Data-Dependent Initialization for ECM-Based Blind Geometry Estimation of a Microphone Array Using Reverberant Speech

Klaus Brümann, Daniel Fejgin, Simon Doclo

Department of Medical Physics and Acoustics, and Cluster of Excellence Hearing4all, University of Oldenburg, Germany
Email: {klaus.bruemann,daniel.fejgin,simon.doclo}@uni-oldenburg.de

Abstract

Recently a method has been proposed to blindly estimate the geometry of an array of distributed microphones using reverberant speech, which relies on estimating the coherence matrix of the reverberation using an iterative expectation conditional-maximization (ECM) approach. Instead of using a data-independent initial estimate of the coherence matrix and a matched beamformer to estimate the initial speech and reverberation power spectral densities, in this paper we propose to use a data-dependent initial estimate of the coherence matrix and a time-varying minimum-power-distortionless-response beamformer. Simulation results show that the proposed ECM initialization significantly improves the estimation accuracy of the microphone array geometry and increases the generalizability for microphone arrays of different sizes.

1 Introduction

Through the technological development and availability of digital devices with built-in microphones, such as mobile phones or smart wearable devices, ad-hoc distributed microphones are becoming more ubiquitously available for signal processing applications such as speaker localization or speech enhancement. In ad-hoc situations, acoustical quantities provide a cheap and convenient way to blindly estimate the microphone array geometry (MAG) (see [1] for an overview). In this paper we focus on blind MAG estimation (also referred to as geometry calibration) using reverberant speech.

An acoustical quantity which is commonly used to blindly estimate the MAG is the time-difference of arrival (TDOA) between the microphones. The TDOA encodes a component of the pairwise distances between the microphones, depending on the direction of arrival of the source. However, it has been shown that estimating the MAG using TDOAs [2, 3] requires either a relatively large number of source signals from different locations or a moving source. In addition, TDOA estimation becomes increasingly unreliable in high reverberation [4].

Alternatively, it has been shown in [5, 6] that the coherence of diffuse or reverberant sound fields can be used for blind MAG estimation. The method proposed in [6] assumes a single speech source in a reverberant environment and consists of three steps. First, the coherence matrix of the reverberation is estimated from the reverberant microphone signals using an expectation conditional-maximization (ECM) approach [7, 8]. The pairwise distances between the microphones are then estimated by comparing the entries of the estimated coherence matrix to a distance-dependent coherence model [9] for a set of frequencies. Finally, the MAG is estimated from the pairwise distances using multi-dimensional scaling [5, 10, 11].

In ECM, the coherence matrix and the speech and reverberation power spectral densities (PSDs) are estimated by iteratively increasing a non-convex likelihood function, which describes how well these parameters represent the reverberant speech signals. If the initial estimates of these parameters are poor, the optimization may converge to a local optimum. In [6], ECM is initialized with a data-independent estimate of the coherence matrix and a matched beamformer is used to estimate the initial speech and reverberation PSDs. To improve these initial estimates, in this paper we propose to initialize ECM with a data-dependent estimate of the coherence matrix and replace the matched beamformer with a time-varying minimum-power-distortionless-response (MPDR) beamformer. We experimentally show that the frequency range used to estimate the pairwise distances plays an important role and demonstrate the advantage of using a data-dependent ECM initialization. In addition, for spatially distributed microphone arrays of different sizes (with microphone distances up to 87 cm), experimental results show that the proposed data-dependent initialization consistently results in significantly smaller MAG estimation errors than using a data-independent initialization.

2 Signal Model

We consider a reverberant environment with one speech source and a spatially distributed, 3-dimensional microphone array with $M \geq 2$ microphones, where $\mathbf{m}_m \in \mathbb{R}^3$ denotes the absolute position of the m-th microphone. The aim is to estimate the MAG, i.e., the *relative* microphone coordinates $\mathbf{M}_{\text{rel}} = [\mathbf{m}_{\text{rel},1}, \mathbf{m}_{\text{rel},2}, \dots, \mathbf{m}_{\text{rel},M}]$, which are related to the absolute microphone positions by an arbitrary rotation, translation, and/or reflection. In the short-time Fourier transform (STFT) domain, the m-th microphone signal $X_m[k,l]$ consists of a direct speech component $X_{d,m}[k,l]$ and a reverberation component $X_{r,m}[k,l]$, i.e.,

