Binaural Reproduction From Microphone Array Signals Incorporating Head-Tracking

Lior Madmoni
School of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer-Sheva, Israel
liomad@gmail.com

Jacob Donley
Facebook Reality Labs Research
Redmond, WA, USA
jdonley@fb.com

Vladimir Tourbabin
Facebook Reality Labs Research
Redmond, WA, USA
vtourbabin@fb.com

Boaz Rafaely
School of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer-Sheva, Israel
br@bgu.ac.il

***Abstract*—Binaural reproduction over headphones is becoming popular with applications in teleconferencing, virtual reality and entertainment. Many binaural reproduction methods are developed and studied extensively for spherical and other specially-designed arrays. However, some applications require simpler arrays, such as wearable or mobile. Recently, a binaural reproduction method was developed for such arrays based on binaural signals matching (BSM). While showing that BSM can produce accurate binaural signals, the study was limited to two array configurations and head-tracking was not incorporated. As a follow-up, this paper studies BSM with a varying number of microphones in a semi-circular configuration and investigates a solution for head-tracking. Simulations show that placing two microphones close to the ear positions of the head related transfer functions significantly improves reproduction with a static head. However, microphones placed away from ear positions are needed for head-tracking. This is then subjectively studied in a listening experiment with a semi-circular array, showing that compensating for head-rotations is important for accurate spatial perception but may introduce timbral artifacts due to the limited number of microphones.**

***Index Terms*—Binaural reproduction, Binaural signals matching, Semi-circular array, Head tracking.**

I. Introduction

Headphone reproduction of acoustic scenes that are captured by microphone arrays is becoming increasingly popular [1]–[3], with many applications in teleconferencing, virtual and augmented reality, and hearing research. One popular method is to convolve high order Ambisonics signals with the head-related transfer functions (HRTFs) [4]. This approach is particularly useful when the Ambisonics signal is estimated accurately for a sufficiently high spherical harmonics (SH) order. Another advantage is the ease of incorporating head-tracking, which enhances the immersion experience of the listener. However, accurate estimation of high order Ambisonics signals often requires special array geometries, such as spherical arrays, and a large number of microphones. This may limit the use of HOA with compact arrays, such as wearable arrays and arrays that are mounted on mobile devices.

This research was supported by Facebook Reality Labs Research.

An alternative approach is beamforming-based binaural reproduction (BFBR), which is more flexible in terms of the array design and may be suitable to compact arrays. With this approach, the microphone signals are spatially filtered with a set of beamformers, and the output signals are further filtered with HRTFs and then summed to reproduce the binaural signals. While BFBR is suitable for more flexible array geometries compared to HOA, recent works [5]–[7] mainly studied this approach with spherical arrays. In addition, a framework for appropriately setting BFBR design parameters, such as the beam number, look-directions, and beamformer weights, was not provided. Such a theoretical framework was recently presented in [8] for spherical arrays. However, for more general array geometries only limited guidelines were suggested, and so a comprehensive design methodology is still unavailable.

A third approach for binaural reproduction is based on binaural signals matching (BSM). This refers to the estimation of the desired binaural signals directly from the array measurements using filters for each ear. This approach was mainly studied in the context of sound field control using loudspeaker arrays [9]–[11]. BSM with headphone reproduction is less common, with a few publications including [7] that only presented limited results, and the parametric method in [12] that only studied a high resolution spherical microphone array. Recently, the design of a BSM system was described in [13] and studied for both spherical and semi-circular arrays. This study was limited to only two array configurations and while it was stated that this method supports head-tracking, its incorporation was not studied.

In this paper, the initial BSM study in [13] is extended to include a varying number of microphones in a semi-circular array and head-tracking. It is shown that the accuracy of BSM is more sensitive to the position of the microphones rather than their quantity. This leads to the study of head rotations of the listener, showing a clear advantage for using more microphones. The perceived spatial and timbral quality of BSM reproduction with head rotation are then subjectively studied with a listening experiment. Furthermore, it is argued that for a full head-tracked binaural audio reproduction over azimuth, a full-circular array may be necessary for an accurate reproduction.

II. Background

This section provides a background for topics that will be studied in this paper including mathematical models for binaural signals and array processing. Throughout the paper, the standard spherical coordinate system is used, denoted by (r, θ, ϕ), where r is the distance to the origin, θ is the elevation angle measured from the Cartesian z axis downwards to the Cartesian xy plane, and ϕ is the azimuth angle measured from the positive x axis towards the positive y axis.

