3D DOA ESTIMATION OF MULTIPLE SOUND SOURCES BASED ON SPATIALLY CONSTRAINED BEAMFORMING DRIVEN BY INTENSITY VECTORS

Despoina Pavlidi [*†], *Symeon Delikaris-Manias* [‡], *Ville Pulkki* [‡], *Athanasios Mouchtaris* [*†]

[*] FORTH-ICS, Heraklion, Crete, GR-70013, Greece
[†] University of Crete, Department of Computer Science, Heraklion, Crete, GR-70013, Greece
[‡] Aalto University, Department of Signal Processing and Acoustics, Espoo, FI-00076, Finland

ABSTRACT

Sound source localization in three dimensions with microphone arrays is an active field of research, applicable in sound enhancement, source separation, and sound field analysis. In this contribution we propose a method for three dimensional multiple sound source localization in reverberant environments. We employ a spatially constrained steered response beamformer on a spherical sector centered at the direction of arrival (DOA) estimates of the intensity vector. Experiments are performed in both simulated and real acoustical environments with a spherical microphone array for multiple sound sources under different reverberation and signal-to-noise ratio (SNR) conditions. The performance of the proposed method is compared with our previously proposed work and a subspace method in the spherical harmonic domain. The results demonstrate a significant improvement in terms of localization accuracy.

Index Terms— direction of arrival, 3D, multiple sources, spherical microphone array processing, sound intensity

1. INTRODUCTION

In direction of arrival (DOA) estimation a wide selection of algorithms is available such as subspace [1], intensity-based [2] and power spectrum methods [3], each of them with different level of complexity. The choice of the algorithm depends on the requirements of the application: the tolerable latency and the required accuracy. Subspace methods such as multiple signal classification (MUSIC) can provide accurate DOA estimates and have been extended in three dimensions for spherical microphone arrays [1]. However, an exhaustive search is required and therefore they can be computationally inefficient for real-time applications. Further improvements in the MUSIC algorithm have been proposed using the direct-path dominance test for highly reverberant environments [4].

Intensity-based methods utilize a pressure and a particle velocity component to analyze the sound field. In practice, the pressure and particle velocity are estimated with an omnidirectional and three dipole microphones respectively [5]. Due to its tolerable latency, the intensity vector is an ideal candidate for real-time DOA estimation and has been previously employed in time-frequency domain spatial sound processing [6]. Its performance has been examined in reverberant environments [7] and a pseudo intensity vector has been formulated in the spherical harmonic domain [2, 8].

In this contribution, an improvement of the DOA estimates using the intensity vector is proposed based on post-processing the short-time estimates with spatially constrained beamforming. We build upon our recent results for 3D DOA estimation using single source zones (SSZs) and the sound intensity vector [9] and demonstrate an improvement in accuracy of DOA estimation. The paper is organized as follows. In Section 2 the background on spherical microphone array processing is presented briefly. Section 3 describes the proposed method of processing the DOA estimates with the use of spherical harmonic domain regular beamformers. Section 4 presents the experimental setup for evaluation and the results using a simulated and a real microphone array in reverberant environments with the presence of multiple speech sources. Section 5 presents our conclusions.

2. SPHERICAL ARRAY PROCESSING

An overview of the process of how to spatially encode the microphone array signals to a set of spherical harmonic signals is presented. For an extended overview of spherical microphone array decomposition and beamforming, the reader is referred to [10–12]. Spatial encoding refers to the process of approximating the spherical harmonic signals, denoted as s_{lm} for order l and degree m, from microphone signals x_q of a microphone array with radius r and microphone positions $\Omega_q = (\theta_q, \phi_q)$. $\Omega = (\theta, \phi)$ denotes the elevation θ and the azimuth ϕ with $\theta \in [-\pi/2, \pi/2]$ and $\phi \in [-\pi, \pi]$. For a microphone array comprised by Q microphones, where the pressure is obtained at discrete points Ω_q, the spherical harmonic coefficients can be approximated by $s_{lm}(k, r) \approx \sum_{q=1}^{Q} g_q(\Omega_q) x_q(k, r, \Omega_q) Y_{lm}^*(\Omega_q)$,

