Spatial Covariance Matrix Estimation for Reverberant Speech with Application to Speech Enhancement

Ran Weisman[1], *Vladimir Tourbabin*[2], *Paul Calamia*[2], *Boaz Rafaely*[1]

[1] School of Electrical and Computer Engineering
Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel
[2] Facebook Reality Labs

ranweis@post.bgu.ac.il, vtourbabin@fb.com, pcalamia@fb.com, br@bgu.ac.il

Abstract

A wide range of applications in speech and audio signal processing incorporate a model of room reverberation based on the spatial covariance matrix (SCM). Typically, a diffuse sound field model is used, but although the diffuse model simplifies formulations, it may lead to limited accuracy in realistic sound fields, resulting in potential degradation in performance. While some extensions to the diffuse field SCM recently have been presented, accurate modeling for real sound fields remains an open problem. In this paper, a method for estimating the SCM of reverberant speech is proposed, based on the selection of time-frequency bins dominated by reverberation. The method is data-based and estimates the SCM for a specific acoustic scene. It is therefore applicable to realistic reverberant fields. An application of the proposed method to optimal beamforming for speech enhancement is presented, using the plane wave density function in the spherical harmonics (SH) domain. It is shown that the use of the proposed SCM outperforms the commonly used diffuse field SCM, suggesting the method is more successful in capturing the statistics of the late part of the reverberation.

Index Terms: Spatial correlation matrix, reverberant speech, spherical arrays, minimum-variance distortionless response

1. Introduction

A wide range of applications in speech and audio signal processing, such as speech enhancement, beamforming, speech dereverberation and spatial audio coding, incorporate speech signals in reverberant environments. Many of the methods developed for these applications require modeling of room reverberation. A common statistical tool for describing reverberant sound fields is the spatial covariance matrix (SCM), also named the power spectral density (PSD) matrix [1]. For example, the design of informed spatial filters for source separation [2, 3, 4, 5], dereverberation and enhancement [1, 6, 7] is based on the SCM. In spatial audio coding, the SCM is used for parametric encoding of the reverberant sound field component [8, 9, 10]. Furthermore, in acoustic scene analysis, the direct-to-reverberant ratio (DRR) is typically estimated by using the structure of the SCM of the direct sound and the reverberant field components [11, 12, 13].

The SCM contains the narrowband second-order spatial statistics of the signal, which encodes information about the spatial distribution and spatial correlation of the sound field. A diffuse field model for the sound field is commonly used [1], as it leads to mathematical tractability and simple analytical formulations. The diffuse model incorporates an assumption of spherical isotropy, which is sometimes replaced with cylindrical isotropy [6]. Nevertheless, the model and its inherent assumptions of isotropy and lack of directional correlation may be inaccurate in practice, e.g. due to differences in the absorption of room boundaries or asymmetry in room geometry [14, 15]. These inaccuracies may lead to errors in the SCM, as demonstrated in this paper, thereby potentially degrading the performance of the corresponding methods. Although there are a few examples in the literature for the use of non-diffuse models for the SCM [4, 16], a general SCM estimation framework that can represent realistic sound fields is not yet available.

In this paper, a data-based method for estimating the SCM of an arbitrary reverberant sound field, which is not necessarily diffuse, is presented. The method operates in the time-frequency (TF) domain, identifying reverberant TF bins that are then used for the SCM estimation. The method's performance is demonstrated by a simple application of minimum variance distortionless response (MVDR) beamforming for dereverberation, using the plane wave density function signals in the spherical harmonics (SH) domain. The method is shown to outperform the baseline method based on the diffuse field assumption, but the gap from oracle performance highlights the need for improved classification of reverberant TF bins.

2. Signal Model

The signal model employed in this work assumes a single speaker in a reverberant room, with sound pressure signals captured by a microphone array. The Q microphone signals are presented in the short-time Fourier transform (STFT) domain, such that $\mathbf{x}(k,n) = [X_1(k,n),\dots,X_Q(k,n)]^T$, where k and n are the frequency and time-frame indices, respectively. The signals are decomposed into three components, using a parametric model, which is commonly used in speech enhancement frameworks [1, 3]:

$$\mathbf{x}(k,n) = \mathbf{x}_e(k,n) + \mathbf{x}_r(k,n) + \mathbf{v}(k,n) \qquad (1)$$

where $\mathbf{x}_e(k,n)$ represents the signal due to the early part of the room impulse response (RIR), $\mathbf{x}_r(k,n)$ represents the late reverberation, and $\mathbf{v}(k,n)$ is an additive noise signal. The early part is formulated as $\mathbf{x}_e(k,n) = \mathbf{g}(k)s(k,n)$, where $s(k,n)$ is the source signal as measured at some reference point (e.g. the array center), and $\mathbf{g}(k)$ is the acoustic transfer-function (ATF) vector, incorporating the direct sound, or the direct sound and early room reflections. Note that $\mathbf{g}(k)$ is assumed to be known, and its estimation is beyond the scope of this paper. If $\mathbf{g}(k)$ models the direct sound, it can be computed by estimating the direction-of-arrival (DOA). Otherwise, it has to be estimated in other ways (see [17] as an example).

