Audio Engineering Society

Conference Paper

Presented at the 2022 International Conference on
Audio for Virtual and Augmented Reality
2022 August 15–17, Redmond, WA, USA

Binaural signal matching with arbitrary array based on a sound field model

Shai Hermon[1], Vladimir Tourbabin[2], Zamir Ben-Hur[2], Jacob Donley[2], and Boaz Rafaely[1]

[1]*School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel*
[2]*Reality Labs Research at Meta, Redmond, WA, USA*

Correspondence should be addressed to Shai Hermon (hermons@post.bgu.ac.il)

ABSTRACT

The capture and reproduction of spatial audio is becoming increasingly important with the mushrooming of applications in teleconferencing, entertainment, and virtual reality. Two popular methods include High Order Ambisonics (HOA) and beamforming-based binaural reproduction (BFBR), which are mainly studied with spherical arrays. Recently, binaural signal matching (BSM) has been proposed for binaural reproduction with arbitrary arrays. This paper investigates improvement of the method, which is based on a model of the sound field separated into direct and reverberant parts. A simulation study shows that the proposed method improved the performance of the BSM algorithm for some tested acoustic scenarios without the need for additional information about the sound field.

1 Introduction

Binaural reproduction has recently become an important and widespread topic of research [1 - 3], following the rapid growth of many applications, including entertainment, augmented reality and virtual reality. A common method of binaural signal reconstruction from a spherical microphone array is based on high order Ambisonics (HOA) [4, 5], together with head-related transfer functions (HRTFs) [6]. The HOA signals are typically synthesized with a computer or estimated from microphone-array measurements [7 - 9]. An accurate estimation of HOA signals is usually required in order to render binaural signals with good perceptual qualities. Hence, many research works have studied methods for capturing HOA signals [10], and combining them for binaural reproduction [11, 6]. These methods have been generalized for other arrays as well, but the best results are obtained for spherical arrays. Another approach is beamforming-based binaural reproduction (BFBR) [12, 13]. In this method, a set of beamformers is applied to the microphone signals with different look directions, followed by convolution with the HRTFs from the same directions. Finally, the output signals are weighted and summed to create the binaural signals. BFBR is more general than HOA, but again most of the algorithms used spherical arrays [12 - 14]. Some used other array types [15, 16], but

a comprehensive design methodology and analysis of performance were not presented.
In practice, it is necessary to find high quality methods for binaural reproduction for general arrays. Some examples include microphones on handheld devices or wearable arrays, which are typically limited by the geometry of the device on which they are mounted. For such arrays, the beam-pattern may depend on the look direction, and their directional resolution may be limited, such that they are not capable of capturing HOA signals. Hence, performing high-quality binaural reproduction with measurements that have been captured with general arrays is still a great challenge. Recently, binaural signal matching (BSM) [17, 18] has been proposed for binaural reproduction with arbitrary arrays. This method assumes that the sound field is diffuse, and is based on a Wiener filter to match between the microphone signals and the binaural signals. The shortcoming of the BSM method is the assumption of a diffuse field, while, in many cases, there is high correlation between the sources because some of the sources are reflections of the direct signal. The aim of this research work is to develop the BSM algorithm by including the assumption of correlation between the sources, and representing the sound field using a model with two components, a direct sound field and a diffuse field.

2 Binaural signal matching

This section presents the BSM method. The signal model employed in this paper is given by:

$$\mathbf{x}(k) = \mathbf{V}(k)\mathbf{s}(k) + \mathbf{n}(k) \tag{1}$$

where $\mathbf{x}(k) = [x_1(k), x_2(k), ..., x_M(k)]$ is the microphone signals (measurements) vector, and k is the frequency index. $\mathbf{V}(k) = [\mathbf{v}(k,\theta_1,\phi_1), \mathbf{v}(k,\theta_2,\phi_2), ..., \mathbf{v}(k,\theta_L,\phi_L)]$ is an $M \times L$ complex matrix, where M represents the number of microphones and L the number of sources. $\mathbf{v}(k,\theta_l,\phi_l)$ represents the array steering vector from the l'th source to the microphone positions for all L sources, $\mathbf{s}(k) = [s_1(k), s_2(k), ..., s_L(k)]$ is the source signals vector, and $\mathbf{n}(k)$ is an additive noise vector.
The model of the binaural signals is given by:

$$p^{r,l}(k) = [\mathbf{h}^{r,l}(k)]^T \mathbf{s}(k) \tag{2}$$

where $p^{r,l}(k)$ is the sound pressure in the right or left ear, and $\mathbf{h}^{r,l}(k) = [h^{r,l}(\theta_1,\phi_1,k), h^{r,l}(\theta_2,\phi_2,k), ...h^{r,l}(\theta_L,\phi_L,k)]$ is the HRTF. In the first step, the array measurements are filtered and combined, in a similar manner to beamforming, where $z^{r,l}$ is defined as:

$$z^{r,l} = [\mathbf{c}^{r,l}]^H \mathbf{x} \tag{3}$$

where $\mathbf{c}$ is an $M \times 1$ vector holding the filter coefficients. Next, $\mathbf{c}$ is chosen to minimize the following mean squared error between equation (3) and $p^{r,l}$, the binaural signals in equation (2), for each ear separately:

$$err^{r,l}_{bin} = E[\mid p^{r,l} - z^{r,l} \mid^2] \tag{4}$$

where, E is the expectation operator. Next, assume that the noise is uncorrelated to the sources signals and is white with a covariance matrix $E[\mathbf{n}\mathbf{n}^H] = \mathbf{I}\sigma_n^2$. Then, substituting equations (2) and (3) in equation 4 leads to the following error formulation:

$$err_{bin} = (\mathbf{c}^H\mathbf{V} - \mathbf{h}^T)\mathbf{R}_{ss}(\mathbf{c}^H\mathbf{V} - \mathbf{h}^T)^H + \sigma_n^2\|\mathbf{c}\|_2^2 \tag{5}$$

where $\mathbf{R}_{ss}$ is the correlation between the sources at each frequency and the parameter $\mathbf{c}$, err_{bin} and $\mathbf{h}$ depend on the left or right ears. In order to produce an accurate binaural signal, equation (5) is minimized over the filter $\mathbf{c}^{l,r}$ for each ear, and it can be shown that the optimal filter can be formulated as:

$$\mathbf{c}^{r,l}_{opt} = (\mathbf{V}\mathbf{R}_{ss}\mathbf{V}^H + \sigma_n^2\mathbf{I})^{-1}\mathbf{V}\mathbf{R}_{ss}[\mathbf{h}^{l,r}]^* \tag{6}$$

where $\mathbf{c}^{l,r}_{opt}$ are the optimal filters for the left and right ears, respectively. For the BSM filter, assume that $\mathbf{R}_{ss} = \mathbf{I}\sigma_s^2$ to get [17]:

$$\mathbf{c}^{r,l}_{BSM} = (\mathbf{V}\mathbf{V}^H + \frac{\mathbf{I}}{SNR})^{-1}\mathbf{V}[\mathbf{h}^{l,r}]^* \tag{7}$$

where $SNR = \frac{\sigma_s^2}{\sigma_n^2}$.

3 Proposed method

To compute the optimal filter we need to estimate the sources correlation matrix, $\mathbf{R}_{ss}$, which may be a challenging task. A more practical approach developed in this paper uses a model-based filter, by assuming that the sound field is represented by a simple model, composed of direct sound and a diffuse field. In other words, the problems is separated into two independent problems – one for a sound field composed of a single source (direct sound) and the other for a sound field

composed of a diffuse field. Combining the solutions for the two problems yields:

$$\mathbf{c}_{prop}^{r,l} = (\alpha \mathbf{v}\mathbf{v}^H + \beta \mathbf{V}\mathbf{V}^H + \gamma \mathbf{I})^{-1}(\alpha \mathbf{v}[h^{r,l}]^* + \beta \mathbf{V}[\mathbf{h}^{r,l}]^*) \quad (8)$$

where $\mathbf{v}$ is the steering vector between the source generating the direct sound field and the microphones, and $\mathbf{V}$ is the steering vector between uniformly distributed sources on a sphere and the microphones. These sources may represent additional direct sound sources and their reflections from the walls of a room, where relevant. Next, *DRR* is defined as the power ratio between the direct and the reverberant signals, and α, β, γ are normalization factors chosen as $\frac{1}{\|\mathbf{v}\|_2}, \frac{1}{DRR \cdot \|\mathbf{V}\|_2}, 10^{-5}$, respectively. Notice that if $DRR \gg 1$ the direct sound term becomes dominant, while when $DRR \ll 1$ the diffuse term becomes dominant and the solution converges to that of the BSM in equation (7).