This research has been partly funded by the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 644283.

where $x_q(k, r, \Omega_q)$ are the separate microphone signals for frequency k, $Y^*_{lm}(\Omega_q)$ are the complex conjugate spherical harmonic functions and $g_q(\Omega_q)$ is selected so that it provides an accurate approximation of the spherical Fourier transform [13, 14]. The accuracy of this approximation depends on how uniformly the microphones are distributed on the surface of the sphere, the type of the array, the radius r and the frequency k [15]. The number of microphones Q defines the highest order L of spherical harmonic signals that can be obtained. For L^{th} order of independent harmonics, the number of microphones to reconstruct L harmonic signals is $Q \geq (L+1)^2$ [11]. For a uniform spherical arrangement of microphones ($g_q = 4\pi/Q$), the equalized spherical harmonic signals can be expressed in matrix form as

$$\mathbf{s} \approx \frac{4\pi}{Q}\mathbf{B}^{-1}\mathbf{Y}^H\mathbf{x}, \tag{1}$$

where $(^H)$ denotes Hermitian transposition, $\mathbf{x} \in \mathbb{C}^{Q\times 1}$ are the microphone array signals,

$$\mathbf{s} = [s_{00}, s_{1-1}, s_{10}, \ldots, s_{LL-1}, s_{LL}]^T \in \mathbb{C}^{(L+1)^2\times 1}, \tag{2}$$

$$\mathbf{B} = \text{diag}\{[b_0, b_1, b_1, b_1, b_2, \ldots, b_L]\} \in \mathbb{C}^{(L+1)^2\times(L+1)^2} \tag{3}$$

and $\mathbf{Y} \in \mathbb{C}^{Q\times(L+1)^2}$ is the matrix containing the spherical harmonics up to order L for the Q microphones [13]. $\mathbf{B}$ depends on the array type, whether it is rigid or open, and is used in Eq. (1) to remove the effect of the microphone array [14].

3. PROPOSED METHOD

In this section we present the proposed method which combines sound intensity vector estimates with spatially constrained beamforming in order to improve the DOA estimation of multiple sound sources. We follow the framework of our previously proposed work [9], which we will briefly describe for completeness and clarity.

3.1. DOA estimation in single-source zones with intensity vector estimates

The sound intensity vector points to the direction of sound energy flow, thus its estimation provides the DOA of the source generating the energy flow as the vector pointing to the opposite direction. The instantaneous active intensity vector can be approximated in the time-frequency (TF) domain with n being the time index as in [8]

$$\mathbf{I}(k,n) = \frac{1}{2}\Re\left\{ \begin{bmatrix} \frac{s^*_{00}(k,n)}{b_0(k)} \end{bmatrix} \begin{bmatrix} s_x(k,n) \\ s_y(k,n) \\ s_z(k,n) \end{bmatrix} \right\}, \tag{4}$$

where s_x, s_y, s_z are averages of 1^{st} order steered eigenbeams with the negative phase towards the x, y and z-axis respectively, calculated as

$$s_\alpha(k,n) = \sum_{m=-1}^{1} Y_{1m}(\Omega_\alpha)s_{1m}(k,n), \quad \alpha = \{x, y, z\}, \tag{5}$$

where $s_{1m}(k,n)$ is defined in Section 2 and Ω_α is $(0,\pi)$, $(0,-\pi/2)$, and $(-\pi/2, 0)$ for each axis.