Assume that a sound field is comprised of D sound sources, such that the d-th source carries the signal $s_d(k)$ with a distinct direction-of-arrival (DOA) (θ_d, ϕ_d) for $d = 1, \dots, D$. Here, $k = \frac{2\pi}{\lambda}$ is the wave-number and λ is the wave-length. Further assume that a listener is positioned with the center of the head coinciding with the origin. Let $h^{l,r}(k, \theta, \phi)$ denote the HRTFs of the left and right ears using the superscripts $(\cdot)^l$ and $(\cdot)^r$, respectively. Then, the complex amplitude of the sound pressure at the left and right ears, denoted $p^l(k)$ and $p^r(k)$, respectively, can be described as follows [14]:

$$p^{l,r}(k) = [\mathbf{h}^{l,r}(k)]^T \mathbf{s}(k), \tag{1}$$

where $\mathbf{h}^{l,r}(k) = \left[h^{l,r}(k, \theta_1, \phi_1), \dots, h^{l,r}(k, \theta_D, \phi_D)\right]^T$ are vectors of length D holding the HRTFs with the corresponding DOAs of the sources, for the left and right ears, $\mathbf{s}(k) = \left[s_1(k), \dots, s_D(k)\right]^T$ is a vector of length D holding the sources signals, and $(\cdot)^T$ is the transpose operator.

Next, assume that an array comprised of M microphones is positioned with its center at the origin, instead of the listener. The position of the m-th microphone is given by (r_m, θ_m, ϕ_m) for $m = 1, \dots, M$. In this case, the signal measured by the array microphones can be described by the following noisy signal model [15]:

$$\mathbf{x}(k) = \mathbf{V}(k)\mathbf{s}(k) + \mathbf{n}(k), \tag{2}$$

where $\mathbf{x}(k) = \left[x_1(k), \dots, x_M(k)\right]^T$ is a vector of length M composed of the pressure amplitude measured by each microphone, $\mathbf{V}(k)$ is the steering matrix of size $M \times D$ such that its d-th column holds the steering vector corresponding to the d-th source [15] for $d = 1, \dots, D$, and $\mathbf{n}(k)$ is a vector of length M holding the additive noise components.

III. Binaural Reproduction using Binaural Signals Matching

This section describes the BSM method for binaural reproduction presented in [13]. With BSM, the binaural signals in (1) are estimated directly from the array measurements model described in (2). At the first stage of the development of BSM, the array measurements are spatially filtered according to [13]:

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

where $\mathbf{c}^{l,r}(k) = \left[c_1^{l,r}(k), \dots, c_M^{l,r}(k)\right]^T$ are vectors of length M holding the filter coefficients for the left and right ears, and $(\cdot)^H$ is the Hermitian transpose operator. In [13] it was suggested to choose these coefficients such that the mean squared error (MSE) $\mathbb{E}[|p^{l,r}(k) - z^{l,r}(k)|^2]$ is minimized, where $\mathbb{E}[\cdot]$ denotes the expectation operator. Minimization of this error defines the objective of BSM - producing signals $z^{l,r}(k)$ that match the binaural signals. It was further assumed in [13] that the sources are uncorrelated with equal powers of σ_s^2 and that the noise is white with power of σ_n^2, which leads to the following solution:

$$\mathbf{c}_{\text{BSM}}^{l,r}(k) = (\mathbf{V}(k)\mathbf{V}^H(k) + \frac{\sigma_n^2}{\sigma_s^2}\mathbf{I}_M)^{-1}\mathbf{V}(k)[\mathbf{h}^{l,r}(k)]^*, \tag{4}$$

where $\mathbf{I}_M$ is the identity matrix of size M, and $(\cdot)^*$ is the complex conjugate operator. Finally, substituting (4) in (3) leads to the estimated binaural signals with BSM, denoted $\hat{p}^{l,r}(k)$:

$$\hat{p}^{l,r}(k) = [\mathbf{c}_{\text{BSM}}^{l,r}(k)]^H \mathbf{x}(k). \tag{5}$$