$$X_m[k,l] = X_{d,m}[k,l] + X_{r,m}[k,l]\,, \tag{1}$$

where $k \in \{1,2,...,K\}$ denotes the frequency bin index and $l \in \{1,2,...,L\}$ denotes the time frame index. In vector notation, the corresponding M-dimensional microphone signal vector $\mathbf{x}[k,l] = [X_1[k,l], X_2[k,l], ..., X_M[k,l]]^{\mathrm{T}}$ can be written as

$$\mathbf{x}[k,l] = \mathbf{x}_d[k,l] + \mathbf{x}_r[k,l]\,. \tag{2}$$

The direct speech component vector $\mathbf{x}_d[k,l]$ can be related to the direct speech component in the first microphone $X_{d,1}[k,l]$ as

$$\mathbf{x}_d[k,l] = \mathbf{g}[k] X_{d,1}[k,l]\,, \tag{3}$$

with $\mathbf{g}[k] = [G_1[k], G_2[k], \dots, G_M[k]]^{\mathrm{T}}$ the relative direct transfer function (RDTF) vector. Assuming free-field transmission, i.e., no object or head between microphones, the elements of the RDTF vector are equal to

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project ID 352015383 (SFB 1330 B2) and Project ID 390895286 (EXC2177/1).

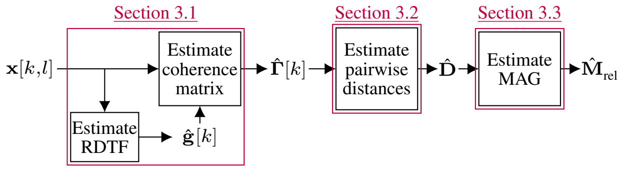


Figure 1: Block-diagram of the state-of-the-art MAG estimation.

$G_m[k] = \exp(-i2\pi\tau_m k/K)$, where the relative time delay between the m-th and the first microphone is defined as $\tau_m = (||\mathbf{q}-\mathbf{m}_m|| - ||\mathbf{q}-\mathbf{m}_1||)f_s/c$, with $\mathbf{q}$ the source position, f_s the sampling frequency, and c the speed of sound.

Assuming that the direct and reverberation components are uncorrelated, the reverberant speech covariance matrix $\mathbf{\Phi}_x[k,l] = \mathbb{E}\{\mathbf{x}[k,l]\mathbf{x}^{\text{H}}[k,l]\}$, with $\mathbb{E}\{\cdot\}$ the expectation operator, can be decomposed into a (rank-1) direct speech covariance matrix $\mathbf{\Phi}_d[k,l] = \mathbb{E}\{\mathbf{x}_d[k,l]\mathbf{x}_d^{\text{H}}[k,l]\}$ and a reverberation covariance matrix $\mathbf{\Phi}_r[k,l] = \mathbb{E}\{\mathbf{x}_r[k,l]\mathbf{x}_r^{\text{H}}[k,l]\}$. Assuming a homogeneous sound field for the reverberation component [9, 12, 13] with a time-invariant spatial coherence matrix $\mathbf{\Gamma}[k]$, the reverberant speech covariance matrix can be written as

$$\mathbf{\Phi}_x[k,l] = \underbrace{\phi_d[k,l]\mathbf{g}[k]\mathbf{g}^{\text{H}}[k]}_{\mathbf{\Phi}_d[k,l]} + \underbrace{\phi_r[k,l]\mathbf{\Gamma}[k]}_{\mathbf{\Phi}_r[k,l]}, \tag{4}$$

where $\phi_d[k,l] = \mathbb{E}\{|X_{d,1}[k,l]|^2\}$ denotes the time-varying direct speech PSD in the first microphone and $\phi_r[k,l]$ denotes the time-varying reverberation PSD. In this paper, the reverberant sound field is assumed to be spherically isotropic [14], such that the frequency-dependent coherence between the a-th and b-th microphone can be modeled as

$$\gamma(k,d_{a,b}) = \text{sinc}\left(\frac{2\pi f_s k d_{a,b}}{cK}\right), \tag{5}$$

where $d_{a,b}$ denotes the microphone distance.

3 MAG Estimation

In this section we describe the MAG estimation procedure proposed in [6], which consists of three steps (see Fig. 1). First, the coherence matrix is estimated from the reverberant microphone signals using the ECM approach (Section 3.1). From the estimated coherence matrix, the pairwise distances are then estimated (Section 3.2). Finally, the MAG is estimated from the estimated pairwise distances using multi-dimensional scaling (Section 3.3).

