COMPRESSION OF HEAD-RELATED TRANSFER FUNCTIONS USING PIECEWISE CUBIC HERMITE INTERPOLATION

Tom Krueger and Julián Villegas

Dept. of Computer Science and Engineering
University of Aizu
Aizu-Wakamatsu, Japan
{m5282010, julian}@u-aizu.ac.jp

ABSTRACT

We present a spline-based method for compressing and reconstructing Head-Related Transfer Functions (HRTFs) that preserves perceptual quality. Our approach focuses on the magnitude response and consists of four stages: (1) acquiring minimum-phase head-related impulse responses (HRIR), (2) transforming them into the frequency domain and applying adaptive Wiener filtering to preserve important spectral features, (3) extracting a minimal set of control points using derivative-based methods to identify local maxima and inflection points, and (4) reconstructing the HRTF using piecewise cubic Hermite interpolation (PCHIP) over the refined control points. Evaluation on 301 subjects demonstrates that our method achieves an average compression ratio of 4.7:1 with spectral distortion ≤ 1.0 dB in each Equivalent Rectangular Band (ERB). The method preserves binaural cues with a mean absolute interaural level difference (ILD) error of 0.10 dB. Our method achieves about three times the compression obtained with a PCA-based method.

1. INTRODUCTION

Head-Related Transfer Functions (HRTFs) capture how sound is transformed from a source to the listener's ears. The individuality of HRTFs stems from the physical features of the listener such as the size and shape of the head, torso, and pinnae, and the location of the sound source relative to the listener. HRTF compression and reconstruction are important for spatial audio, virtual reality, and hearing aid applications. Dense HRTF datasets (i.e., those comprising measurements at many locations) are essential for accurate sound localization but also face practical challenges. Each HRTF is represented by numerous frequency bins [1] and the storage of an entire database can exceed several hundred megabytes [2]. This is problematic in resource-limited environments, such as mobile or wearable audio devices, where memory limitations and processing capabilities directly impact real-time performance [3]. Furthermore, personalized HRTFs greatly improve localization accuracy compared to generic HRTFs [4], but are expensive and difficult to capture. So, finding data representations that could be used for interpolating existing HRTFs as a way of personalization is desirable. As general awareness of spatial audio and HRTF personalization increases, efficient methods for storing and adapting HRTFs to a given listener become a key point to consider for practical implementation [5].

Compression and reconstruction of HRTFs should not degrade their perceptual attributes. In addition, reconstruction (decoding) should be fast and simple to implement. Recent research on HRTF compression has shown multiple methods aimed at reducing data size efficiently while maintaining perceptual quality. Marentakis et al. [6] showed that Principal Component Analysis (PCA) applied to HRTF magnitude spectra significantly improves the compression ratio compared to time domain representations. Arévalo and Villegas [7] proposed a method based on Eigen decomposition of HRTFs, capable of achieving a compression ratio of 15:1 with < 1 dB spectral distortion in the range of 100 Hz and 16kHz. In that case, different Eigen values for each measurement and subject were used. Additionally, the “TT-Tucker” method proposed by Wang et al. [8] employs multidimensional tensor decomposition, achieving approximately 98% data compression with superior perceptual fidelity compared to simpler PCA-based methods, according to the authors.

Spherical harmonics have also been used. Lie et al. [9] introduced a frequency-dependent harmonic decomposition of HRTFs using a mixed-order approach (i.e., a 8-order for low frequencies and a 22-order for high frequencies). They claimed to reduce the number of coefficients without degrading the accuracy of perceptual localization.

In different domains, spline interpolation has been employed for its smoothness and precision in representing continuous data from discrete points [10]. For HRTF reconstruction, splines have been explored by Carlile et al. [11], who used spherical thin-plate splines to create continuous virtual audio from a small number of measurements. Their research indicates that a high-fidelity Virtual Auditory Space (VAS) could be achieved with only 150 HRTF measurements. Their approach focused mainly on spatial interpolation rather than compression, and did not explore a minimal set of control points needed to preserve spectral features. Völkering et al. [12] showed that cubic splines on sparse horizontal-plane HRTFs reduce the mean spectral error by up to 1 dB compared to linear interpolation for 10° to 20° sampling.