Note that the solution in equation (8) requires the estimation of two parameters, i.e. the source direction of arrival (DOA) and the *DRR*. The latter, in particular, may not always be available. To estimate the DOA, it is possible to use the direct path dominant (DPD) algorithm that is based on the MUSIC method [19], or any other similar method. The estimation of *DRR* can be performed by a dedicated algorithm as in [20]. Alternatively, in this paper it is proposed to estimate *DRR* based on the effective rank of the correlation matrix of the microphone signals, $\mathbf{R}_{xx}$, where the effective rank of a matrix with real and positive eigenvalues is defined as [21]:

$$erank(\mathbf{A}) = \exp\left(H(p_1, p_2, ..., p_Q)\right) \quad (9)$$

where Q is the number eigenvalue of matrix. Here H is the Shannon entropy given by:

$$H(p_1, p_2, ..., p_Q) = -\sum_{i=1}^{Q} p_i log(p_i) \quad (10)$$

where p_i is the normalized eigenvalue of matrix $\mathbf{A}$ and defined as:

$$p_i = \frac{\lambda_i}{\sum_{n=1}^{Q} \lambda_n} \quad (11)$$

where λ is the eigenvalue of matrix $\mathbf{A}$. In this paper it is shown that $erank(\mathbf{R}_{xx})$ can be related to *DRR* for the data employed in the simulation study through a 4th order polynomial fit. The insight behind this useful relation is explained next. From inspection of Eq. (1), one can infer that, when ignoring noise, $\mathbf{V}\mathbf{R}_{ss}\mathbf{V}^H$ approximately equals $\mathbf{R}_{xx}$. Recall that $\mathbf{V}$ holds the DOA of the sound sources and $\mathbf{R}_{ss}$ is the correlation matrix between the sources. Therefore, $erank(\mathbf{R}_{xx})$ can be seen as a measure of the complexity (degrees of freedom or independent components) of the sound field. For example, when *DRR* is high the sound field behaves like a single source field and when *DRR* is low the sound field behaves like a reverberant or a multiple source field. The value of $erank(\mathbf{R}_{xx})$ is expected to behave similarly - when the sound field is composed of single source the rank is expected to be one, and for multiple uncorrelated sources it is expected to be full rank.

4 Simulation study

This section presents a simulation study of the proposed method for binaural reproduction.

4.1 Setup

This simulation study incorporates an array with $M = 6$ microphones that are uniformly distributed on a semi-circle in the horizontal plane, at elevation angles $\{\theta_m = \frac{\pi}{2}\}_{m=1}^{M}$ and azimuth angles $\{\varphi_m = \frac{\pi}{2} - \frac{(m-1)}{M-1}\}_{m=1}^{M}$, which are mounted on a rigid sphere with a radius of r = 10 cm. This is an example of an array which can be mounted on a wearable head device, e.g. AR glasses. The array was positioned in a room of size $6 \times 5 \times 3$ m at the location $[3 \times 3 \times 1.7]$m. A point source was also positioned in the room, producing a Gaussian white noise signal, such that the distance between the source and the array was selected to be - 1 m, 0.3 m and 0.1 m. The reverberation time of the room was designed to be $T60 = 0.37$ sec. The impulse response between the source and the microphones was simulated using the image method [22] and Matlab (version R2020a). Microphone signals were then computed by convolving the source signal with the room impulse responses. The sampling frequency of the signal recorded by the array was fs = 8 kHz. For the HRTFs, the Neumann KU100 manikin measurements from the Cologne database [23] were used.