3.1 Coherence Estimation Using ECM

In this section, the frequency bin index k is omitted for conciseness, however, the parameter estimation is carried out in each frequency bin independently. Using ECM [7, 8], the time-invariant coherence matrix $\mathbf{\Gamma}$ and the time-varying speech and reverberation PSDs $\phi_d[l]$ and $\phi_r[l]$ are iteratively estimated given the reverberant speech signals $\mathbf{x}[l]$ and an estimate of the RDTF vector $\hat{\mathbf{g}}$. The procedure is derived in batch-mode, i.e., it is assumed that $\mathbf{x}[l]$ is available for all time-frames, i.e., $\mathbf{X} = [\mathbf{x}[1],\mathbf{x}[2],...,\mathbf{x}[L]]$. The batch direct speech PSD $\boldsymbol{\phi}_d$, and reverberation PSD $\boldsymbol{\phi}_r$ are defined similarly. The set of parameters to be estimated is defined as $\boldsymbol{\theta} = \{\boldsymbol{\phi}_d,\boldsymbol{\phi}_r,\mathbf{\Gamma}\}$. The speech and reverberation components are modeled as independent complex Gaussians with zero-mean and variance $\phi_d[l]$ and covariance matrix $\mathbf{\Phi}_r[l]$, respectively.

Assuming independent time-frames, the estimated parameter set $\hat{\boldsymbol{\theta}}^{(i)}$ is updated in each iteration i by increasing the expectation of the log-likelihood function (dependent on the probability density function $\psi(X_{d,1}[l],\mathbf{x}_r[l];\boldsymbol{\theta})$, defined as in [6] Eq. (14)), conditioned on the estimated parameters, i.e.,

$$Q\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}^{(i-1)}\right) = \mathbb{E}\left\{\sum_{l=1}^{L}\log\psi(X_{d,1}[l],\mathbf{x}_r[l];\boldsymbol{\theta})|\mathbf{X},\hat{\mathbf{g}};\hat{\boldsymbol{\theta}}^{(i-1)}\right\}. \tag{6}$$

In the expectation step (E-step), estimates of the direct speech PSD $\hat{\sigma}_d^{(i)}[l]$, the covariance matrix of the reverberant speech $\hat{\mathbf{\Phi}}_x^{(i)}[l]$ and of the reverberation $\hat{\mathbf{\Phi}}_r^{(i)}[l]$ are updated as

$$\hat{\mathbf{\Phi}}_x^{(i)}[l] = \hat{\phi}_d^{(i-1)}[l]\hat{\mathbf{g}}\hat{\mathbf{g}}^{\text{H}} + \hat{\phi}_r^{(i-1)}[l]\hat{\mathbf{\Gamma}}^{(i-1)}, \tag{7a}$$

$$\begin{aligned}\hat{\sigma}_d^{(i)}[l] = {} & |\hat{\phi}_d^{(i-1)}[l]\hat{\mathbf{g}}^{\text{H}}(\hat{\mathbf{\Phi}}_x^{(i)}[l])^{-1}\mathbf{x}[l]|^2 + ... \\ & \hat{\phi}_d^{(i-1)}[l]\left[\mathbf{I} - \hat{\phi}_d^{(i-1)}[l]\hat{\mathbf{g}}^{\text{H}}\hat{\mathbf{\Gamma}}^{(i-1)}\hat{\mathbf{g}}\right],\end{aligned} \tag{7b}$$

$$\hat{\mathbf{x}}_r^{(i)}[l] = \hat{\phi}_r^{(i-1)}[l]\hat{\mathbf{\Gamma}}^{(i-1)}(\hat{\mathbf{\Phi}}_x^{(i)}[l])^{-1}\mathbf{x}[l], \tag{7c}$$

$$\begin{aligned}\hat{\mathbf{\Phi}}_r^{(i)}[l] = {} & \hat{\mathbf{x}}_r^{(i)}[l](\hat{\mathbf{x}}_r^{(i)}[l])^{\text{H}} + \hat{\phi}_r^{(i-1)}[l]\,\hat{\mathbf{\Gamma}}^{(i-1)} ... \\ & \left[\mathbf{I} - (\hat{\mathbf{\Phi}}_x^{(i)}[l])^{-1}\hat{\mathbf{\Gamma}}^{(i-1)}\hat{\phi}_r^{(i-1)}[l]\right],\end{aligned} \tag{7d}$$

