditory research [10]. In this approach, the HRIR measurement set is encoded in the spatial dimension of the data rather than the other signal dimension which is in the time or frequency domain. We followed the sPCA decomposition approach of [10] and [11] in the discrete time domain. The indices for the ear are removed for simplicity. In equation (1), $w_q(\theta, \varphi)$ is the basis function that only depends on the direction, it can be identical or different for each ear as well. Besides, $g_q(t, s)$ is the time domain and individual dependent weights.

$$h(\theta, \varphi, t, s) = \sum_q g_q(t, s) w_q(\theta, \varphi). \qquad (1)$$

The matrix calculations to decompose HRIRs into matrices g and w, and recovery of the HRIRs from a subset of spatial PCs, are explained in detail in [10].

Evaluations

As sPCA reduces the spatial dimensionality of the HRTF data by projecting it onto orthogonal axes defined by the eigenvectors of the covariance matrix, the eigenvalues associated with these components represent the variance explained by each. By summing these eigenvalues cumulatively, the cumulative energy curve unveils the principal components increases. This representation delineates the trade-off between dimensionality reduction and information preservation, aiding in the selection of an optimal number of components. The cumulative energy is calculated as:

$$Cumulative\ energy = \frac{\sum_{q=1}^{M} \lambda_q}{\sum_{q=1}^{Q} \lambda_q} \times 100\% \qquad (2)$$

Where λ_q: denotes the eigenvalue corresponding to each principal component sorted in descending order, Q is the total number of directional PCs, and M is the selected subset of PCs.

Results and Discussion

To evaluate the reconstruction efficiency of HRIRs from a different subset of spatial principal components, the cumulatuive energy is calculated for each of the time-alignment preprocessing methods using Formula (1) and the results are plotted in **Figure *1***. Two hard threshold values of 90 and 99 percent of the total energy are considered to compare the encoding efficiency. Recent work by Gavish and Donoho [14] also provides a theory-based approach to determining the optimal threshold for the singular value truncation. In the Gavish-Donoho criterium calculation and on our PCA results, the noise profile in the method was estimated from the least significant principal component values, and the truncation threshold values were then calculated. To compare the efficiency of the preprocessing methods across 92 subjects, the encoding efficiency for the aforementioned criteria is depicted in Figure 2 for seven candidate preprocessing methods. The results show that at 90 percent of the HRIR energy can be reconstructed with a mean of 8.05 (std: 1.18) spatial PCs using -10 dB relative threshold preprocessing while the original HRIRs need 19.30 PCs (std: 2.36). To reconstruct 99 percent of the HRIR energy, using the -10 dB threshold method 32.08 (std: 3.76) spatial PCs are required, while for the original, unprocessed data 49.73 (std: 3.77) PCs are needed. To satisfy the Gavish-Donoho criterium, 20.09 (std: 3.67) PCs are required using the -10 dB threshold method; but 31.11 PCs for the unprocessed data (std: 6.42). The reason for the efficiency of the -10 dB threshold over other methods can be explained by being less sensitive to early peaks and multipath propagation effects in the HRIR signals.

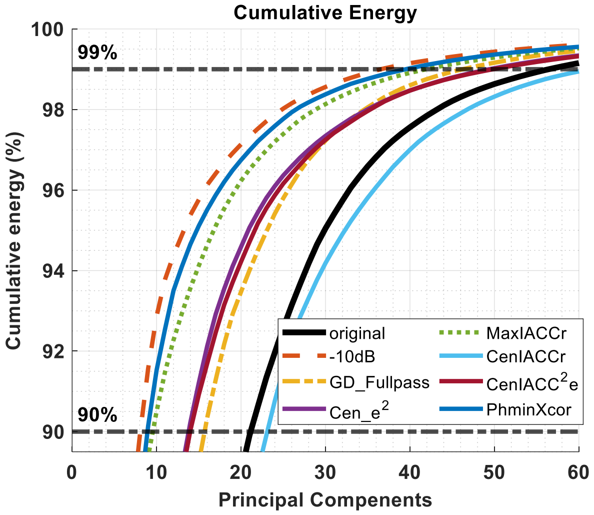

Figure 1: The plot represents the cumulative energy at specific numbers of principal components calculated from the PCA results of the original and preprocessed HRIRs using -10 dB relative threshold, group-delay, centroid of the squared envelope, maximum of the cross-correlation, centroid of the cross-correlation and its squared, and also cross-correlation of the minimum phase components of the signals. The cumulative energy is plotted for 60 principal components (out of 440) for ease of demonstration.

