3D LOCALIZATION OF MULTIPLE SOUND SOURCES WITH INTENSITY VECTOR ESTIMATES IN SINGLE SOURCE ZONES

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

This work proposes a novel method for 3D direction of arrival (DOA) estimation based on the sound intensity vector estimation, via the encoding of the signals of a spherical microphone array from the space domain to the spherical harmonic domain. The sound intensity vector is estimated on detected single source zones (SSZs), where one source is dominant. A smoothed 2D histogram of these estimates reveals the DOA of the present sources and through an iterative process, accurate 3D DOA information can be obtained. The performance of the proposed method is demonstrated through simulations in various signal-to-noise ratio and reverberation conditions.

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

1. INTRODUCTION

Direction of arrival estimation (DOA) for multiple sound sources using microphone arrays has gained the interest of the research community over the last decades. State of the art examples of such algorithms are the multiple signal classification and the estimation of signal parameters via rotational invariance techniques [1]. Such DOA estimators have been widely used and formulated for different types of microphone arrays ranging from linear to arbitrary configurations. 3D DOA estimation plays an important role in a variety of applications, namely speech enhancement, smart home automation and source separation. Accurate estimation of the 3D position of a sound source increases the robustness of the aforementioned applications.

Recently, many of the state of the art DOA estimators have been reformulated using the spherical harmonic framework [2,3]. 3D DOA estimation using up to first order spherical harmonic signals has been initially proposed in directional analysis of sound fields for sound reproduction [4]. The analysis was performed using a b-format signal to calculate the instantaneous intensity vector. The intensity vector points to the direction of the net flow of energy. Based on the same principle, in [5] the authors proposed the use of a pseudointensity vector, formulated in the spherical harmonic domain, for estimating the DOA of a single source by averaging the estimated intensity vector at each time-frequency (TF) point across all the effective spectrum range. This algorithm has been extended to multiple sound source localization in [6], by using clustering to extract the position of sound sources and reflections. An approach for highly reverberant environments is proposed in [7], also in the spherical harmonic domain, by identifying the time-frequency bins where only one source is dominant with a direct-path dominance test.

Motivated by previous research that utilizes the intensity vector, in our proposed work we take advantage of the sparsity of speech signals in the TF domain. We apply a single source confidence measure [8] in order to locate the TF regions where only one source is active, i.e., single source zones (SSZs). We estimate the intensity vector at selected TF points in each detected SSZ and then we form 2D histograms which reveal the DOA of all the present active sources. By processing these 2D histograms we retrieve accurate 3D DOA estimation of the sources as shown in the extended simulation results.

The remainder of the paper is organized as follows. In Section 2 we review the theory of encoding the microphone signal from the space domain to the spherical harmonic domain. In Section 3 we present the proposed method including the single source zone detection and DOA estimation with 2D histogram processing. In Section 4 we present the evaluation results with multiple sound sources in different signal-to-noise ratio (SNR) and reverberation conditions. Conclusions are drawn in Section 5.

2. BACKGROUND

2.1. Spatial encoding

An overview of the process of how to spatially encode the microphone array sensor signals to a set of spherical harmonic signals is provided here for the sake of completeness. For

This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES-MusiNet.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

an extended overview of this process the reader is referred to [9]. Spatial encoding refers to the process of approximating the spherical harmonic signals, denoted as s_{lm}, where l is the order and m is the degree with $-l \leq m \leq l$, from microphone signals x_q of a microphone array with radius r and microphone positions $\Omega = [\theta_q, \phi_q]$ for elevation θ and azimuth ϕ, where $\theta \in [-\pi/2, \pi/2]$ and $\phi \in [0, 2\pi)$. The accuracy of this approximation depends on the number and how the microphones are distributed on the surface of the sphere, the radius r and the frequency k [10]. For a microphone array with $q = 1, \ldots, Q$ microphones the spherical harmonic signals can be approximated as [9]

$$s_{lm}(k) \approx \frac{1}{Q} \sum_{q=1}^{Q} g_{lm}^{q}(k) x_q(k, r, \Omega), \tag{1}$$

where $x_q(k, r, \Omega)$ is the q^{th} microphone signal in the frequency domain and $g_{lm}^{q}(k)$ is selected so that it provides an accurate approximation of the spherical Fourier transform [11].