with $\mathbf{I}$ the $M \times M$-dimensional identity matrix. In the conditional-maximization step (CM-step), the parameters in $\boldsymbol{\theta}$ are estimated as

$$\hat{\phi}_d^{(i)}[l] = \hat{\sigma}_d^{(i)}[l], \tag{8a}$$

$$\hat{\mathbf{\Gamma}}^{(i)} = \frac{1}{L}\sum_{l=1}^{L}\frac{\hat{\mathbf{\Phi}}_r^{(i)}[l]}{\text{Tr}\{\hat{\mathbf{\Phi}}_r^{(i)}[l]\}}, \tag{8b}$$

$$\hat{\phi}_r^{(i)}[l] = \text{Tr}\{\hat{\mathbf{\Phi}}_r^{(i)}[l]\,(\hat{\mathbf{\Gamma}}^{(i)})^{-1}\}. \tag{8c}$$

with $\text{Tr}\{\cdot\}$ the trace-operator. The E- and CM-steps are iterated I times.

3.2 Pairwise Distance (PD) Estimation

The distance $d_{a,b}$ between microphones a and b is estimated as in [5, 6] by determining the distance d for which the coherence model in (5) is most similar to the estimated coherence $\hat{\Gamma}_{a,b}^{(I)}[k]$ in (8b) between microphones a and b, for a set $\mathscr{K}$ of frequency bins, i.e.,

$$\hat{d}_{a,b} = \underset{d}{\text{argmin}}\sum_{k\in\mathscr{K}}|\hat{\Gamma}_{a,b}^{(I)}[k] - \gamma(k,d)|^2. \tag{9}$$

3.3 Geometry Estimation

As shown in [5, 10], the $P \times M$-dimensional MAG (with $P = \min(M,3)$ in $\mathbb{R}^3$) can be inferred from the pairwise distances between the microphones with multi-dimensional scaling. Using the $M \times M$-dimensional Euclidean distance matrix, i.e., the matrix of squared distances, $\mathbf{D} = [d_{a,b}^2]$, a geometric centering operation $\mathbf{G} = -\frac{1}{2}(\mathbf{I} - \frac{1}{M}\mathbf{1}\mathbf{1}^{\text{T}})\mathbf{D}(\mathbf{I} - \frac{1}{M}\mathbf{1}\mathbf{1}^{\text{T}})$ is first applied, with $\mathbf{1}$ an M-dimensional vector of ones. Using the eigenvalue decomposition $\mathbf{G} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\text{T}}$, with $\mathbf{U}$ a matrix containing the eigenvectors and $\mathbf{\Lambda}$ a diagonal matrix containing P

non-negative eigenvalues $\lambda_1 \geq ... \geq \lambda_P \geq 0$, it can be shown that the MAG $\mathbf{M}_{\text{rel}}$ is related to the eigenvectors and eigenvalues as

$$\mathbf{M}_{\text{rel}} = [\text{diag}(\sqrt{\lambda_1},...,\sqrt{\lambda_P}), \mathbf{0}_{P\times(M-P)}]\mathbf{U}^{\text{T}}. \quad (10)$$

When estimating the MAG using the estimated Euclidean distance matrix $\hat{\mathbf{D}} = [\hat{d}^2_{a,b}]$, the corresponding $\hat{\mathbf{G}}$ might not be positive semi-definite and may have more than P non-zero eigenvalues due to estimation errors, so only the P largest positive eigenvalues are used in (10).

4 Data-based ECM Initialization

The ECM algorithm, described in Section 3.1, needs to be initialized using initial estimates of the coherence matrix $\hat{\mathbf{\Gamma}}^{(0)}[k]$ and the reverberation and direct speech PSDs $\hat{\phi}^{(0)}_r[k,l]$ and $\hat{\phi}^{(0)}_d[k,l]$. Similarly as in [15, 16], these initial PSDs are estimated as

$$\hat{\phi}^{(0)}_r[k,l] = \frac{1}{M-1}\mathbf{x}^{\text{H}}[k,l](\hat{\mathbf{\Gamma}}^{(0)}[k])^{-1}\left[\mathbf{I}-\hat{\mathbf{g}}[k]\mathbf{h}^{\text{H}}[k,l]\right]\mathbf{x}[k,l], \quad (11a)$$

$$\hat{\phi}^{(0)}_d[k,l] = \mathbf{h}^{\text{H}}[k,l]\left[\mathbf{x}[k,l]\mathbf{x}^{\text{H}}[k,l] - \hat{\phi}^{(0)}_r[k,l]\hat{\mathbf{\Gamma}}^{(0)}[k]\right]\mathbf{h}[k,l], \quad (11b)$$

with $\mathbf{h}[k,l]$ a beamformer.