The Spatially Oriented Format for Acoustics (SOFA) by Majdak et al. [13] is a standard for HRTF datasets. In its current version, SOFA supports two compressed representations in addition to raw impulse responses: spherical harmonic decompositions and cascaded second-order section (SOS) filter chains. Implying that the previously HRTF compression alternatives (or the one proposed here) cannot be leveraged by the standard.

The purpose of this research is to evaluate the compression and reconstruction of HRTF magnitude spectra using spline-based methods. We aim to reduce data storage, while ensuring that the spectral distortion does not exceed 1.0 dB for each Equivalent Rectangular Band (ERB), which model the frequency discrimination performed by the basilar membrane. Furthermore, we impose an absolute magnitude error between the original and reconstructed HRTFs of ≤ 1.0 dB across frequency bands, as explained in the following sections. While our current concern is in HRTF compression, the proposed representation could pave the way for HRTF personalization by standardizing the number of spline control points and modifying only their relative weights.

2. METHOD

Overall, our method comprises the four stages shown in Figure 1: (1) Data acquisition: loading minimum phase, no interaural time difference (ITD) head-related impulse responses (HRIRs) from a database [14], (2) Computing their respective HRTFs and applying adaptive Wiener filtering [15] to each magnitude spectrum, (3) extracting control points using derivative-based methods with iterative segmentation, pruning, and (4) reconstructing the HRTF magnitude response using spline interpolation (PCHIP) [16]. A freely available implementation of this method is found at `https://github.com/tomKruegerJapan/HRTF_CompInter`.

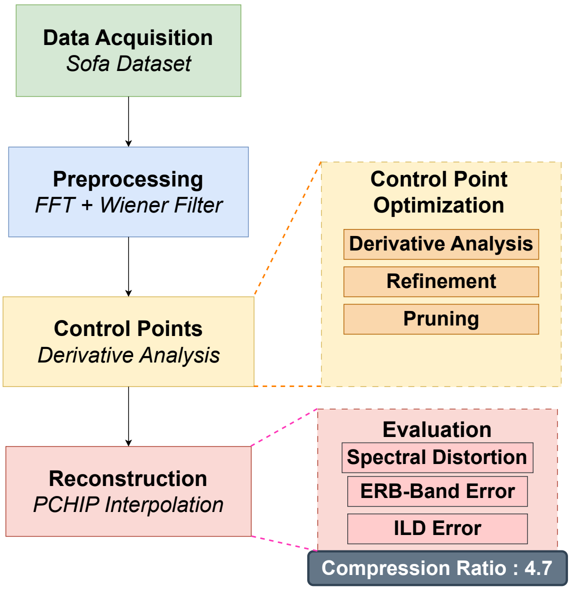

Figure 1: *Four stage pipeline of the proposed method.*

2.1. Data and Acquisition

We use the publicly available SONICOM HRTF database [14]. Specifically, we used 793 minimum-phase (with no ITD) HRIRs of 301 subjects, sampled at 48 kHz, for a total of 238, 693 measurements. The minimum-phase representation eases the separate treatment of spectral and temporal characteristics of HRTFs.

The corresponding SOFA files of these HRIRs were loaded using the "sofa" library (version 0.2.0) in Python 3.11. These files contain the sampling rate, impulse responses for both channels (left and right), and source location in spherical coordinates relative to the center of the head.