The SCMs of the different signal components are formulated as:

$$\mathbb{E}\left[\mathbf{x}_e(k,n)\mathbf{x}_e(k,n)^H\right] = \sigma_s^2(k,n)\mathbf{g}(k)\mathbf{g}(k)^H \tag{2}$$

$$\mathbb{E}\left[\mathbf{x}_r(k,n)\mathbf{x}_r(k,n)^H\right] = \sigma_r^2(k,n)\mathbf{\Gamma}(k) \tag{3}$$

$$\mathbb{E}\left[\mathbf{v}(k,n)\mathbf{v}(k,n)^H\right] = \mathbf{\Sigma}_v(k) \tag{4}$$

where $\mathbb{E}[\cdot]$ is the expectation operator, and $\sigma_s^2(k,n)$ and $\sigma_r^2(k,n)$ are non-negative scalars representing the power of the early and reverberant signal components. The reverberation SCM is given by a time-invariant matrix $\mathbf{\Gamma}(k)$, scaled by the time-varying coefficient $\sigma_r^2(k,n)$, representing the change in the reverberation signal power over time. Note that the modeling of $\mathbf{\Gamma}(k)$ and $\mathbf{g}(k)$ as time-invariant is based on the stationarity of the acoustic scene, with constant source and array positions. $\mathbf{\Sigma}_v(k)$ is assumed to be time-invariant as well. Due to the short temporal autocorrelation of speech, $\mathbf{x}_e(k,n)$ and $\mathbf{x}_r(k,n)$ are assumed to be uncorrelated. Assuming also that the noise is uncorrelated with the source, then:

$$\begin{aligned}&\mathbb{E}\left[\mathbf{x}(k,n)\mathbf{x}(k,n)^H\right] = \\ &\sigma_s^2(k,n)\mathbf{g}(k)\mathbf{g}(k)^H + \sigma_r^2(k,n)\mathbf{\Gamma}(k) + \mathbf{\Sigma}_v(k)\end{aligned} \tag{5}$$

The focus of this paper is the matrix $\mathbf{\Gamma}(k)$, which will be referred to as the reverberation normalized spatial covariance matrix (RNSCM). This matrix encodes the structure of the reverberation SCM, and is usually assumed to be known, mostly by assuming an ideal diffuse field.

A common representation of a reverberant sound field is the plane wave amplitude density (PWD) function, denoted by $a(k,n,\mathbf{\Omega})$ [18]. k and n are the frequency and time indices respectively, $\mathbf{\Omega} \in \mathbb{S}^2$ represents a direction in 3D space, and $\mathbb{S}^2$ is the unit sphere. The PWD function can be represented in the SH domain, using a set of SH coefficients, denoted by a_{nm} [18]. In the following, the microphone signals are replaced by these PWD coefficients in the SH domain, i.e. $\mathbf{x}(k,n) \triangleq [a_{00}(k,n),\ldots,a_{NN}(k,n)]^T$, where a finite SH order N is used, leading to $Q=(N+1)^2$ coefficients. This assumption may require the use of a spherical microphone array [18], but it is believed that the method and results of this work could be generalized to other arrays configurations.

3. Validity of the Diffuse Model Assumption

A diffuse field is commonly described by an infinite number of plane waves of equal magnitude, arriving uniformly from all directions with random phase [19]. Under these certain assumptions, the RNSCM of the PWD function in the SH domain is given by $\mathbf{\Gamma}_{\text{diff}}(k) = \mathbf{I}$ [8].