In [6], a data-independent initial estimate of the coherence matrix $\hat{\mathbf{\Gamma}}^{(0)}[k]$ and a data-independent minimum-variance-distortionless-response (MVDR) beamformer $\mathbf{h}[k]$ were used, i.e.,

$$\hat{\mathbf{\Gamma}}^{(0)}_{\mathbf{I}}[k] = \mathbf{I}, \forall k, \quad (12a)$$

$$\mathbf{h}_{\text{MVDR}}[k] = \frac{(\hat{\mathbf{\Gamma}}^{(0)})^{-1}[k]\hat{\mathbf{g}}[k]}{\hat{\mathbf{g}}^{\text{H}}[k](\hat{\mathbf{\Gamma}}^{(0)})^{-1}[k]\hat{\mathbf{g}}[k]} = \frac{\hat{\mathbf{g}}[k]}{||\hat{\mathbf{g}}[k]||^2}, \quad (12b)$$

where the time-invariant MVDR beamformer in (12b) simplifies to a matched filter [17] since an uncorrelated sound field was assumed for the reverberation in (12a).

In this paper, we propose a data-dependent initialization of the coherence matrix estimate and beamformer to estimate the direct speech and reverberation PSDs. The initial estimate of the coherence matrix $\hat{\mathbf{\Gamma}}^{(0)}[k]$ is obtained by averaging the trace-normalized covariance matrix of the reverberant speech signals over all time frames, i.e.,

$$\boxed{\hat{\mathbf{\Gamma}}^{(0)}_x[k] = \frac{1}{L}\sum_{l=1}^{L}\frac{\mathbf{x}[k,l]\mathbf{x}^{\text{H}}[k,l]}{\frac{1}{M}\text{Tr}(\mathbf{x}[k,l]\mathbf{x}^{\text{H}}[k,l])}} \quad (13)$$

The data-dependent (but time-invariant) coherence matrix initialization in (13) could be used in (12b) to compute the beamformer $\mathbf{h}[k,l]$. However, to take into account the time-varying nature of the speech and reverberation, we propose to use a time-varying MPDR beamformer instead, i.e.,

$$\boxed{\mathbf{h}_{\text{MPDR}}[k,l] = \frac{(\hat{\mathbf{\Phi}}^{(0)}_x)^{-1}[k,l]\hat{\mathbf{g}}[k]}{\hat{\mathbf{g}}^{\text{H}}[k](\hat{\mathbf{\Phi}}^{(0)}_x)^{-1}[k,l]\hat{\mathbf{g}}[k]}} \quad (14)$$

This is motivated by the fact that unlike for the coherence matrix of the reverberation, an unbiased, time-varying estimate of the covariance matrix of the reverberant speech signals can be easily estimated as

$$\hat{\mathbf{\Phi}}^{(0)}_x[k,l] = \rho\,\hat{\mathbf{\Phi}}^{(0)}_x[k,l-1] + (1-\rho)\mathbf{x}[k,l]\mathbf{x}^{\text{H}}[k,l], \quad (15)$$

with ρ a recursive smoothing parameter. In speech enhancement applications, the MVDR beamformer is typically preferred over the MPDR beamformer since the MPDR beamformer results in speech distortion in case of estimation errors in the RDTF vector $\mathbf{g}[k]$ [18, 19]. However, in this application, avoiding target cancellation is not as important as accurately estimating the initial speech and reverberation PSDs.

5 Experimental Evaluation

In this section, we experimentally compare the influence of the proposed data-dependent initialization scheme on the PD estimation error for different frequency ranges (Section 5.2) and the MAG estimation error for different microphone array sizes (Section 5.3).