2.2. Intensity vector

The proposed DOA algorithm is based on the sound intensity [12] and it has been utilized in parametric sound reproduction systems [4]. As in [5], the instantaneous active intensity vector can be approximated in the TF domain with n being the time index, as

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

where s_{00}^{*} is the complex conjugate of the 0^{th} order spherical harmonic signal, $b_0(k)$ is the mode strength compensation and s_x, s_y and s_z are the 1^{st} order spherical harmonic signals with their positive phase towards the x, y and z-axis respectively. Each of these signals is calculated as

$$s_\alpha(k, n) = \sum_{m=-l}^{l} \frac{Y_{lm}(\Omega_\alpha)}{b_l(k)} s_{lm}(k, n), \quad \alpha = \{x, y, z\}, \tag{3}$$

where $Y_{lm}(\Omega_\alpha)$ is the spherical harmonic base function of order $l = 1$ and degree m, Ω_α is set to $(0, 0)$, $(0, \pi/2)$ and $(\pi/2, 0)$ for each axis and $b_l(k)$ is the mode strength compensation for the specific order and depends on the type of the array [13].

3. PROPOSED METHOD

In the previous section we recalled the basic encoding of the microphone signals to the spherical harmonic domain as well as the definition and estimation of the intensity vector $\mathbf{I}(k, n)$, which indicates the direction of sound flow at a TF point. The vector pointing to the opposite direction of the intensity vector indicates the DOA. In this section we introduce the proposed method, which is based on the estimation of the intensity vector over selected TF points, such that each $\mathbf{I}(k, n)$ estimation reveals the DOA of each of the simultaneously active sound sources.

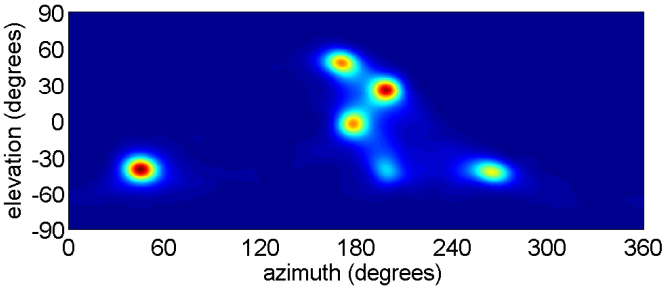

Fig. 1. Smoothed 2D histogram of six sources at RT_{60}=0.2 sec and SNR=45 dB.

3.1. Single source zones detection

We propose the use of the single-source zones (SSZs) framework as in [8]. A SSZ is a series of K frequency-adjacent TF points (K, n) where one source is dominant over any other active source and satisfies the following criterion:

$$\overline{\rho}(K, n) \geq 1 - \epsilon, \tag{4}$$

where $\overline{\rho}(K, n)$ is the average correlation coefficient between pairs of observations of adjacent microphones and ϵ is a small user-defined threshold.

The correlation coefficient $\rho_{i,j}(K, n)$ is defined as

$$\rho_{i,j}(K, n) = \frac{R_{i,j}(K, n)}{\sqrt{R_{i,i}(K, n) \cdot R_{j,j}(K, n)}}, \tag{5}$$

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). Thus, SSZ detection takes place in the TF domain. $X_i(k, n)$ and $X_j(k, n)$ are the microphone signals of the i^{th} and the j^{th} microphones respectively in the TF domain. Note that $x_q(k, r, \Omega)$ 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) parameters.

3.2. 3D DOA estimation via 2D histogram processing

After the detection of a SSZ, we estimate the intensity vector $\mathbf{I}(k, n)$ at d selected frequency components belonging to the SSZ, i.e., those TF points that correspond to the indices of the d highest peaks of the magnitude of the cross-power spectrum over all microphone pairs. In this manner we have d DOA estimates at each detected SSZ .

Algorithm 1: 2D Histogram Processing for 3D DOA estimation

1 Set the loop index $g = 1$.
2 Find $(\theta_g, \phi_g) = \arg\max_{\theta,\phi} \mathbf{y}_s^g(\theta, \phi)$, where $\mathbf{y}_s^g(\theta, \phi)$ is the smoothed histogram at the current iteration. The DOA of this source is (θ_g, ϕ_g).
3 Calculate the contribution of the current source as

$$\delta_g = \mathbf{y}_s(\theta, \phi) \odot \mathbf{w}_{\mathrm{C}}(\theta - \theta_g, \phi - \phi_g),$$

where the operator $\odot$ stands for element-wise multiplication.
4 Remove the contribution of this source as

$$\mathbf{y}_s^{g+1}(\theta, \phi) = \mathbf{y}_s^g(\theta, \phi) - \delta_g.$$

5 Increment g .
6 If $g \leq G$ go to step 2.

Once we have estimated all the local DOAs in the SSZs (Sections 2.2 & 3.1), 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 smooth the 2D histogram by applying an averaging filter, e.g., a circularly symmetric Gaussian window $\mathbf{w}_{\mathrm{A}}(\theta, \phi)$ of zero mean and standard deviation (std) equal to σ_{A}.