The validity of this diffuse model in representing real sound fields is examined using real data from the ACE challenge [20], which contains recorded impulse responses of various rooms. The data was recorded by the MH Acoustics Eigenmike, a 32-channel spherical microphone array. Two rooms from this database were employed, with parameters shown in table 1. The first 50 ms of each RIR were truncated, keeping only the late reverberation part of the RIR. Each truncated RIR was convolved with 5 seconds of white noise. Then, the order 3 SH coefficients of the PWD function were computed using the robust-PWD algorithm [21], after filtering the signal to a frequency band centered at 3000 Hz, with a bandwidth of 100 Hz. The SCM from Eq. (3), denoted by $\mathbf{R}$, and the directional variance of the sound field, $\mathbb{E}\left[|a(f_k,\mathbf{\Omega})|^2\right]$, were computed from the measured PWD signals by time averaging. The validity of the diffuse model was investigated by computing the projection error of $\mathbf{R}$ on the diffuse RNSCM, using:

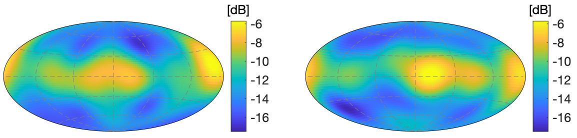
$$\epsilon \triangleq \frac{||\mathbf{R} - \frac{1}{Q}\text{tr}(\mathbf{R})\cdot\mathbf{I}||_F}{||\mathbf{R}||_F} \tag{6}$$

where $||\cdot||_F$ is the Frobenius norm and tr $(\cdot)$ is the trace operator.

The results are shown in Figure 1. The non-isotropic distribution of the field can be observed, as more energy is concentrated in the horizontal plane. Also, the error ϵ is relatively high (around 0.5), which shows the inaccuracy of the diffuse field model, providing motivation for using more accurate models.

Table 1: *Room parameters from ACE challenge data, including mean T_{60} and DRR*

Room	Dimensions [m]	T_{60} [s]	DRR [dB]
1	$5.1 \times 4.5 \times 3.2$	0.57	1.6
2	$6.9 \times 9.7 \times 3$	0.66	3.3



(a) Room 1 ($\epsilon = 0.45$)

(b) Room 2 ($\epsilon = 0.49$)

Figure 1: $\mathbb{E}\left[|a(k,\mathbf{\Omega})|^2\right]$ *and diffuse projection error ϵ of the late reverberation of the two rooms (ACE challenge data)*

4. Proposed Method

The proposed method estimates the RNSCM, $\mathbf{\Gamma}(k)$, based on a subset of TF bins that is classified as highly reverberant, ignoring the other bins. For this classification, a measure is proposed for quantifying reverberation in each TF bin. This measure is referred to as the reverberant field dominance (RFD), denoted by $\gamma(k,n)$. Assuming $\mathbf{v}(k,n) \equiv 0$ for simplicity, TF bins for which the RFD measure is high enough, satisfy:

$$\mathbf{x}(k,n) \approx \mathbf{x}_r(k,n) \tag{7}$$

The set of TF bins with a sufficiently high RFD is denoted as:

$$\mathcal{A}_{\text{RFD}} \triangleq \left\{(k,n) \mid \gamma(k,n) > \gamma_0\right\} \tag{8}$$

Defining a weight function $w(k,n)$, the proposed RNSCM estimator is formulated as:

$$\hat{\mathbf{\Gamma}}(k) \triangleq \frac{\sum_{(k,n)\in\mathcal{A}_{\text{RFD}}} w(k,n)\mathbf{x}(k,n)\mathbf{x}(k,n)^H}{||\sum_{(k,n)\in\mathcal{A}_{\text{RFD}}} w(k,n)\mathbf{x}(k,n)\mathbf{x}(k,n)^H||_F} \tag{9}$$

Note that $||\hat{\mathbf{\Gamma}}(k)||_F = 1$. The weight function is chosen to be $w(k,n) = 1$ to simplify the discussion, and the choice of the threshold γ_0 depends on the specific RFD measure, as detailed in section 5.

In this paper, two RFD measures are proposed and examined. First, a measure based on the early-to-late reverberation ratio, the RFD-ELR, is proposed:

$$\gamma_{\text{ELR}}(k,n) \triangleq -10\log\left(\frac{|X_{\text{e-ref}}(k,n)|^2}{|X_{\text{r-ref}}(k,n)|^2}\right) \tag{10}$$

where $X_{\text{e-ref}}(k,n)$ and $X_{\text{r-ref}}(k,n)$ are the early and reverberant components of the signal at a reference point, calculated separately by using the RIR. Note that this data is not directly accessible in practice. However, since the estimation is performed with the observed data only, the usage of this measure can show the estimation potential and the existence of bins with relevant information for the RNSCM, as demonstrated in section 5. Note that γ_{ELR} is higher when the ELR is lower, i.e. when the early component is less dominant. Also note that in the case where the early component models the direct sound, the ELR reduces to the DRR measure.