In order to analyze the accuracy of the BSM method, the following normalized error is defined [13]:

$$\text{err}_{\text{BSM}}^{l,r}(k) = \frac{\mathbb{E}[|p^{l,r}(k) - \hat{p}^{l,r}(k)|^2]}{\mathbb{E}[|p^{l,r}(k)|^2]}. \tag{6}$$

This error can be reformulated by substituting (1), (4) and (5) in (6):

$$\text{err}_{\text{BSM}}^{l,r}(k) = \frac{\sigma_s^2 \left\|\mathbf{V}^H(k)\mathbf{c}^{l,r}(k) - [\mathbf{h}^{l,r}(k)]^*\right\|^2 + \sigma_n^2 \left\|\mathbf{c}^{l,r}(k)\right\|^2}{\sigma_s^2 \left\|[\mathbf{h}^{l,r}(k)]^*\right\|^2}, \tag{7}$$

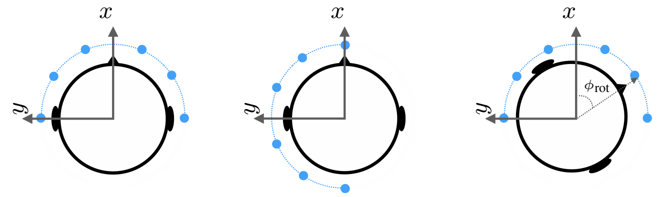
where $\|\cdot\|$ is the l^2-norm. As was explained in [13], the suggested BSM method is designed to accurately reproduce binaural signals due to sound fields that are comprised of D uncorrelated sources. However, when D is sufficiently large and the error in (7) is sufficiently small, the same BSM solution may also accurately reproduce more general sound fields with a different number of sources which may also be correlated. This is because (7) actually matches transfer functions $\mathbf{V}^H(k)$ to $[\mathbf{h}^{l,r}(k)]^*$, independent of the source signals. However, a more detailed analysis of the reproduction accuracy with other sound field types is left for future work.

IV. The Effect of Microphones Number and Their Positions on Performance

This section presents an analysis of the normalized error of the BSM binaural reproduction method, defined in (7), as a function of frequency and for different array configurations. For this purpose, a sound field comprised of $D = 240$ sources is assumed, with DOAs that correspond to a spirally-nearly uniform distribution on the sphere [16]. With this number of sources, the sound field can be relatively complex, and hence, accurate reproduction may suggest that the method is capable of accurately reproducing many other complex sound fields. The array studied here is comprised of M microphones distributed on a semi-circle in the horizontal plane, where $r_m = 10\,\text{cm}$, $\theta_m = \frac{\pi}{2}$ rad, and $\phi_m = \frac{\pi}{2} - \frac{\pi(m-1)}{M-1}$ rad for $m = 1, \ldots, M$, and which are mounted on a rigid sphere. These can represent for example a wearable array on a pair of glasses or a headset device such as those used in AR and VR applications. A schematic diagram of the head position and the array with $M = 6$ is presented in Fig. 1(a). The steering vectors for this array, required to calculate the filters $\mathbf{c}^{l,r}_{\text{BSM}}(k)$ in (4), were calculated analytically in the SH domain according to Section 4.2 in [17]. The HRTFs in (4) were taken from the Cologne database [18] measurements of the Neumann KU100 manikin, which has height$\times$width$\times$depth dimensions of approximately 28cm $\times$ 18cm $\times$ 22cm, and with a sampling frequency of 48 kHz. The HRTFs for the exact DOAs of the sources were interpolated in the SH domain up to an order of 30. Then, the filter coefficients $\mathbf{c}^{l,r}_{\text{BSM}}(k)$ in (4) were calculated assuming a signal-to-noise ratio (SNR) of 20 dB, for frequencies in the range $[75, 20000]$ Hz with a 75 Hz resolution. Finally, the normalized error was calculated according to (7).