4.2 Methodology

This section presents how the data was processed in order to compare between the performance of the optimal filter in equation (6), BSM filter in equation (7) and the proposed filter in equation (8). All filters were designed to be of length 50 msec. Selected HRTFs

were used, with directions corresponding to the steering vectors. This was performed by first calculating the HRTFs in the spherical harmonics domain up to an order of $N_h = 30$, and then applying the inverse spherical Fourier transform at the corresponding directions of the steering vector. Next, the optimal filter was calculated assuming σ_n^2 is equal to 0.1. The $\mathbf{R}_{ss}$ and $\mathbf{V}$ as follows, the steering vectors of the rigid array were calculated in the spherical harmonics domain, as described in [7] (section 4.2). $\mathbf{R}_{ss}$ was computed from

$$\mathbf{R}_{ss}(k) = \frac{1}{N}\sum_{m=1}^{N} \mathbf{s}(m,k)\mathbf{s}(m,k)^H \qquad (12)$$

where $\mathbf{s}(m,k)$ is the sources signals vector in the time - frequency domain, computed by applying the short-time Fourier transform (STFT) to $\mathbf{s}(t)$ with 50% overlapping windows of length 50 msec. N is an averaging parameter, while $\mathbf{s}(t)$ represents the microphone signals in the time domain. The BSM filter was calculated by assuming that $\mathbf{V}$ is a steering vector between 400 nearly-uniformly distributed sources (on a sphere) such that $\mathbf{V} \in R^{6\times 400}$ and the signal-to-noise ratio (SNR) is 10. The proposed filter was calculated by using the parameters presented in section 3 and $\mathbf{V}$ calculate like in the BSM filter, and by estimation the DRR. The error was calculated for the left binaural signal:

$$Error(k) = \sqrt{\frac{E(|p^l(k) - z^l(k)|^2)}{E(|p^l(k)|^2)}} \qquad (13)$$

where $z^l(k)$ is the estimated binaural signal given by equation (3). In order to investigate the performance of the filters under different acoustic environments, each simulation was repeated for diffrent values of direct-to-reverberant ratios (DRR) by changing the source - array distance as presented in section 4.1.

4.3 DRR estimation

The proposed method uses the DRR as parameter. As a DRR may not always be available, in this paper it is suggested that DRR is estimated from the effective rank of matrix $\mathbf{R}_{xx}$, as presented in Section 3. A room simulation was generated as described above. Then, the effective rank value was fitted to a $4th$ order polynomial, as we discussed in section 3, with the aim of approximating DRR. For each simulation and acoustic condition the variables (DRR, $erank(\mathbf{R}_{xx})$) were calculated. Then, $erank(\mathbf{R}_{xx})$ was averaged over the frequency range of [780 Hz, 3580 Hz], leading to a scalar value. Finally, the set (DRR, $E(erank(\mathbf{R}_{xx}))$) was fitted to 4th order polynomial. Figure 1 shows the estimation of the DRR under all conditions. The figure shows a relatively good fit. The estimated DRR values were employed in the study presented in the section 3, leading to relatively good performance of the proposed method. However, to validate this approach, it is suggested to be studied under a wider range of conditions

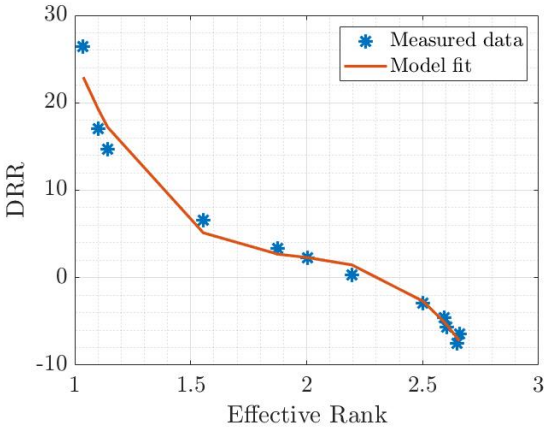

Fig. 1: DRR vs. effective rank: blue points represent the ground truth data and the orange plot is the 4th order polynomial fit.

4.4 Performance under various acoustic conditions

In this section the error defined in Eq. (13) was compared between the three filters under various acoustic environments. Figures $2-4$ present the error for three environments, differing by their DRR values, from $17\,dB$ in Fig. 1 to $-4\,dB$ in Fig. 3. The figures show that as DRR reduces and the sound field becomes more reverberant, the difference between the optimal filter and the BSM filter becomes smaller. This is expected, as the BSM assumes a diffuse field, or a large set of uncorrelated sources. The figures also show that the proposed method achieved errors that are very close to the optimal. This is particularly evident in Fig. 2, where significant improvement over the BSM method is observed. This is explained by the sound field model employed in the proposed method, specifically modeling the dominance of direct sound from the source.