$$\mathbf{y}_{\mathrm{s}}(\theta, \phi) = \sum_i \sum_j \mathbf{y}(i, j) \mathbf{w}_{\mathrm{A}}(\theta - j, \phi - i), \qquad (6)$$

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, $\mathbf{y}(\theta, \phi)$ is the original 2D histogram and $\mathbf{y}_{\mathrm{s}}(\theta, \phi)$ is the smoothed one. An example of a smoothed histogram of six sources at $(-40, 45)^\circ$, $(-3, 177)^\circ$, $(-42, 265)^\circ$, $(49, 170)^\circ$, $(-42, 199)^\circ$, and $(26, 199)^\circ$ at RT_{60}=0.2 sec and 45 dB SNR of additive white Gaussian noise is shown in Figure 1.

In order to extract the final DOA estimates of the sources, we proceed further by processing the 2D smoothed histogram. We detect the highest peak of the smoothed histogram and we identify its index as the DOA of the first source. Then we remove its contribution from the histogram by applying a Gaussian window $\mathbf{w}_{\mathrm{C}}(\theta, \phi)$ of zero mean and std equal to σ_{C}. We proceed to the detection of the second peak and the removal of its contribution and iteratively to the next peak until we reach the number G of sources. The steps of the aforementioned iterative procedure are described in detail in Algorithm 1 and in Figure 2 we show an example for a scenario with 4 sources at $(54, 82)^\circ$, $(43, 118)^\circ$, $(-58, 307)^\circ$, and $(-22, 172)^\circ$ at RT_{60}=0.2 sec and 45 dB SNR of additive white Gaussian noise.

The core steps of our method are summarized as follows:

1. The encoding of the microphone signals to the spherical harmonic domain (Section 2.1).
2. The detection of all the SSZs (Section 3.1).
3. The $\mathbf{I}(k, n)$ estimation leading to DOA estimates in the SSZs (Section 2.2).
4. The generation and smoothing of the 2D histogram of a block of DOA estimates (Section 3.2).
5. The processing of the smoothed 2D histogram to extract the 3D DOA estimates (Section 3.2, Algorithm 1).

4. EVALUATION

The performance of the proposed method is investigated by conducting extended simulations in anechoic and reverberant environments. We have employed a rigid spherical microphone array comprising 32 microphones, placed according to the angular positions of the 3D sphere covering problem solutions [14] and radius equal to $r = 0.042$ m. We used the spherical microphone array room impulse response generator by Jarrett et al [15] which is based on the image method of Allen and Berkley [16] to simulate a room of $8 \times 8 \times 6$ meters. The spherical array was placed in the centre of the room, and the simulated sound sources were placed 1.5 m away from the centre of the array. The speed of sound was $c = 343$ m/s while the frequency range used was 500-3800 Hz to avoid aliasing phenomena [7]. In each simulation the sound sources had equal power and the signal-to-noise ratio at each microphone was estimated as the ratio of the power of each source signal to the power of the noise signal. Any other parameters and their corresponding values can be found in Table 1. The performance of the proposed algorithm was 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 ($\hat{\mathbf{v}}$) [5] over all sound sources, all positions and all the frames of the

Table 1. Simulation parameters

parameter	notation	value
sampling frequency	f_s	48000 Hz
frame size		2048 samples
overlapping in time		50%
FFT size		2048 samples
TF zones width	K	375 Hz
overlapping in frequency		50%
SSZ threshold	ϵ	0.2
frequency bins/SSZ	d	2
histogram bin size		$0.5^\circ \times 0.5^\circ$
averaging window std	σ_{A}	5°
localization window std	σ_{C}	20°