In [9] we proposed the estimation of $\mathbf{I}(k,n)$ and consequently of the DOA in TF points of SSZs, i.e., series of K frequency-adjacent TF points where only one source dominates, adopting a relaxed sparsity assumption of the sources in the TF domain. In this manner we avoid obscure areas of the TF spectrum where more than one sources are simultaneously active. The SSZs are selected as those areas of the TF spectrum that exhibit a mean correlation coefficient higher than a predefined threshold. The correlation coefficient is defined as

$$\rho_{i,j}(K,n) = \frac{R_{i,j}(K,n)}{\sqrt{R_{i,i}(K,n)\cdot R_{j,j}(K,n)}}, \tag{6}$$

where $R_{i,j}(K,n) = \sum_{k\in K} |X_i(k,n)\cdot X_j(k,n)|$ is the cross-correlation of the magnitude of the TF transform over an analysis zone for any pair of signals (x_i, x_j). $X_i(k,n)$ and $X_j(k,n)$ are the signals of the i^{th} and the j^{th} microphones respectively in the TF domain. Note that $x_q(k, r, \Omega_q)$ in Section 2 is now expressed in the TF domain as $X_q(k,n)$ for the q^{th} microphone by omitting the (r, Ω_q) parameters.

3.2. From intensity vector estimates to spatially constrained beamforming

DOA estimation through sound intensity vector possesses low computational complexity, since it can provide instantaneous time-frequency estimates. However, by definition, the intensity vector estimation exploits the spherical harmonic analysis of the sound field up to the first order even though the available microphone array may provide higher spherical harmonic orders. On the other hand, DOA estimation relying on steered-response beamforming, even though it can exhibit high accuracy and exploits the full potential of the recording device, it suffers from high computational complexity due to the exhaustive search of the 3D space. These two different approaches motivated this work to propose a hybrid methodology that takes advantage of the simplicity of the intensity vector estimation and the accuracy of beamforming in order to lead to enhanced DOA estimation compared to [9].

Assume that, although $-\mathbf{I}(k,n)$ might not point exactly to the DOA of a source, it will point towards the "neighbourhood" of a true source, i.e., it aims near the true direction. We call this a coarse DOA estimation, $\Omega_c = (\theta_c, \phi_c)$. We could then beamform around the area where $-\mathbf{I}(k,n)$ is pointing, i.e., perform spatially constrained beamforming (SCB), and thus obtain a refined DOA estimation. The beamforming is performed over the spherical sector defined by $-\mathbf{I}(k,n)$ and a vector of angle distance equal to γ_b from $-\mathbf{I}(k,n)$ (see also Fig. 1). The DOA, $\Omega_f = (\theta_f, \phi_f)$, is then estimated as the index where the power of the SCB gets maximized, i.e.,

$$\Omega_f = \arg\max_{\Omega_s} |p(k,n,\Omega_s)|^2, \tag{7}$$

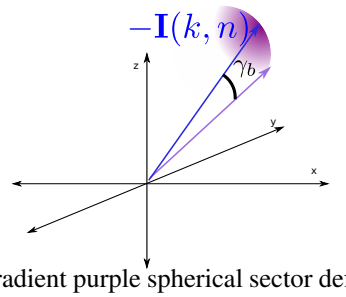

Fig. 1. The gradient purple spherical sector defines the beamforming area

where Ω_s belongs to the set of points in the spherical sector to be scanned, and $p(k,n,\Omega_s)$ is the beamformer's output for a regular beampattern steered at direction Ω_s [1] with $p(k,n,\Omega_s)$ given by

$$p(k,n,\Omega_s) = \mathbf{y}^T(\Omega_s)\mathbf{s}, \tag{8}$$

where $\mathbf{y}(\Omega_s) = [Y_{00}(\Omega_s), Y_{1-1}(\Omega_s), \ldots, Y_{LL}(\Omega_s)]^T \in \mathbb{C}^{(L+1)^2\times 1}$. Once we have estimated all the refined DOAs in the SSZs we form a 2D histogram from the set of estimations in a block of N consecutive time frames. This constant size block slides one frame each time. We process the 2D histogram as in [9] in order to extract the final DOA estimates. We first smooth the 2D histogram by applying a circularly symmetric Gaussian window $\mathbf{w}_{\mathrm{A}}(\theta,\phi)$ of zero mean and standard deviation (std) equal to σ_{A}, leading to