5.1 Scenario and Algorithm Parameters

For the simulations we considered a rectangular room with dimensions $6\times6\times2.4$ m and simulated room impulse responses (RIRs) using the image method [20, 21], assuming equal reflection coefficients for all walls. For each result shown, 50 acoustic scenarios were simulated, where one scenario consists of a unique combination of random source location, random microphone array location and geometry, and random 5s anechoic speech signal from [22], convolved with simulated RIRs. For each acoustic scenario, the reflection coefficient was chosen such that the direct-to-reverberant ratio in the first microphone (arbitrarily chosen) was approximately 0 dB.

The sampling frequency was equal to 16 kHz. For the STFT framework, the frame length was 512 samples (corresponding to 32 ms), with 50% overlap between frames, and a square-root-Hann analysis window was used. The recursive smoothing parameter in (15) was set to $\rho = 0.9$ and $I = 3$ ECM iterations were performed.

5.2 Frequency Range Sensitivity Analysis

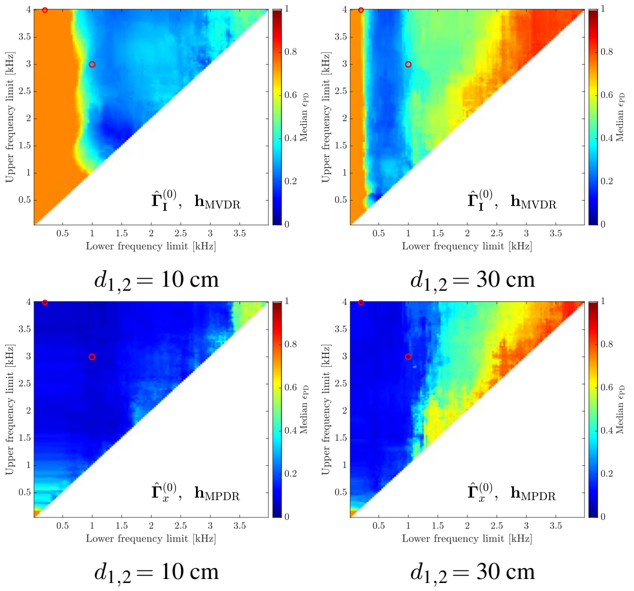
To analyse the influence of the frequency range $\mathscr{K}$ in (9) for both initialization schemes, we considered acoustic scenarios with $M = 2$ microphones and evaluated the normalized PD error, defined as

$$\epsilon_{\text{PD}} = \frac{|\hat{d}_{1,2} - d_{1,2}|}{d_{1,2}}, \quad (16)$$

with $\hat{d}_{1,2}$ the estimated distance between both microphones. For this analysis, the oracle RDTF vector $\mathbf{g}[k]$ was used (computed as in [23] using the anechoic RIR).

For two distances ($d_{1,2}$ = 10 cm, $d_{1,2}$ = 30 cm) and for both initialization schemes, Fig. 2 depicts the median normalized PD error for all scenarios, where the lower and upper frequency limits f_{low} and f_{up} in (9) were varied in the range 0 to 4 kHz.

For the data-independent initialization (top plots), it can be observed that the median PD error is large for low lower frequency limits (below 1 kHz for $d_{1,2}$ = 10 cm and below 300 Hz for $d_{1,2}$ = 30 cm) and high lower frequency limits (above 3 kHz for $d_{1,2}$ = 10 cm and above 1 kHz for $d_{1,2}$ = 30 cm) but does not depend very much on the upper frequency limit. By selecting the optimal frequency range, the minimum achievable median PD error was equal to 12% for $d_{1,2}$ = 10 cm (f_{low} = 1.55 kHz and f_{up} = 1.65 kHz) and 14% for $d_{1,2}$ = 30 cm (f_{low} = 450 Hz and f_{up} = 550



$d_{1,2} =$ 10 cm

$d_{1,2} =$ 30 cm

$d_{1,2} =$ 10 cm

$d_{1,2} =$ 30 cm

Figure 2: Median PD error between a pair of microphones for different frequency ranges (top plots: data-independent initialization, bottom plots: data-dependent initialization) The two red circles indicate the frequency ranges considered in Section 5.3.

Hz). It can be clearly observed that the frequency range (especially the lower frequency limit), for which a low median PD error is obtained, highly depends on the distance $d_{1,2}$ such that this frequency range would need to be tuned depending on the microphone array size, which is obviously not desired.