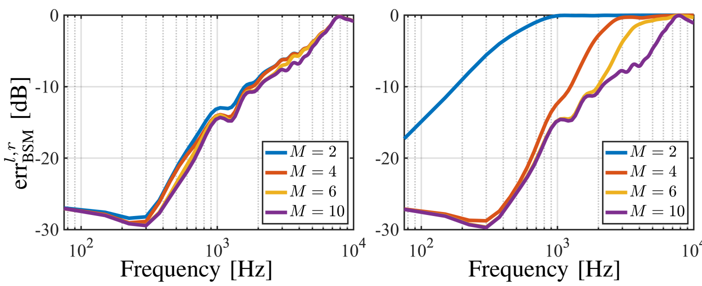
The normalized errors corresponding to the left ear signal, calculated according to (7), and with $M = 2, 4, 6$, and 10 microphones are presented in Fig. 2(a) as a function of frequency. First, note that in all cases the errors are relatively small at frequencies below 1.5 kHz. This means that the BSM reproduction of the left ear signal is expected to be relatively accurate at these frequencies. The right ear errors are very similar in this case and are not presented here for brevity. Further, note that the performance improvement when increasing the number of microphones from $M = 2$ is relatively minor. This behaviour may be explained by the geometry of this semi-circular array that is relatively similar to the dimensions of the KU100 manikin, and hence, at these low frequencies the effect of both bodies on arriving plane-waves may be very similar. In addition, for all configurations, there is a microphone which is very close to the position of the left ear of the KU100 manikin, which may further explain these low errors.

In order to study the affect of the proximity between the array microphones and ear positions on the accuracy of reproduction, each array configuration was rotated 90° anti-clockwise in azimuth relative to its center. A schematic diagram of this rotation is presented in Fig. 1(b) with the original head position. As can be seen, with this rotation the distance between the right ear and the microphones is now much larger compared to the array orientation prior to the rotation. The normalized errors of the left ear in this case are presented in Fig. 2(b). There is a performance degradation with the two-microphone rotated array compared to the original array orientation in Fig. 2(a). A similar but less significant degradation can also be seen for the four-microphone array. When six microphones are used with the rotated array, the error is relatively stable at the lower frequencies and with



(a) Original (b) Rotated array (c) Head rotation

Fig. 1. Schematic diagram of a head and the semi-circular array with $M = 6$ (blue dots). (a) Original orientation, (b) array rotated 90° anti-clockwise in azimuth, and (c) head rotation by ϕ_{rot} clockwise in azimuth.



(a) (b)

Fig. 2. The error $\text{err}^{l}_{\text{BSM}}$ in (7) as a function of frequency, and for varying microphone number M. (a) Left ear errors with the original array orientation, (b) left ear errors with the array rotated 90° anti-clockwise in azimuth.

ten microphones it is stable for the entire frequency range, compared to the original array orientation. The behaviour will significantly affect the accuracy of BSM binaural reproduction with head-tracking during playback, and hence, it is further studied next.

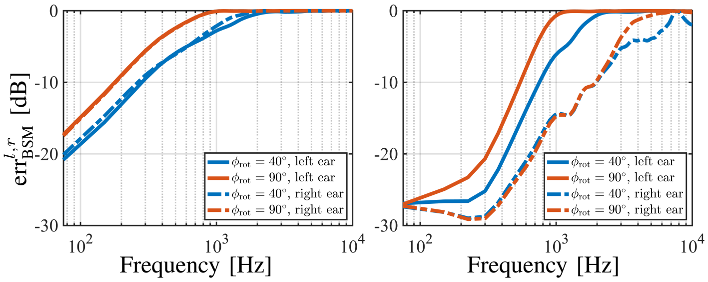
V. Binaural Signals Matching with Head-tracking

When it is desired to fix the acoustic environment relative to the listener's head position, compensation for head movements of the listener is required. In this paper, head movements are limited to head-rotations over azimuth only, for simplicity, and it is also assumed that the exact degree of head rotation is known at any moment via a head-tracking device. A schematic diagram of a head rotation by ϕ_{rot} degrees clockwise in azimuth is presented in Fig. 1(c) with the original array orientation. In order to compensate for this head-rotation, the entries of the HRTFs vectors $\mathbf{h}^{l,r}(k)$, which are used for the calculation of the filters $\mathbf{c}^{l,r}_{\text{BSM}}(k)$ in (4), should be rotated in the opposite direction. Thus, a head rotation by ϕ_{rot} degrees clockwise in azimuth is compensated by the filters by replacing the HRTFs $\mathbf{h}^{l,r}(k)$ in (4) by their following rotated version:

$$\mathbf{h}^{l,r}_{\text{rot}}(k) = \left[h^{l,r}(k, \theta_1, \phi_1 + \phi_{\text{rot}}), \ldots, h^{l,r}(k, \theta_D, \phi_D + \phi_{\text{rot}})\right]^T. \tag{8}$$