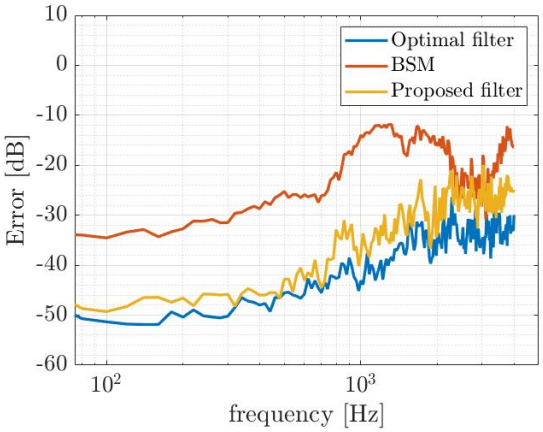

Fig. 2: The binaural signal error as in Eq. (13) computed for each filter for $DRR = 17dB$ with the distance between the microphone and white noise source position set to 0.1 m.

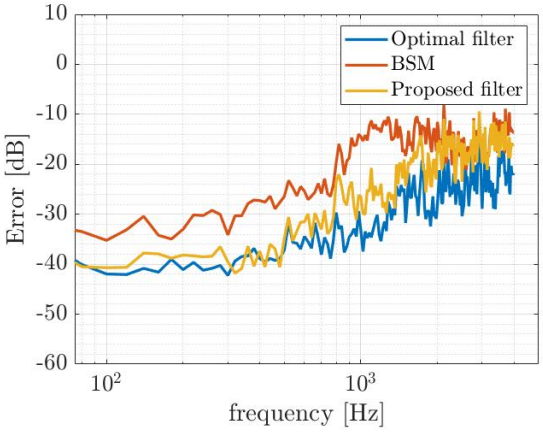

Fig. 3: The binaural signal error as in Eq. (13) computed for each filter for $DRR = 7.5dB$ with the distance between the microphone and white noise source position set to 0.3 m.

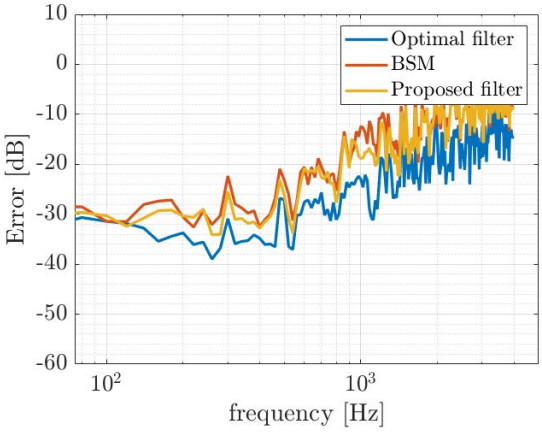

Fig. 4: The binaural signal error as in Eq. (13) computed for each filter for $DRR = -4dB$ with the distance between the microphone and white noise source position set to 1 m.

5 Conclusion

This paper studied a method for binaural reproduction with general arrays. The method is based on estimating the binaural signal directly from array measurements by assuming the sound field model is composed from direct and reverberant parts. A simulation study showed that the proposed method yields improved results over the BSM algorithm. As predicted by the theory, the improvement is more significant when the effective rank of the microphones correlation matrix is low, which occurs when DRR is high. In addition, as DRR decreases, the proposed filter converges to the BSM filter. The improvement in performance required information about parameters i.e. DRR and source DOA. These can be estimated from array measurements. The limitations of this study are that it considered only a single speaker in the room, Extension to multiple speakers, and validation with a listening test are proposed for future work.

References

[1] Politis, A., McCormack, L., and Pulkki, V., "Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing," in 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 379–383, IEEE, 2017.

[2] Noisternig, M., Sontacchi, A., Musil, T., and Holdrich, R., “A 3D ambisonic based binaural sound reproduction system,” in Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality, Audio Engineering Society, 2003.

[3] Ben-Hur, Z., Brinkmann, F., Sheaffer, J., Weinzierl, S., and Rafaely, B., “Spectral equalization in binaural signals represented by ordertruncated spherical harmonics,” The Journal of the Acoustical Society of America, 141(6), pp. 4087–4096, 2017.