$$\mathbf{h}_{\mathrm{s}}(\theta,\phi) = \sum_i \sum_j \mathbf{h}(i,j)\mathbf{w}_{\mathrm{A}}(\theta-i,\phi-j), \tag{9}$$

where $\mathbf{w}(\theta,\phi) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{\theta^2+\phi^2}{\sigma^2}}$ is the Gaussian window, $\mathbf{h}(\theta,\phi)$ is the original 2D histogram and $\mathbf{h}_{\mathrm{s}}(\theta,\phi)$ is the smoothed one. We then iteratively detect the highest peak of the smoothed histogram $\mathbf{h}_s^g(\theta,\phi)$, identify its index as the DOA of a source, $(\theta_g,\phi_g) = \arg\max_{\theta,\phi} \mathbf{h}_s^g(\theta,\phi)$ and remove its contribution from the histogram, $\delta_g = \mathbf{h}_s(\theta,\phi) \odot \mathbf{w}_{\mathrm{C}}(\theta-\theta_g,\phi-\phi_g)$ by applying a second Gaussian window $\mathbf{w}_{\mathrm{C}}(\theta,\phi)$ of zero mean and std equal to σ_{C} until we reach the number G of sources. Thus the smoothed histogram at each next iteration would be $\mathbf{h}_s^{g+1}(\theta,\phi) = \mathbf{h}_s^g(\theta,\phi) - \delta_g$. The core steps of our method are summarized as follows:

1. Encode the microphone signals to the spherical harmonic domain.
2. Detect all the SSZs.
3. Estimate $\mathbf{I}(k,n)$ so as to obtain coarse DOA estimates in the SSZs.
4. Beamform in neighborhoods of intensity-based DOA estimates for refined DOA estimation in the SSZs.
5. Generate and smooth the 2D histogram of a block of refined DOA estimates.
6. Process the smoothed 2D histogram to extract the final 3D DOA estimates.

4. EVALUATION

The performance of the proposed method is investigated by extended simulations and real measurements in reverberant environments. A rigid spherical microphone array is utilized with radius equal to $r = 0.042$ m, comprising 32 microphones, placed at the center of the faces of a truncated icosahedron. For the simulations we used the room impulse response (RIR) generator by Jarrett et al [16] which is based on the image method of Allen and Berkley [17] to simulate a room of $5.6 \times 6.3 \times 2.7$ m^3, having the same dimensions as the room where we conducted the real experiments. The spherical array was placed in the center of the room, and the sound sources were placed 1 m away from the center of the array. The sampling frequency was equal to 48 kHz and the time frame and FFT size was 2048 samples. We applied 50% overlapping in time and K was equal to 375 Hz for the detection of SSZs with the mean correlation coefficient threshold set to 0.8. The angle for the SCB was set to $\gamma_b = 10^\circ$ and the beamformer's order was $L = 3$. The windows used at the histogram processing had std equal to $\sigma_{\mathrm{A}} = 5^\circ$ and $\sigma_{\mathrm{C}} = 20^\circ$. The speed of sound was $c = 343$ m/s while the frequency range used was 500-3800 Hz to avoid aliasing phenomena [18]. The minimum separation between the sources was 20°. The sources had equal power and the SNR at each microphone was estimated as the ratio of the power of each source signal to the power of the noise signal. The performance of the proposed algorithm is demonstrated by the mean estimation error (MEE) which measures the angular distance between a unit vector pointing at the true DOA ($\mathbf{v}$) and a unit vector pointing at the estimated DOA ($\hat{\mathbf{v}}$) [8] over all sound sources and frames of the source signals. The error is defined as

$$\mathrm{MEE} = \frac{1}{N_{\mathrm{F}} G} \sum_n \sum_g \cos^{-1}\left(\mathbf{v}_{ng}^T \hat{\mathbf{v}}_{ng}\right), \tag{10}$$

where $\cos^{-1}\left(\mathbf{v}_{ng}^T \hat{\mathbf{v}}_{ng}\right)$ expresses the angular distance between the true DOA of the g^{th} active source in the n^{th} frame and the estimated one. The association between the true and the estimated DOA of a source is based on the permutation