Secondly, a more practical RFD measure, that does not require the RIR, is proposed. This measure is based on a transfer function vector similarity test, the RFD-TFVS:

$$\gamma_{\text{TFVS}}(k,n) \triangleq$$

$$1-\frac{\sum\limits_{n'=-n_0}^{n_0}\sum\limits_{k'=-k_0}^{k_0}|\mathbf{x}(k+k',n+n')^H\mathbf{g}(k+k')|^2}{\sum\limits_{n'=-n_0}^{n_0}\sum\limits_{k'=-k_0}^{k_0}||\mathbf{x}(k+k',n+n')||^2\cdot||\mathbf{g}(k+k')||^2} \tag{11}$$

where n_0, k_0 are parameters controlling the amount of time and frequency averaging, respectively. This measure quantifies the similarity between a TF bin and its neighborhood to the transfer function vector, and is normalized such that $0 \leq \gamma_{\text{TFVS}} \leq 1$. When there is less similarity, higher values are obtained, suggesting that the reverberant part is more dominant.

The estimation process relies on the existence of bins with dominant reverberation energy, which depends on the nature of the RIR and the source signal. In the case of speech signals, the non-stationarity in time leads to onset and offset time segments, with the latter giving rise to reverberation-dominant segments. Such behavior may require that the tail of the RIR does not decay too fast (i.e. T_{60} is high enough).

5. Experimental Study

The performance of the proposed method is demonstrated by an application of MVDR beamforming to dereverberation of speech.

5.1. Setup

An acoustic scene of a single speaker in a reverberant room was simulated. A room of dimensions $10 \times 8 \times 3$ m and $T_{60} = 1$s was chosen, and the source and array positions were randomly drawn such that the steady-state DRR was -8 dB. The RIR was simulated by the image method [22], providing directly the SH coefficients of the PWD function, with SH order $N = 3$, i.e. $Q = 16$ coefficients, and a sampling frequency of 16 kHz. The signal $\mathbf{x}$ was then obtained by convolving the RIR with a speech signal of 5 seconds length. STFT was performed using a Hanning window of 512 samples, 25% hop size and FFT size of 1024. This simulation process was repeated 50 times, each time drawing different source and array positions and different speech signals, leading to 50 generated acoustic scenes with the same DRR and T_{60}.

5.2. Methodology

In each simulation, MVDR beamforming was employed with the aim of reducing the reverberant signal component, and the results were evaluated as detailed next. Since the interference is composed of reverberation only, the MVDR beamformer is given by [23]:

$$\mathbf{w}(k)^H \triangleq \frac{\mathbf{g}(k)^H\mathbf{\Gamma}(k)^{-1}}{\mathbf{g}(k)^H\mathbf{\Gamma}(k)^{-1}\mathbf{g}(k)} \tag{12}$$

where the output signal is $y(k,n) = \mathbf{w}(k)^H\mathbf{x}(k,n)$. Generally speaking, if the RNSCM is more accurate, then better performance is expected. Therefore, the performance of 4 different RNSCMs was compared: diffuse RNSCM, estimated RNSCM using RFD-DRR and RFD-TFVS measures, and oracle RNSCM.

The diffuse RNSCM is signal independent and given by $\mathbf{\Gamma}_{\text{diff}}(k) = \mathbf{I}$. The other RNSCMs were computed for each simulation, using the reverberant speech. The estimated RNSCM was computed according to (9). First, TF bins with low energy were ignored in both RFD measures, by taking into account only bins with an energy level higher than the median. The RFD-DRR was calculated according to (10), using the signal at the array center, and γ_0 was set to 15, which is equivalent to choosing bins with DRR below -15 dB. The RFD-TFVS was calculated according to (11), with n_0, k_0 taken to be equivalent to averaging over 30ms and 100Hz, and γ_0 was adjusted such that 20% of the bins were chosen. The oracle RNSCM was computed by $\mathbf{\Gamma}_{\text{oracle}}(k) \triangleq \sum_n \mathbf{x}_r(k,n)\mathbf{x}_r(k,n)^H$ using all bins, as $\mathbf{x}_r(k,n)$ can be calculated separately using the RIR. Moreover, in all the simulations the ATF vector $\mathbf{g}(k)$ was assumed to model the direct part only, and was calculated according to the DOA, which was assumed to be known. Therefore, the early reflections were included in $\mathbf{x}_r(k,n)$ in practice.