For the proposed data-dependent initialization (bottom plots), it can be observed that the median PD error is less sensitive to the selection of the lower and upper frequency limits, allowing a fixed frequency range to be used for estimating the geometry of different-sized microphone arrays. As well as reducing the frequency range sensitivity, the minimum achievable median PD error was reduced to 6% for $d_{1,2} = 10$ cm ($f_{\mathrm{low}} = 1.10$ kHz and $f_{\mathrm{up}} = 4$ kHz) and 8% for $d_{1,2} = 30$ cm ($f_{\mathrm{low}} = 400$ Hz and $f_{\mathrm{up}} = 3.80$ kHz). Although 1.1 kHz and 400 Hz constitute the optimal lower frequency limits for the tested microphone distances, it mainly seems to be important to capture the main-lobe of the coherence (in the case of a sinc-shaped coherence or similar), for which we select an upper frequency of 4 kHz in the following section.

5.3 MAG Estimation Error for Different Array Sizes

To investigate the MAG estimation performance for both initialization schemes, we considered 3-dimensional arrays with $M = 6$ microphones, randomly distributed within randomly positioned cubes with different cube lengths CL $\in \{10,20,\dots,50\}$ cm (for the largest cube length, PDs up to 87 cm are possible if microphones are located at opposite corners of the cube). To compute the MAG error, the estimated MAG $\hat{\mathbf{M}}_{\mathrm{rel}}$ in (10) was first aligned with the true coordinates $\mathbf{M}$ using the Procrustes analysis solution [1, 11, 24]. For each acoustic scenario, we evaluated the normalized MAG error per microphone, defined as

$$\epsilon_{\mathbf{m}_m} = \frac{||\hat{\mathbf{m}}_m - \mathbf{m}_m||_2}{\mathrm{CL}}. \qquad (17)$$

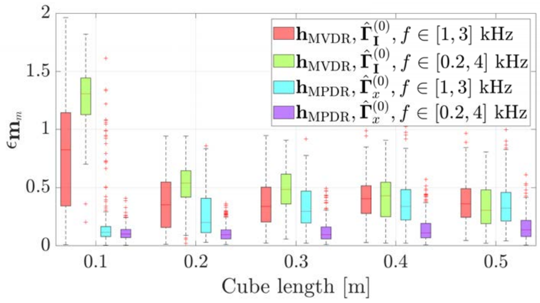


Figure 3: Normalized MAG error $\epsilon_{\mathbf{m}_m}$ vs. cube length for different ECM initializations and frequency ranges.

We compared the normalized MAG error for two frequency ranges: the frequency range $f \in [1,3]$ kHz, used in [6], and $f \in [0.2,4]$ kHz based on the results from the previous section and given that a lower frequency limit of 200 Hz generally leads to the smallest errors when using the data-dependent initialization, especially for increasing PDs. For these experiments, the RDTF vector $\mathbf{g}[k]$ was estimated using the generalized cross-correlation with phase transforms (GCC-PHAT) [25]. For the considered cube lengths and for both initialization schemes, Fig. 3 shows box plots of the normalized MAG error for all microphones and acoustic scenarios. For the data-independent initialization, it can be observed that the normalized MAG error is very high for CL = 10 cm (for both considered frequency ranges). For larger cube lengths, the normalized MAG error becomes smaller, where the median MAG error is in the range 15-55% for the frequency range [1,3] kHz and 50-65% for the frequency range [0.2, 4] kHz.

For the proposed data-dependent initialization, it can be observed that with a frequency range [1,3] kHz, the normalized MAG error is smaller than when using the data-independent initialization for CL $\leq$ 20 cm and similar for CL $>$ 20 cm. However, when selecting a more appropriate frequency range based on Fig. 2, i.e., [0.2, 4] kHz, the normalized MAG error is significantly smaller than when using data-independent initializations for all cube lengths. Even more importantly, the normalized MAG error is relatively constant for all considered array sizes, namely in the range 5-20%. This corresponds with the observation in Fig. 2 that the selection of the frequency range for the data-dependent initialization is far less dependent on the microphone distance than for the data-independent initialization. The significantly reduced normalized MAG error and the improved generalizability for different array sizes shows the importance of properly initializing the estimated parameters in ECM.