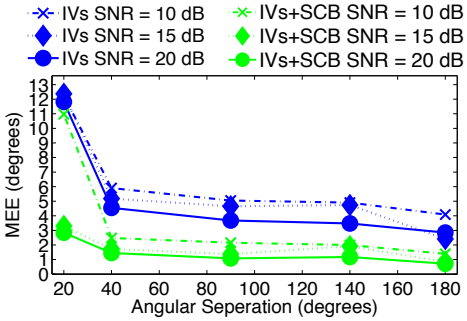

Fig. 2. MEE versus angular separation between 2 sound sources for $\mathrm{RT}_{60} = 0.4$ sec and various SNR conditions.

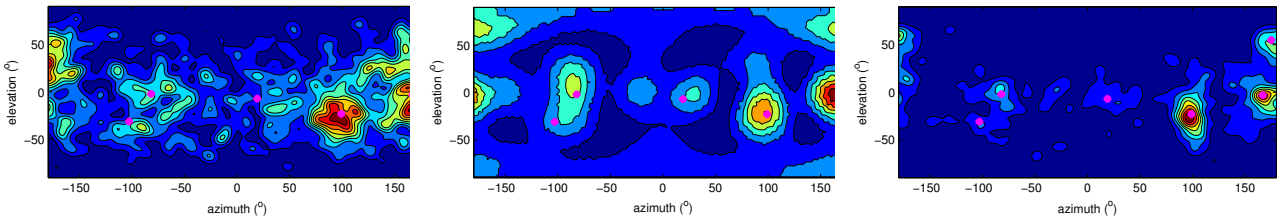

Fig. 3. 2D histogram for six sound sources with the intensity vector (left), the corresponding pseudospectrum for the MUSIC method with direct-path dominance test (middle) and the 2D histogram for intensity vector + SCB (right).

that leads to the minimum error. N_{F} is the total number of frames after subtracting $N-1$ frames of the initialization period and G is the number of active sources, assumed to be known. In all the simulations speech files were used of duration approximately 7 seconds, leading to $N_{\mathrm{F}} = 282$ frames. Any gaps or silent periods were removed. The block size is equal to 1 second, i.e., $N = 46$ frames, which was found to be a good compromise between the accuracy and the responsiveness of the algorithm.

4.1. Results with simulated room impulse responses

In our first set of simulations we investigate the performance of our proposed method (denoted as "IVs+SCB") for several angular distances between two continuously active sources for SNR=$\{10, 15, 20\}$ dB and $\mathrm{RT}_{60} = 0.4$ s in comparison with our previously proposed method of [9] (denoted as "IVs"). We show the results in Fig. 2. In all examined cases, IVs+SCB exhibits better performance than the previous one for all SNR conditions and angular separations. In our second set of simulations we compare the performance of the proposed work with the MUSIC algorithm as implemented in [4] and denoted as "DPD-MUSIC". In Fig. 3 the 2D histograms for the IVs and the IVs+SCB methods and the pseudospectrum of the DPD-MUSIC are shown for a case of 6 simultaneous speech sources in a simulated reverberant environment with $\mathrm{RT}_{60} = 0.6$ sec. The pink marks denote the true position of the sources. The processing of these representations of 2D estimates is based on one second history for all three methods assuming a known number of sources and follows the steps described in Section 3.2. Results in different acoustical conditions are shown in Fig. 4, for scenarios involving one, three and six simultaneously active speakers in highly reverberant conditions of $\mathrm{RT}_{60} = 0.6$ s and SNR=$\{10, 15, 20\}$ dB. DPD-MUSIC and IVs demonstrate similar performance while IVs+SCB exhibits a clear advantage especially for higher signal-to-noise ratios.