In each simulation, the performance was evaluated by the frequency-weighted segmental signal-to-noise ratio (fwSegSNR) [24] and the PESQ [25] measures, using the direct sound as the reference signal. The difference in each measure was calculated, comparing the beamformer output signal, and the input signal at the array center, given by the first entry of $\mathbf{x}$, which contains the a_{00} coefficient. Then, the final results were computed by averaging all simulations.

5.3. Results and Discussion

Performance in terms of ΔfwSegSNR and ΔPESQ is presented in Figure 2. As can be seen, for both measures the diffuse model yields the worst performance, while the oracle RNSCM yields the best performance. This is expected, since the diffuse RNSCM is a scene-invariant model, while the oracle RNSCM is tailored to the specific scene, and computed directly using the true reverberant signal, which cannot be separately observed in practice. Moreover, the relatively large performance gap, may indicate that the reverberation is not perfectly diffuse. Compared to these, the RFD method performs better than the diffuse case and worse than the oracle case, implying that the estimated RNSCM manages to incorporate some non-diffuse characteristics of the reverberation. In addition, the RFD-DRR outperforms the RFD-TFVS. However, recall that this measure is based on the true DRR, which has to be estimated in practice. An example of the TF bins selected by each RFD measure is shown in Figure 3. As can be seen, many of the bins selected by the RFD-TFVS are also selected by the RFD-DRR. However, due to the calculation method of the RFD-TFVS, the selected

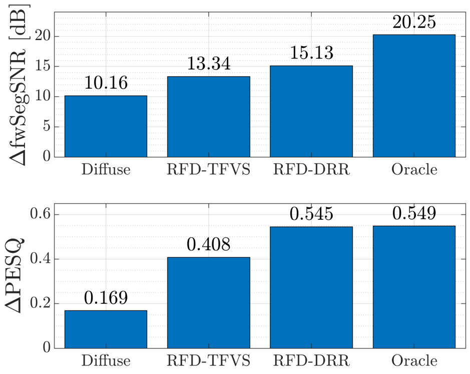
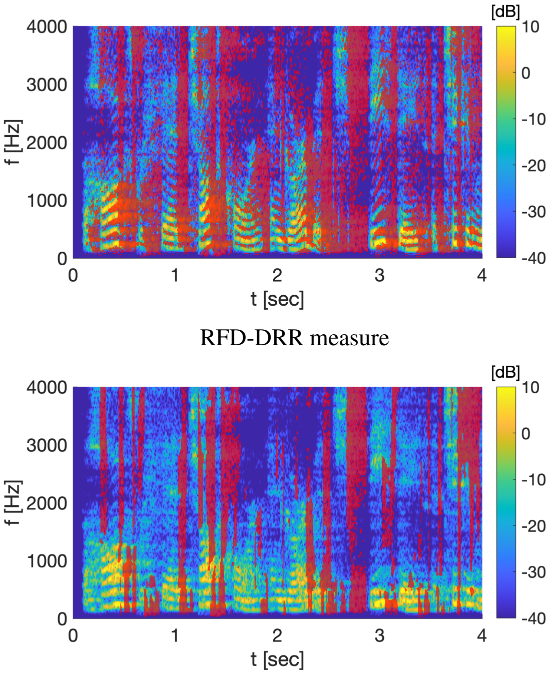


Figure 2: *ΔfwSegSNR and ΔPESQ for all three methods and the oracle reference*

areas it yields are more smooth in the TF domain, which leads to selecting bins with higher DRR as well. Note that compared to the RFD-DRR, fewer bins are selected by the RFD-TFVS, due to the different thresholds, which were adjusted separately such that each measure will yield the best performance for the given data.