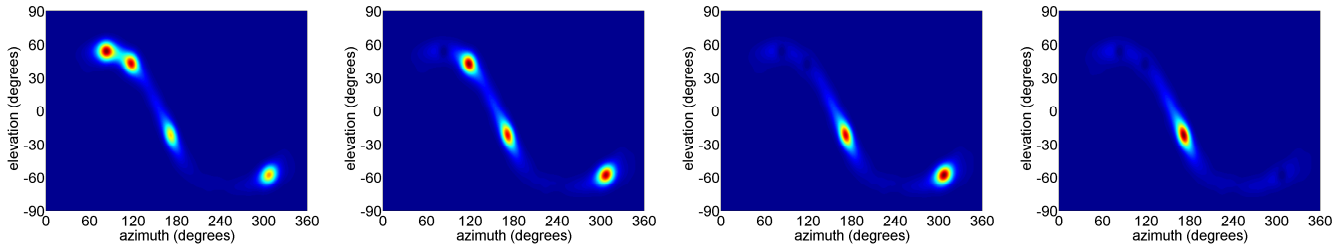

Fig. 2. Visualization of Algorithm 1. (a) The 2D histogram given as input to Algorithm 1. Four sources are clearly visible. (b) The 2D histogram after the first iteration. The contribution of the first detected source at $(54, 82)^\circ$ has been removed while the DOAs of the three remaining sources are highlighted. (c) The 2D histogram after the second iteration. (d) The 2D histogram after the third iteration where only the contribution the fourth source at $(-22, 172)^\circ$ is present.

source signals. The error is defined as

$$\text{MEE} = \frac{1}{N_\text{P} N_\text{F}} \sum_{p,n} \frac{1}{G} \sum_{g} \cos^{-1}\left(\mathbf{v}_{png}^T \hat{\mathbf{v}}_{png}\right), \quad (7)$$

where $\cos^{-1}\left(\mathbf{v}_{png}^T \hat{\mathbf{v}}_{png}\right)$ expresses the angular distance between the true DOA of the g^{th} active source in the p^{th} positioning in the n^{th} frame and the estimated one. The association between the true and the estimated DOA of a source is determined based on the permutation that leads to the minimum error, given the permutations between the true DOAs and the estimated ones. N_P is the total number of different positions of the sound sources around the array, i.e., the sound sources were placed in different and random orientations in each simulation and the total number of different positions is $N_\text{P} = 10$. N_F is the total number of frames after subtracting $N - 1$ frames of the initialization period and G is the total number of active sources, which is assumed to be known. In all the simulations speech sound files were used of duration approximately equal to 9 seconds, leading to $N_\text{F} = 375$ frames. Any gaps or silent periods were manually removed. The block size is equal to 1 second, i.e., $N = 46$ frames.

In our first set of simulations we investigated the spatial resolution of our proposed method, i.e., how close two sources can be while accurately estimating their DOA. Figure 3 shows the MEE against the angular separation of two continuously active sound sources, one male and one female speaker, for SNR=$\{15, 20, 45\}$ dB of additive white Gaussian noise and reverberation time $\text{RT}_{60} = \{0.4, 0.8\}$ sec. The MEE is very low even when the sources are very close to each other, e.g., for angular separation equal to 20°, RT_{60}=0.8 sec and SNR=15 dB, the MEE is equal to 5.74°. With increasing SNR and decreasing reverberation time the MEE is improved as expected and shown in Figure 3.

Aiming at highlighting the impact of the SSZ selection on the DOA estimation robustness, we compared our proposed method against the W-disjoint orthogonality (WDO) assumption [17], where each TF point is assumed to be dominated by a single source, therefore $\mathbf{I}(k, n)$ is estimated for every TF point in the frequency range of interest. Figure 4 shows the MEE versus the reverberation time, $\text{RT}_{60} = \{0, 0.2, 0.4, 0.6, 0.8\}$ sec, where RT_{60}=0 sec corresponds to the anechoic case at SNR=$\{15, 20, 45\}$ dB for scenarios with four simultaneously and continuously active sources, two male and two female speakers. The minimum angular separation between the sources was 19.94°. As it is shown in Figure 4 the proposed SSZ method experiences lower error for all SNR and RT_{60} conditions which was expected since our method selects only TF points belonging to single-source zones and avoids the use of TF points with spurious $\mathbf{I}(k, n)$ information.

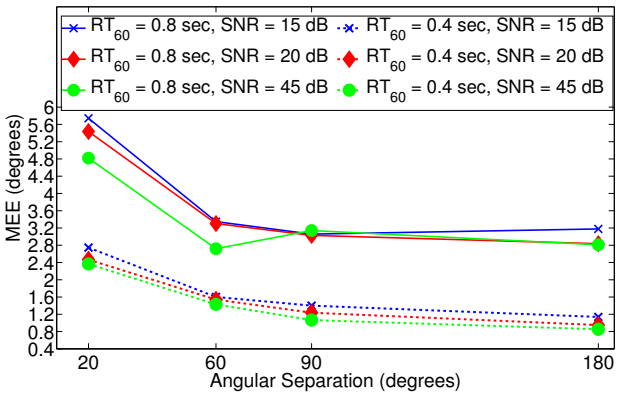

Fig. 3. MEE versus angular separation between 2 sound sources in various SNR and reverberation conditions.