6 Conclusions

Aiming at improving the MAG estimation performance using an iterative ECM approach, in this paper we proposed a data-dependent initialization of the coherence matrix estimate and the beamformer for estimating the speech and reverberation PSDs. Simulation results showed that the proposed data-dependent initialization significantly reduced the PD estimation error compared to a data-independent initialization and allowed to select the frequency range for PD estimation independently of the microphone distance. In addition, the data-dependent ECM initialization resulted in a significantly reduced MAG estimation error for all considered array sizes.

References

[1] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. Fink, “Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms,” *IEEE Signal Process. Mag.*, vol. 33, no. 4, pp. 14–29, 2016.

[2] M. Pollefeys and D. Nister, “Direct computation of sound and microphone locations from time-difference-of-arrival data,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (Las Vegas, Nevada, USA), pp. 2445–2448, 2008.

[3] N. Gaubitch, B. Kleijn, and R. Heusdens, “Auto-localization in ad-hoc microphone arrays,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, (Vancouver, BC, Canada), pp. 106–110, 2013.

[4] B. Champagne, S. Bédard, and A. Stéphenne, “Performance of time-delay estimation in the presence of room reverberation,” *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 148–152, 1996.

[5] I. McCowan, M. Lincoln, and I. Himawan, “Microphone array shape calibration in diffuse noise fields,” *Proc. IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 3, pp. 666–670, 2008.

[6] O. Schwartz, A. Plinge, E. Habets, and S. Gannot, “Blind microphone geometry calibration using one reverberant speech event,” in *Proc. IEEE Workshop Appl. Signal Process. Audio, Acoust.*, pp. 131–135, 2017.

[7] X.-L. Meng and D. B. Rubin, “Maximum likelihood estimation via the ECM algorithm: A general framework,” *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[8] O. Schwartz, S. Gannot, and E. Habets, “An expectation-maximization algorithm for multimicrophone speech dereverberation and noise reduction with coherence matrix estimation,” *Proc. IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1495–1510, 2016.

[9] E. Habets and S. Gannot, “Generating sensor signals in isotropic noise fields,” *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, 2007.

[10] S. T. Birchfield, “Geometric microphone array calibration by multidimensional scaling,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, (Hong Kong (cancelled)), pp. V–157–160, 2003.

[11] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, “Euclidean distance matrices: essential theory, algorithms, and applications,” *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 12–30, 2015.

[12] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A consolidated perspective on multimicrophone speech enhancement and source separation,” *Proc. IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, 2017.

[13] I. Kodrasi and S. Doclo, “Late reverberant power spectral density estimation based on an eigenvalue decomposition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 611–615, 2017.

[14] R. Cook, R. Waterhouse, R. Berendt, S. Edelman, and M. Thompson Jr, “Measurement of correlation coefficients in reverberant sound fields,” *J. Acoust. Soc. Amer.*, vol. 27, no. 6, pp. 1072–1077, 1955.

[15] H. Ye and R. DeGroat, “Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise,” *IEEE Trans. Speech Audio Process.*, vol. 43, no. 4, pp. 938–949, 1995.

[16] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, “Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids,” in *Proc. IEEE Europ. Signal Process. Conf.*, (Lisbon, Portugal), pp. 61–65, 2014.

[17] E.-E. Jan and J. Flanagan, “Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, (Atlanta, Georgia, USA), pp. 917–920, 1996.

[18] B. Van Veen and K. Buckley, “Beamforming: A versatile approach to spatial filtering,” *ASSP*, vol. 5, no. 2, pp. 4–24, 1988.

[19] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, “Sensitivity analysis of MVDR and MPDR beamformers,” in *Proc. IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, (Eliat, Israel), pp. 416–420, 2010.

[20] E. Habets, *RIR-Generator*. Available at `https://github.com/ehabets/RIR-Generator`.

[21] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[22] P. Kabal, *TSP Speech Database*. Available at `http://www-mmsp.ece.mcgill.ca/Documents/Data/`.

[23] N. Gößling and S. Doclo, “Relative transfer function estimation exploiting spatially separated microphones in a diffuse noise field,” in *Proc. IEEE Int. Workshop Acoustic Echo, Noise Control*, (Tokyo, Japan), pp. 146–150, 2018.

[24] P. Schoenemann, *A solution of the orthogonal Procrustes problem with applications to orthogonal and oblique rotation*. PhD thesis, University of Illinois at Urbana-Champaign, 1964.

[25] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *Proc. IEEE Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 320–327, 1976.