2.2. Preprocessing and Filtering

Each impulse response was transformed into the frequency domain using the Discrete Fourier Transform (DFT). We compute the magnitude frequency response (in dB) up to 20 kHz. The resulting HRTFs present spectral noise (i.e., local peaks and troughs), which are unlikely to affect spatial perception, but pose a hurdle for the spline decomposition. To smooth such local spectral "ripples," several methods have been proposed in the past, including cepstral smoothing and moving average over third-octave (or smaller) bands [17], however, preliminary experiments indicated that applying an adaptive Wiener filter [15] to each magnitude response outperformed them, therefore we opted for it. Note also that although the use of Wiener filtering for HRTF is uncommon, it has been successfully applied to reduce artifacts in other audio related problems such as source separation [18]. The smoothed magnitude M_{smooth} for a given frequency bin f of an HRTF is computed as

$$M_{\text{smooth}}(f) = \text{Wiener}\big(M_{\text{orig}}(f); w_{\text{base}}, \sigma_{\text{noise}} = 0.1\big), \qquad (1)$$

where M_{orig}, w_{base}, and σ_{noise} represent the original magnitude, the base window size, and the noise-power estimate, respectively. The actual base window size depended on an error threshold $T = 1.0$ dB. When equation 2 holds

$$\big|M_{\text{orig}}(f) - M_{\text{smooth}}(f)\big| \leq T, \qquad (2)$$

11 frequency bins were used, otherwise, 7 bins were used instead. These settings were empirically determined.

2.3. Control Points Refinement and Pruning

After smoothing, we computed a set of control points for the reconstruction of the HRTFs. Control point candidates were identified with the first frequency derivative of the magnitude response to locate local extrema. Additional control point candidates (related to inflection points) were found with the second derivative. These derivative-based methods yield a great number of points in some cases. To reduce them, we implemented an iterative process of segmentation and pruning. First, we simulated the reconstruction with the current set of control points. We then segmented frequency regions inside individual ERBs that showed high Spectral Distortion (SPD). In these segments, additional points were inserted at the highest error value of that specific segment. Once the segmentation is finished, we apply a multi-stage pruning process to remove redundant or overly dense points while preserving reconstruction accuracy. First, a pruning function iterates over the control points, and for each candidate temporarily removes it and reconstructs the response on a sampled frequency grid. The resulting reconstruction error is compared to a 1.0 dB threshold and if the removal does not yield a larger error, the point is pruned. Next, an ERB pruning step checks the spectral distortion of the reconstruction. A control point is removed only if the reconstruction error remains below the 1.0 dB threshold for each ERB. This control point selection is inspired by the optimal changepoint detection methods described by Killick et al. [19].

2.4. Spline Interpolation and Reconstruction

Spline interpolation was used to reconstruct the complete HRTF from the calculated control points. We evaluated two reconstruction methods: 1) cubic spline [20] and 2) piecewise cubic Hermite interpolating polynomial (PCHIP) interpolation [16]. Both methods were compared using the full SONICOM dataset on 42 ERB bands by comparing the error and SPD of the original HRTF magnitude response and the reconstructed response. We also computed the mean absolute interaural level difference (ILD) error. In agreement with previous works [2, 21], it was found that the PCHIP method provided better reconstruction results, so we limit our discussion to this method only.

3. RESULTS

3.1. Compression ratio

Each original magnitude response was represented using 427 frequency bins (up to 20 kHz), multiplied by 2 channels and 4 bytes per sample, yields an uncompressed size of 3 416 bytes per HRTF. The compressed size is obtained by multiplying the number of control points (both channels) by 2 (frequency and magnitude) and by 4 bytes per value. Using the entire dataset (301 subjects, each with 793 measurements), the proposed method reaches an average compression ratio of $4.7{:}1$ (SD $= 1.19$).