We show the performance of the proposed method when the number of simultaneously active sources is considered to be relatively high, i.e., $G = 6$ in Figure 5. Three male and three female speech sound files were used with the minimum angular separation between them being at 21.14°. Once again the MEE versus the reverberation time for various SNR conditions is shown. The MEE is higher compared to the 4 sources scenarios shown in Figure 4. The method performs robustly in moderate reverberation conditions, experiencing MEE equal to 19.57° in the worst case scenario, i.e., at $\text{RT}_{60} = 0.8$ sec and SNR=15 dB.

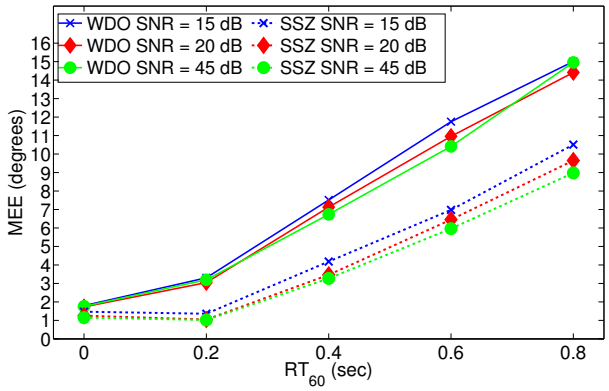

Fig. 4. MEE versus RT_{60} for scenarios with 4 simultaneously active sound sources in various SNR conditions.

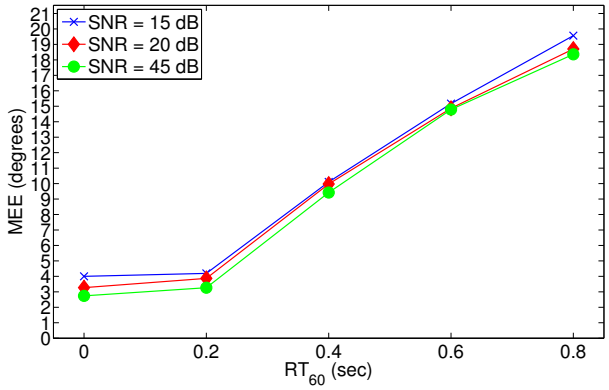

Fig. 5. MEE versus RT_{60} for scenarios with 6 simultaneously active sound sources in various SNR conditions.

5. CONCLUCION

A multiple sound source 3D DOA estimator is proposed in this paper based on the intensity vector. The algorithm detects single source zones and the DOAs of the sound sources are estimated through 2D histogram processing. The accuracy of the proposed algorithm in estimating simultaneously active sound sources of different number and placements was evaluated with the mean error estimator and demonstrated its effectiveness at different reverberation and SNR conditions. In future work, we will examine the joined estimation of the number of active sources and their corresponding 3D DOAs and proceed with experiments in real environments.

REFERENCES

[1] Harry L Van Trees, *Detection, estimation, and modulation theory, optimum array processing*, John Wiley & Sons, 2004.

[2] H.l Sun et al, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," in *IEEE ICASSP*, 2011, pp. 117–120.

[3] X. Li, S. Yan, X. Ma, and C. Hou, "Spherical harmonics MUSIC versus conventional MUSIC," *Applied Acoustics*, vol. 72, no. 9, pp. 646–652, 2011.

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

[5] D. P. Jarrett, E. Habets, and P. A Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *EUSIPCO, Aalborg, Denmark*, 2010, pp. 442–446.

[6] 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.

[7] 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, October 2014.

[8] D. Pavlidi et al, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 2193–2206, Oct. 2013.

[9] 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.

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

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

[12] F. Fahy, *Sound intensity*, CRC Press, 2002.

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

[14] J. H. Conway et al, *Sphere packings, lattices and groups*, vol. 3, Springer-Verlag New York, 1993.

[15] D. P. Jarrett et al, "Rigid sphere room impulse response simulation: Algorithm and applications," *JASA*, vol. 132, no. 3, pp. 1462–1472, audiolabs-erlangen.de/fau/professor/habets/software/smirgenerator.

[16] 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.

[17] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.