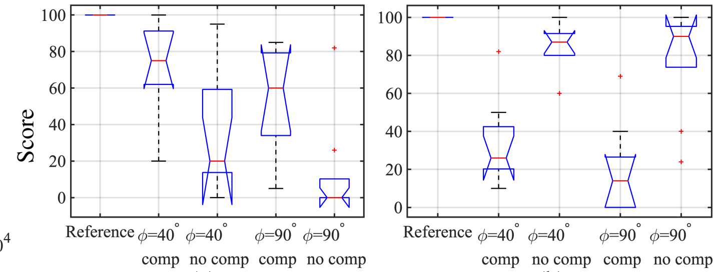
Note that (8) is suitable for azimuth rotations only, and for rotations with more degrees-of-freedom, the compensation should be set appropriately.

The estimation errors for compensating anti-clockwise head rotation of $\phi_{\text{rot}} = 40°$ and $\phi_{\text{rot}} = 90°$ are presented in Fig. 3. These are presented for both ears signals in order



(a) (b)

Fig. 3. The BSM reproduction error $\mathrm{err}^{l}_{\mathrm{BSM}}$ in (7) as a function of frequency, and for anti-clockwise head rotation of $\phi_{\mathrm{rot}} = 40°, 90°$. These correspond to the semi-circular array with (a) $M = 2$, and (b) $M = 6$ microphones.



(a) (b)

Fig. 4. Box-plots of the listening experiment results, with medians marked by the horizontal red line, outliers marked by red plus signs, 25th and 75th percentiles are marked by the bottom and top colored edges, respectively, and the minimal and maximal grades marked by the black lines. The scores are based on (a) source location, (b) timbre variation.

to study the accuracy of binaural reproduction with BSM and head-tracking. Fig. 3(a) shows that for $M = 2$ the errors are relatively large, since the rotation has increased the distance between the microphones and ear positions. It can be concluded that this array is not capable of reproducing accurate binaural signals with head-tracking. Fig. 3(b) shows that for $M = 6$ and for both head rotations, the right ear signal is estimated with very similar errors that are relatively low below approximately $1.5\,\mathrm{kHz}$. However, these head rotations significantly increase the distance between the left ear and the microphone positions, and hence the left ear signal errors are much larger. This decreases the frequency range at which accurate reproduction is possible. Because of this behaviour, degradation is expected for many head rotations. This may be improved by using a full-circular array geometry and a sufficient number of microphones. However, the study of such an array is outside the scope of this paper and is suggested for future work.

VI. Listening Experiment

In order to further understand the reproduction quality of BSM with these head rotations, a listening experiment was conducted. For this purpose, a point source was simulated inside a room of dimensions $15.5 \times 9.8 \times 7.5\,\mathrm{m}$ and with reverberation time of $T_{60} = 0.6\,\mathrm{s}$ using the image method [19]. The source was located at $(8.5, 6.3, 1.7)\,\mathrm{m}$ inside the room. A five seconds long segment of female speech was used as the source signal, taken from the TIMIT database [20], and upsampled to $48\,\mathrm{kHz}$ to match the sampling rate of the HRTFs. This signal was used in the calculation of the pressure measured by the previously described semi-circular array with $M = 6$ microphones and with its center located at $(5, 5, 1.7\,)\mathrm{m}$. For each degree of head rotation ($40°$ and $90°$ in this experiment), the binaural signals were calculated according to (5) with the rotation-compensated BSM filters that include $\mathbf{h}^{l,r}_{\mathrm{rot}}(k)$ described in (8) and with the non-compensated BSM filters. This corresponds to four sets of test signals.

The experiment was based on the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [21], but without an anchor signal, since such a signal is not clear in this case. Two MUSHRA screens were generated with the same four test signals and a reference HOA signal of order $N = 12$. In each screen, the scoring criterion for evaluating the similarity between the test signals and the reference signal was different. The first is based on the perceived location of the source, and the second is based on timbral variation. Eight subjects with no known hearing impairments participated in the experiment and were instructed to rate the similarity of the test signals to the reference on a scale of 0-100 for each criterion separately. Prior to the listening test, a training stage was performed with a single screen, introducing the signals to the subjects as a familiarization stage.