3.2. Spectral Distortion (SPD) and Reconstruction Accuracy

The resulting control points, reconstruction, and error between original and reconstruction are illustrated in Figure 2. The top panel of Figure 2 shows the original frequency response (solid line), the reconstructed frequency response after PCHIP spline interpolation (dashed line), and the calculated control points (black marks). The bottom panel shows the magnitude error per frequency bin. As shown in this panel, the error never exceeds ± 1.0 dB. PCHIP achieves a mean SPD $= 0.37$ dB (SD $= 0.03$) with 97 (SD $=$ 23.4) control points per measurement, while cubic splines achieve a similar result with slightly more control points due to overshoot in some regions. Across all measurements, the reconstruction errors are ≤ 1.0 dB in each ERB (42), as shown in Figure 3.

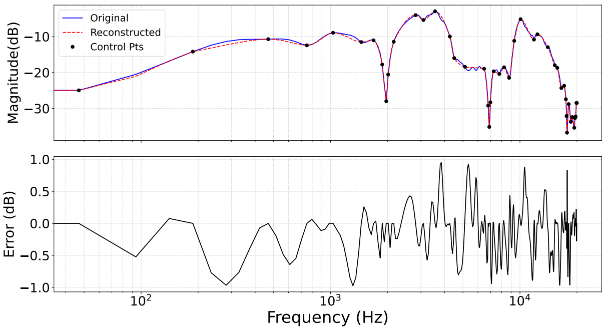

Figure 2: *Magnitude reconstruction of the left channel of subject P0137 for azimuth $\theta = 0°$ and elevation $\phi = -30.0°$, showing the error maintaining the limit of ± 1 dB.*

The total full-band spectral distortion, averaged over all $238,693$ reconstructions, is 0.334 dB (SD $= 0.03$). Likewise, across all 42 ERBs, the reconstruction errors are < 1.0 dB, as shown in Figure 3.

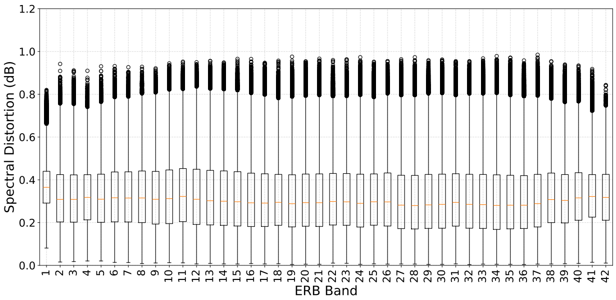

Figure 3: *Spectral distortion over 42 ERBs across all $238,693$ reconstructions (301 subjects $\times$ 793 directions).*

3.3. Other results

Interaural level differences were obtained by subtracting the magnitude response of the right channel from that of the left channel. The ILD error is then calculated by the absolute difference between the original ILD and the reconstructed ILD. The resulting mean absolute ILD error across all measurements is 0.1 dB (SD $= 0.07$). Note that error in interaural time differences (ITDs) cannot be computed with our dataset selection, and that for actual audio spatialization, we rely on the ITDs provided by SONICOM.

Figure 4 shows the absolute error introduced by the Wiener filter. On average, this error was of 0.30 dB (SD $= 0.05$). Regarding the number of control points, we were able to remove an average of 13 and 10 control points for the left and right channel, respectively. This represents a reduction of about 20%. This resulted in an average of $48, 5$ control points for each channel. In our current implementation, control points for the left and right HRTF channels are selected independently to best preserve each channel's unique spectral cues. Consequently, the average number of control points differs between channels, resulting in 47 and 50 control points for the left and right channel, respectively.

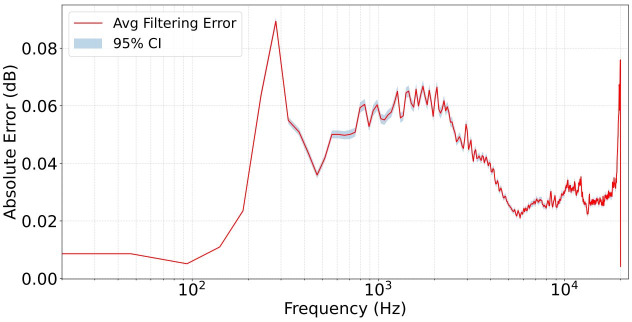

Figure 4: Average absolute Wiener filtering error computed as $|M_{\text{smooth}}(f) - M_{\text{orig}}(f)|$, and averaged across 301 subjects.