4.2. Results with real room impulse responses

For the conduction of real experiments we recorded RIRs using the em32 EigenMike [19] in a reverberant room of the same dimensions and reverberation as in the simulations. We show our results in Fig. 5 at the left plot, while at the right we plot a simulated counterpart. Our proposed method shows high accuracy for medium and higher SNR conditions even

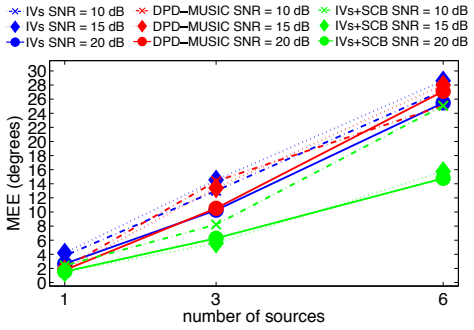

Fig. 4. MEE versus number of sources for $\mathrm{RT}_{60} = 0.6$ sec and various SNR conditions.

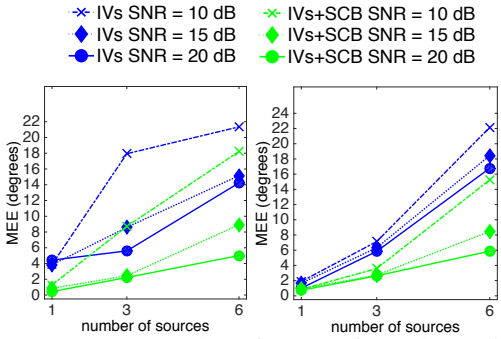

Fig. 5. MEE versus number of sources for real RIRs in a room of $\mathrm{RT}_{60} = 0.3$ sec and various SNR conditions (left) and its simulated counterpart (right)

with six simultaneously active sources. For lower SNR and as the number of sources increases the performance degrades as expected, following similar tendency between the simulated and real results.

5. CONCLUSION

We presented a method which significantly improves the accuracy of DOA estimation of multiple sound sources in the 3D space. Our method utilizes intensity vector estimates to trigger beamforming in a spatially constrained sector, leading to an efficient hybrid algorithm. The evaluation was conducted in simulated reverberation and SNR conditions and with real RIRs for various number of sources and the method was compared to our previous approach and the state-of-the-art, exhibiting superior performance.

6. REFERENCES

[1] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication*, pp. 281–305. Springer, 2010.

[2] C. Evers, A. H Moore, and P. A Naylor, "Multiple source localisation in the spherical harmonic domain," in *14th IEEE IWAENC*, 2014, pp. 258–262.

[3] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 117–120.

[4] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[5] S. Tervo, "Direction estimation based on sound intensity vectors," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2009, pp. 700–704.

[6] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.

[7] D. Levin, E. A. P. Habets, and S. Gannot, "On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1800–1811, 2010.

[8] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2010, pp. 442–446.

[9] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, and A. Mouchtaris, "3D localization of multiple sound sources with intensity vector estimates in single source zones," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1556–1560.

[10] B. Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8, Springer, 2015.

[11] T. D. Abhayapala, "Generalized framework for spherical microphone arrays: Spatial and frequency decomposition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2008, pp. 5268–5271.

[12] J. Meyer and G. Elko, "Spherical harmonic modal beamforming for an augmented circular microphone array," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2008, pp. 5280–5283.

[13] E. G. Williams, *Fourier acoustics: sound radiation and nearfield acoustical holography*, Acad. press, 1999.

[14] H. Teutsch, *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, vol. 348, Springer Science & Business Media, 2007.

[15] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 135–143, 2005.

[16] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: Algorithm and applications," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1462–1472, 2012.

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating smallroom acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[18] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1494–1505, 2014.

[19] mh acoustics, "EM32 eigenmike microphone array release notes (v17. 0)," Tech. Rep., mh acoustics, Summit, NJ, USA, 2013.