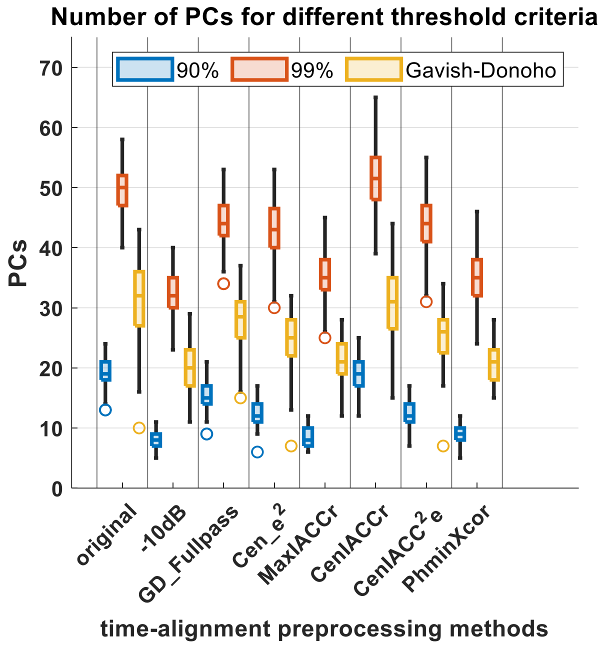

Figure 2: The distribution of the minimum number of PCs calculated for the three threshold criteria of 90 and 99 percent of cumulative energy and Gavish-Donoho across 92 subjects. Results are given for the seven delay calculation methods.

Conclusion

In this study, sPCA was employed to assess various time-alignment preprocessing methods for HRIRs. The impact of different delay calculation methods on encoding efficiencies was evaluated by comparing the number of PCs required to achieve the same cumulative energy level after preprocessing. The findings indicated that aligning HRIRs to -10 dB relative threshold level consistently yielded the lowest average number of PCs necessary for reconstructing HRIRs to achieve equivalent energy levels across all three truncation criteria. Thus, the -10 dB relative threshold time alignment emerges as the preferred choice for optimized encoding of HRIR signals.

References

[1] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1637-1647, 1992, doi: 10.1121/1.402444.

[2] T. Shouichi, "Spatial Principal Component Analysis of Head-Related Transfer Functions and Its Domain Dependency," in *Advances in Principal Component Analysis*, M. Fausto Pedro García Ed. Rijeka: IntechOpen, 2022, p. Ch. 9.

[3] F. Brinkmann and S. Weinzierl, *Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition*. 2018.

[4] J. M. Arend, C. Pörschmann, S. Weinzierl, and F. Brinkmann, "Magnitude-Corrected and Time-Aligned Interpolation of Head-Related Transfer Functions," *arXiv e-prints,* p. arXiv:2303.09966, 2023/3 2023, doi: 10.48550/arXiv.2303.09966.

[5] E. Imani, W. Hu, and M. White, "Representation Alignment in Neural Networks," *arXiv e-prints,* p. arXiv:2112.07806, 2021, doi: 10.48550/arXiv.2112.07806.

[6] Y. Lou, C. E. Mingard, and S. Hayou, "Feature Learning and Signal Propagation in Deep Neural Networks," presented at the Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research, 2022.

[7] B. F. G. Katz and M. Noisternig, "A comparative study of interaural time delay estimation methods," *The Journal of the Acoustical Society of America,* vol. 135, no. 6, pp. 3530-3540, 2014, doi: 10.1121/1.4875714.

[8] J. Bernhard, G. Gomez, and B. Seeber, "Time-domain interpolation of head-related transfer functions with correct reproduction of notch frequencies," in *Fortschritte der Akustik--DAGA'15*, 2015, pp. 1126-1127.

[9] J. Chen, B. D. Van Veen, and K. E. Hecox, "A spatial feature extraction and regularization model for the head-related transfer function," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 439-452, 1995, doi: 10.1121/1.413110.

[10] B. Xie, "Recovery of individual head-related transfer functions from a small set of measurements," *The Journal of the Acoustical Society of America,* vol. 132 1, pp. 282-94, 2012, doi: 10.1121/1.4728168.

[11] B. Xie, R. P. I. Dr. Ning Xiang, and J. Blauert, *Head-Related Transfer Function and Virtual Auditory Display: Second Edition*. J. Ross Publishing, 2013.

[12] S. Takane, "Spatial Principal Component Analysis of Head-Related Transfer Functions using their complex logarithm with unwrapping of phase," 2019.

[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," *journal of the audio engineering society,* vol. 67, no. 9, pp. 705-718, 2019/9 2019, doi: 10.17743/jaes.2019.0024.

[14] M. Gavish and D. L. Donoho, "The Optimal Hard Threshold for Singular Values is 4/sqrt(3)," *arXiv e-prints,* p. arXiv:1305.5870, 2013, doi: 10.48550/arXiv.1305.5870.