3.4. Comparison with other methods

We implement a PCA-based HRTF compression baseline using the "scikit-learn" (Version 1.6.1) library, following Grijalva et al. [22]. For each measurement, we select the minimal number of principal components needed to keep full-band spectral distortion below 1.0 dB. Table 1 compares the results of our spline method with this PCA baseline.

Table 1: Mean performance comparison of PCA-based compression versus our spline control point method (301 subjects). Standard deviation in parenthesis.

Metric	**PCA Baseline**	**Proposed Method**
Compression Ratio	$1.58(0.10)$:1	$\mathbf{4.7}(\mathbf{1.19})$:1
full-band spectral distortion	$\mathbf{0.273}(\mathbf{0.170})$ dB	$0.370(0.030)$ dB
No. Parameters	$120.00(9.00)$	$\mathbf{97}(\mathbf{23.38})$
ILD Error	$\mathbf{0.011}(\mathbf{0.01})$ dB	$0.1(0.07)$ dB

As shown in Table 1, our method achieves about three times the compression of the PCA-based method while reducing the control points (components) by approximately 20%. The proposed method achieves a slightly higher full-band SPD and ILD errors, however they remain very low, ensuring faithful spatial cues.

3.5. Computational Performance

All timing experiments were run on Windows 11 using AMD 64 having 8 physical cores and 16 logical cores with up to 3.2 GHz and a 16 GB of RAM. We compared two processing pipelines over 793 measurements per subject: 1) SOFA pipeline: HRIR→FFT→complex spectrum; and 2) Control-point pipeline: control points→PCHIP magnitude reconstruction→complex spectrum.

The SOFA pipeline achieves a mean computation time of $24\,\mu$s per measurement (total 18.8 ms), while the control point approach requires approximately 0.171 ms (total 135.8 ms) respectively. I.e., control point pipeline is 0.14 times slower than the SOFA pipeline.

4. DISCUSSION

While our proposed method achieves less compression than Arévalo and Villegas [7], we present control points that are simple to manipulate and thus, potentially enabling personalization.

The Wiener filter smoothing reduces unwanted ripples while preserving important resonances. As mentioned before, spectral-smoothing techniques have been proposed in the past [23]. We compared our adaptive Wiener filter against cepstral and fractional octave smoothing on a randomly selected subject. Cepstral smoothing produced large local filtering errors ($>$ 20 dB) in some frequency bins. Similarly, third-octave smoothing introduced absolute errors above the self-imposed 1 dB threshold. Consequently, the Wiener filtering was used instead. For practical spatial audio applications, the reconstruction from this compressed representation is computationally efficient. The stored control points are interpolated using PCHIP splines to recover the magnitude spectrum. To spatialize sound using our compressed HRTFs, we convert the dB magnitude to linear magnitude and combine that with the phase response assumed to be linear [23]. Thus, we can convolve the reconstructed HRTF with a target sound by spectral multiplication. Applying an inverse FFT to this yields the spatialized audio.

To provide a better view of the reconstruction performance, independent of specific hardware, we can replace the raw timing with the following analysis. Since control points extraction, segmentation, and pruning are all performed offline, the following cost analysis refers only to the decoding (reconstruction) step, which is the more relevant part. Building the reconstruction with PCHIP from C control points has $\mathcal{O}(C)$. Evaluating it at N query frequencies requires locating each spline interval (via binary search) in $\mathcal{O}(\log C)$, for a total of $\mathcal{O}(N \log C)$. Converting the N reconstructed magnitudes into complex-valued spectrum samples adds $\mathcal{O}(N)$. Hence, the composite complexity is