[4] Rafaely, B. and Avni, A., “Interaural cross correlation in a sound field represented by spherical harmonics,” The Journal of the Acoustical Society of America, 127(2), pp. 823–828, 2010.

[5] Poletti, M. A., “Three-dimensional surround sound systems based on spherical harmonics,” Journal of the Audio Engineering Society, 53(11), pp. 1004–1025, 2005.

[6] Avni, A., Ahrens, J., Geier, M., Spors, S., Wierstorf, H., and Rafaely, B., “Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution,” The Journal of the Acoustical Society of America, 133(5), pp. 2711–2721, 2013.

[7] Rafaely, B., Fundamentals of spherical array processing, volume 8, Springer, 2015.

[8] Wabnitz, A., Epain, N., McEwan, A., and Jin, C., “Upscaling ambisonic sound scenes using compressed sensing techniques,” in 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 1–4, IEEE, 2011.

[9] Alon, D. L., Sheaffer, J., and Rafaely, B., “Robust plane-wave decomposition of spherical microphone array recordings for binaural sound reproduction,” The Journal of the Acoustical Society of America, 138(3), pp. 1925–1926, 2015.

[10] Abhayapala, T. D. and Ward, D. B., “Theory and design of high order sound field microphones using spherical microphone array,” in 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pp. II–1949, IEEE, 2002.

[11] Davis, L. S., Duraiswami, R., Grassi, E., Gumerov, N. A., Li, Z., and Zotkin, D. N., “High order spatial audio capture and its binaural headtracked playback over headphones with HRTF cues,” in Audio Engineering Society Convention 119, Audio Engineering Society, 2005.

[12] O’Donovan, A. M., Zotkin, D. N., and Duraiswami, R., “Spherical microphone array based immersive audio scene rendering,” International Community for Auditory Display, 2008.

[13] Song, W., Ellermeier, W., and Hald, J., “Using beamforming and binaural synthesis for the psychoacoustical evaluation of target sources in noise,” The Journal of the Acoustical Society of America, 123(2), pp. 910–924, 2008.

[14] Jiang, J., Xie, B., and Mai, H., “The Number of Virtual Loudspeakers and the Error for Spherical Microphone Array Recording and Binaural Rendering,” in Audio Engineering Society Conference: 2018 AES International Conference on Spatial Reproduction-Aesthetics and Science, Audio Engineering Society, 2018.

[15] Calamia, Davis, S., Smalt, C., and Weston, C. "A conformal, helmet-mounted microphone array for auditory situational awareness and hearing protection" 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 96–100.

[16] Ifergan, and Rafaely, B. "On the selection of the number of beamformers in beamforming-based

binaural reproduction" 2022 EURASIP Journal on Audio, Speech, and Music Processing, 2022(1), 6–6.

[17] L. Madmoni, J. Donley, V. Tourbabin, and B. Rafaely, “Beamforming-based binaural reproduction by matching of binaural signals,” presented at the Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality, 2020.

[18] H. Beit-On et al., "Audio Signal Processing for Telepresence Based on Wearable Array in Noisy and Dynamic Scenes," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8797-8801, doi: 10.1109/ICASSP43922.2022.9747583.

[19] Nadiri, and Rafaely, B. "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test". 2014 IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(10), 1494–1505.

[20] P. Samarasinghe, T. Abhayapala, H. Chen, P. Samarasinghe, T. Abhayapala, and H. Chen, “Estimating the Direct-to-Reverberant Energy Ratio Using a Spherical Harmonics-Based Spatial Correlation Model,” vol. 25, no. 2, pp. 310–319, 2017, doi: 10.1109/TASLP.2016.2633811.

[21] F. R. O’Keefe, J. A. Meachen, and P. D. Polly, “On Information Rank Deficiency in Phenotypic Covariance Matrices,” 2021, doi: 10.1093/sysbio/syab088.

[22] J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics. J Acoust Soc Am. 65(4), 943–50 (1979)

[23] Bernschütz, B., “A spherical far field HRIR/HRTF compilation of the Neumann KU 100,” in Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics, p. 29, AIA/DAGA, 2013.