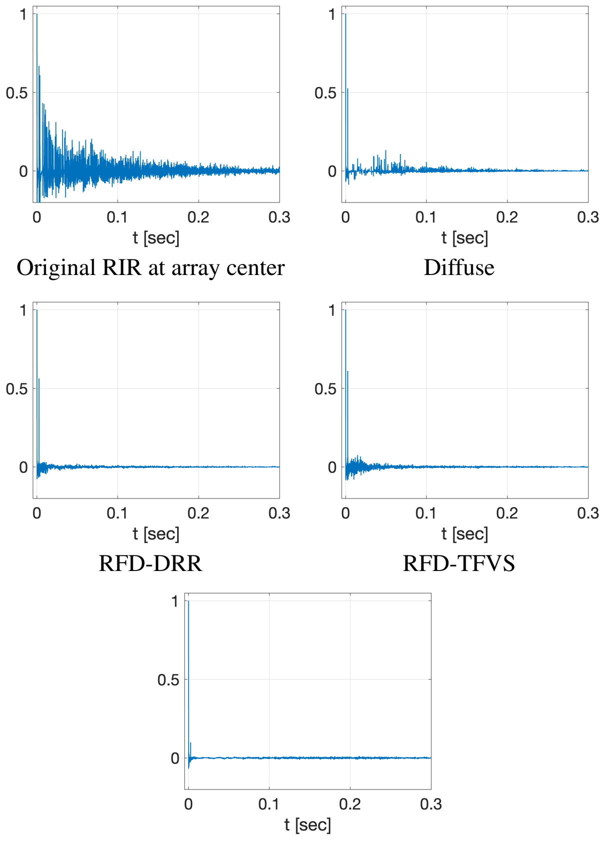
For further detailed analysis, Figure 4 shows an example of the RIR at the output of the MVDR beamformer for each method. Note that these are the same beamformers that were computed based on the speech signals. It can be observed that the beamformers that use the estimated RNSCM based on the RFD-DRR and RFD-TFVS attenuate better the late reverberation, compared to the diffuse RNSCM case. This result verifies that the observed improvement in Figure 2 is due to better attenuation of the late reverberation. Moreover, it seems that the oracle-based MVDR attenuates better the early reflections, compared to the RFD-DRR and RFD-TFVS, which may explain the performance gap between them.



RFD-DRR measure

RFD-TFVS measure

Figure 3: *An example of the selected TF bins (marked in red) for the SCM estimation, using the two different RFD measures*



Original RIR at array center

Diffuse

RFD-DRR

RFD-TFVS

Oracle

Figure 4: *An example of the RIRs at the beamformer output, using different RNSCMs, showing only the first 0.3 s*

6. Conclusion

In this work, a data-based approach for modeling the SCM of reverberant speech was presented. The method estimates the SCM by using TF bins with dominant reverberation, and it is apparently more effective in modeling the late reverberation compared to a diffuse field based model. A simple application of MVDR beamforming was presented, confirming the existence of vital information in reverberant speech signals, that can be exploited to more accurately estimate the SCM of the reverberant part. Future work will focus on incorporating the early reflections into the estimation framework, and examining the SCM estimation performance under noisy conditions. Additional bin-selection methods will also be examined, in order to reduce the performance gap from the oracle.

7. Acknowledgements

This research was supported by Facebook Reality Labs.

8. References

[1] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and com-

parison of late reverberation power spectral density estimators," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1056–1071, 2018.

[2] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[4] A. Fahim, P. N. Samarasinghe, and T. D. Abhayapala, "PSD estimation and source separation in a noisy reverberant environment using a spherical microphone array," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1594–1607, 2018.

[5] N. Q. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, no. 1, p. 149, 2013.

[6] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1599–1612, 2016.

[7] S. Braun and E. A. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE, 2013, pp. 1–5.

[8] A. Politis, S. Tervo, and V. Pulkki, "Compass: Coding and multidirectional parameterization of ambisonic sound scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6802–6806.

[9] C. Schörkhuber and R. Höldrich, "Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of ambisonic signals," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[10] A. Politis, J. Vilkamo, and V. Pulkki, "Sector-based parametric sound field reproduction in the spherical harmonic domain," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 852–866, 2015.

[11] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, 2011.

[12] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.

[13] D. P. Jarrett, O. Thiergart, E. A. Habets, and P. A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*. IEEE, 2012, pp. 1–5.

[14] M. Hodgson, "When is diffuse-field theory applicable?" *Applied Acoustics*, vol. 49, no. 3, pp. 197–207, 1996.

[15] Y. IZUMI and M. OTANI, "Direction-of-arrival distribution analysis of reflected sounds using spherical microphone array."

[16] O. Schwartz, S. Gannot, and E. A. Habets, "An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1495–1510, 2016.

[17] R. Varzandeh, M. Taseska, and E. A. Habets, "An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017, pp. 11–15.

[18] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

[19] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[20] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge—corpus description and performance evaluation," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[21] D. L. Alon, B. Rafaely, V. Pulkki, S. Delikaris-Manias, and A. Politis, "Spatial decomposition by spherical array processing," *Parametric time-frequency domain spatial audio*, vol. 25, 2017.

[22] J. B. Allen and D. 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.

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

[24] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Ninth international conference on spoken language processing*, 2006.

[25] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.