$$\mathcal{O}\big(C + N \log C + N\big) = \mathcal{O}\big(N \log C\big),$$

and across all M measurements

$$\mathcal{O}\big(M\,N \log C\big).$$

Future work could explore the identification of control points that are consistent in different spatial locations, which could potentially reduce storage requirements. Additionally, the 1 dB spectral distortion threshold could be adjusted based on perceptual relevance. E.g., increasing this threshold for frequencies bands where hearing is less sensitive [24]. In addition, since ERBs model regions where we are unable to discriminate between two frequencies, which makes it possible to further eliminate control points within the same ERB, and keep only one per band. Furthermore, the core segmentation and pruning used to select control points offer additional room for compression. Merging redundant control points into single representative points whenever their combined interpolation error stays below the thresholds could further increase the compression. The primary limitation of our approach is that we did not conduct subjective testing. We will tackle this limitation in the near future.

5. CONCLUSION

This work presents an efficient and accurate method for reconstructing HRTFs using optimized control point calculation combined with PCHIP spline interpolation. Our approach effectively identifies and preserves spectral features through a derivative-based process, ensuring reconstruction accuracy consistently within a 1.0 dB error margin across all ERBs. Taking advantage of critical points and inflection points of the magnitude response, we achieve an average compression ratio of 4.7:1 while maintaining full-band spectral distortion of 0.37 with a SD of 0.030 dB and a mean absolute ILD error of 0.10 and a SD of 0.070 dB over $238,693$ measurements. Compared to a PCA-based baseline, the control point representation reduces stored parameters by approximately 20 % for similar perceptual quality. Although interpolation itself is computationally efficient, further optimizations in control point calculation and refinement such as more aggressive pruning, segmentation, and merging, also offer room for improvement. Beyond compression, the compact control point format enables the potential of direct manipulation of spectral features for personalized HRTFs.

6. ACKNOWLEDGMENTS

The authors thank Mr. Mohamed Maatar and Mr. Timoté Moreaux for their support and the anonymous reviewers for their invaluable comments and criticism.

7. REFERENCES

[1] Fabian Brinkmann, Manoj Dinakaran, Robert Pelzer, Peter Grosche, Daniel Voss, and Stefan Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and head-

phone impulse responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718, 2019, DOI: 10.14279/depositonce-15233.

[2] Hannes Gamper, "Head-related transfer function interpolation in azimuth, elevation, and distance," *J. Acoust. Soc. Am.*, vol. 134, no. 6, pp. EL547–EL553, 2013, DOI: 10.1121/1.4828983.

[3] Johannes M Arend, Fabian Brinkmann, and Christoph Pörschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *J. Audio Eng. Soc.*, vol. 69, no. 1/2, pp. 104–117, 2021, DOI: 10.17743/jaes.2020.0070.

[4] Elizabeth M Wenzel, Marianne Arruda, Doris J Kistler, and Frederic L Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123, 1993, DOI: 10.1121/1.407089.

[5] Johannes M Arend, Christoph Pörschmann, Stefan Weinzierl, and Fabian Brinkmann, "Magnitude-corrected and time-aligned interpolation of head-related transfer functions," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3783–3799, 2023, DOI: 10.1109/taslp.2023.3313908.

[6] Georgios Marentakis and Josef Hôlzl, "Compression efficiency and signal distortion of common PCA bases for HRTF modelling," in *Proc. 18 Sound and Music Comput. Conf.*, 2021.

[7] Camilo Arévalo and Julián Villegas, "Compressing head-related transfer function databases by Eigen decomposition," in *IEEE Int. Wkshp. on Multimedia Signal Process. (MMSP)*, Tampere, Finland, 2020, DOI 10.1109/MMSP48831.2020.9287134.

[8] Jing Wang, Min Liu, Xiang Xie, and Jingming Kuang, "Compression of HRTF based on TT-Tucker decomposition," *IEEE Access*, vol. 7, pp. 39639–39651, 2019, DOI: 10.1109/ACCESS.2019.2906364.