The scores given by the subjects are presented using box-plots in Fig. 4. The scores corresponding to the perceived location of the source are presented in Fig. 4(a), where it can be seen that the binaural signals corresponding to the compensated BSM reproduction of both head rotation values received generally much higher scores than the non-compensated reproduction. The scores based on timbral variation are presented in Fig. 4(b), where now the compensated BSM reproduction received relatively low scores, which may be due to the large errors at the right ears described previously. However, the non-compensated BSM reproduction received relatively high scores in this criterion. These results highlight the trade-off between the perceived location and timbral variation when compensating for head rotations with BSM reproduction, which may be utilized in the design of such a system.

VII. Conclusions

In this work, binaural reproduction with BSM was studied with a semi-circular array in varying configurations. It was shown that the method can produce accurate binaural signal at a limited frequency range with only two microphones close to the ears of the listener. However, the performance degrades significantly as the distance between the microphones and the ear positions in the HRTFs increases, even with more microphones. For head-tracking applications, increasing the number of microphones with a semi-circular array improves performance. For full head-tracking over azimuth a full-circular array may be required. Subjective evaluation showed that with a semi-circular array, compensating for head rotation may be useful for spatial perception and source localization with a trade-off introducing timbral artifacts. It is suggested that future work studies BSM with other array geometries, develops a design framework that guarantees accurate binaural reproduction and extends the listening experiment performed in this work.

References

[1] Archontis Politis, Leo McCormack, and Ville Pulkki, "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)*. IEEE, 2017, pp. 379–383.

[2] Markus Noisternig, Alois Sontacchi, Thomas Musil, and Robert Holdrich, "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] Zamir Ben-Hur, Fabian Brinkmann, Jonathan Sheaffer, Stefan Weinzierl, and Boaz Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017.

[4] Michael A Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society*, vol. 33, no. 11, pp. 859–871, 1985.

[5] Adam M O'Donovan, Dmitry N Zotkin, and Ramani Duraiswami, "Spherical microphone array based immersive audio scene rendering," International Community for Auditory Display, 2008.

[6] Wookeun Song, Wolfgang Ellermeier, and Jørgen Hald, "Using beamforming and binaural synthesis for the psychoacoustical evaluation of target sources in noise," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 910–924, 2008.

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

[8] Itay Ifergan, "Theoretical framework for beamformer distribution in beamforming based binaural reproduction," *MS.c. Thesis, School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Submitted*, 2020.

[9] Ji-Ho Chang, Chan-Hui Lee, Jin-Young Park, and Yang-Hann Kim, "A realization of sound focused personal audio system using acoustic contrast control," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2091–2097, 2009.

[10] P-A Gauthier, C Camier, Y Pasco, A Berry, E Chambatte, R Lapointe, and M-A Delalay, "Beamforming regularization matrix and inverse problems applied to sound field measurement and extrapolation using microphone array," *Journal of Sound and Vibration*, vol. 330, no. 24, pp. 5852–5877, 2011.

[11] Ferdinando Olivieri, Filippo Maria Fazi, Mincheol Shin, and Philip Nelson, "Pressure-matching beamforming method for loudspeaker arrays with frequency dependent selection of control points," in *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[12] Symeon Delikaris-Manias, Juha Vilkamo, and Ville Pulkki, "Parametric binaural rendering utilizing compact microphone arrays," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 629–633.

[13] Lior Madmoni, Jacob Donley, Vladimir Tourbabin, and Boaz Rafaely, "Beamforming-based binaural reproduction by matching of binaural signals," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020.

[14] Dylan Menzies and Marwan Al-Akaidi, "Nearfield binaural synthesis and ambisonics," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1559–1563, 2007.

[15] Harry L Van Trees, *Optimum array processing: Part IV of detection, estimation, and modulation theory*, John Wiley & Sons, 2004.

[16] Edward B Saff and Amo BJ Kuijlaars, "Distributing many points on a sphere," *The mathematical intelligencer*, vol. 19, no. 1, pp. 5–11, 1997.

[17] Boaz Rafaely, *Fundamentals of spherical array processing*, vol. 8, Springer, 2015.

[18] Benjamin Bernschütz, "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*. AIA/DAGA, 2013, p. 29.

[19] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[20] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, Nancy L Dahlgren, and Victor Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, vol. 10, no. 5, pp. 0, 1993.

[21] ITU-Recommendation, "Method for the subjective assessment of intermediate quality level of coding systems," *ITU-R BS*, pp. 1534–1, 2003.