[9] Junfeng Li, Biao Wu, Dingding Yao, and Yonghong Yan, "A mixed-order modeling approach for HRTFs in the spherical harmonic domain," *App. Acoust.*, vol. 176, pp. 107828, 2021, DOI: 10.1016/j.apacoust.2020.107828.

[10] Frederick N Fritsch and Ralph E Carlson, "Monotone piecewise cubic interpolation," *SIAM J. Num. Analysis*, vol. 17, no. 2, pp. 238–246, 1980, DOI: 10.1137/0717021.

[11] S Carlile, C Jin, and V Van Raad, "Continuous virtual auditory space using HRTF interpolation: Acoustic and psychophysical errors," in *Proc. IEEE Pacific-Rim Conf. Mltmed.*, 2000, pp. 220–223.

[12] Jan Völkering, Eugen Rasumow, and Matthias Blau, "Examination of different HRTF interpolation methods," in *DAGA Conference, Oldenburg, Germany*, 2014.

[13] Piotr Majdak, Fabian Brinkmann, Julien De Muynke, Michael Mihocic, and Markus Noisternig, "Spatially oriented format for acoustics 2.1: Introduction and recent advances," *J. Audio Eng. Soc.*, vol. 70, pp. 565–584, 2022.

[14] Engel Isaac, Daugintis Rapolas, Vicente Thibault, Hogg Aidan O. T., Pauwels Johan, Tournier Arnaud J., and Picinali Lorenzo, "The SONICOM HRTF dataset," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 241–253, 2023, DOI: 10.17743/jaes.2022.0066.

[15] Jingdong Chen, Jacob Benesty, Yiteng Huang, and Simon Doclo, "New insights into the noise reduction wiener filter," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, 2006, DOI: 10.1109/TSA.2005.860851.

[16] F. N. Fritsch and J. Butland, "A method for constructing local monotone piecewise cubic interpolants," *SIAM J. on Sci. and Statistical Comput.*, vol. 5, no. 2, pp. 300–304, 1984, DOI: 10.1137/0905021.

[17] Huopaniemi Jyri and SmithIII Julius O., "Spectral and time-domain preprocessing and the choice of modeling error criteria for binaural digital filters," in *Proc. 16 Audio Eng. Soc. Int. Conf.*, 1999, pp. 301–312.

[18] Emmanuel Vincent, "An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation," in *Inter. Conf. on Latent Variable Analy. and Signal Separation*. Springer, 2010, pp. 157–164, DOI: 10.1007/978-3-642-15995-4_20.

[19] Rebecca Killick, Paul Fearnhead, and Idris A Eckley, "Optimal detection of changepoints with a linear computational cost," *J. Am. Stat. Assoc.*, vol. 107, no. 500, pp. 1590–1598, 2012, DOI: 10.1080/01621459.2012.737745.

[20] Carl d. Boor, *A Practical Guide to Splines*, Springer Verlag, New York, 1978.

[21] Hugeng Hugeng, Jovan Anggara, and Dadang Gunawan, "Implementation of 3d HRTF interpolation in synthesizing virtual 3d moving sound," *Int. J. Tech.*, vol. 8, no. 1, pp. 186–195, 2017, DOI: 10.14716/ijtech.v8i1.6859.

[22] Felipe Grijalva, Boris Escobar, Byron Alejandro Acuna Acurio, and Robin Álvarez, "Analysis and synthesis of hrtfs using principal component analysis," in *IEEE 4 Ecuador Technical Chapters Meeting (ETCM)*, 2019, pp. 1–6.

[23] Jyri Huopaniemi, Nick Zacharov, and Matti Karjalainen, "Objective and subjective evaluation of head-related transfer function filter design," *J. Audio Eng. Soc.*, vol. 47, no. 4, pp. 218–239, 1999.

[24] John C Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1493–1510, 1999, DOI: 10.1121/1.427147.