IEEE Open Journal of Signal Processing

Received 5 September 2023; revised 22 November 2023; accepted 26 November 2023. Date of publication 29 November 2023; date of current version 29 December 2023. The review of this article was arranged by Associate Editor Ang Li.

Digital Object Identifier 10.1109/OJSP.2023.3337721

A Flexible Framework for Expectation Maximization-Based MIMO System Identification for Time-Variant Linear Acoustic Systems

TOBIAS KABZINSKI (Graduate Student Member, IEEE), AND PETER JAX (Member, IEEE)

Institute of Communication Systems (IKS), RWTH Aachen University, 52056 Aachen, Germany

CORRESPONDING AUTHOR: TOBIAS KABZINSKI (e-mail: kabzinski@iks.rwth-aachen.de)

This work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant 509806277.

ABSTRACT Quasi-continuous system identification of time-variant linear acoustic systems can be applied in various audio signal processing applications when numerous acoustic transfer functions must be measured. A prominent application is measuring head-related transfer functions. We treat the underlying multiple-input-multiple-output (MIMO) system identification problem in a state-space model as a joint estimation problem for states, representing impulse responses, and state-space model parameters using the expectation maximization (EM) algorithm. We address limitations of prior work by imposing different model structures, especially for dependencies within a (transformed) state vector. This results in block diagonal matrix structures, for which we derive M-step update rules. Making assumptions about this model structure and choosing a block size for a given application define the computational complexity. In examples, we found that applying this framework yields improvements of up to 10 dB in relative system distance in comparison to a conventional method.

INDEX TERMS Expectation maximization, system identification, state-space model.

I. INTRODUCTION

Identifying acoustic systems is required in many audio signal processing applications. Characterizing linear acoustic systems or transfer paths using measurements is an essential step before designing filters for applications, such as feed-forward active noise control (ANC) [1], personal sound zones [2], crosstalk cancellation [3], or when measuring head-related transfer functions (HRTFs) for binaural synthesis [4], [5]. In these measurements, control over the playback signal, which is fed into the system, can be assumed. In some applications, such as acoustic echo control (AEC), it is required to identify acoustic systems at runtime, and the playback signal cannot be controlled; these applications are not the primary focus of this contribution.

Two main approaches for measurements of linear acoustic systems exist. On the one hand, in static measurements acoustic impulse responses (IRs) are measured for one specific configuration between transmitter(s) and receiver(s). This type of measurement is often conducted using exponential sweeps [6], [7], [8], [9]. On the other hand, quasi-continuous measurements aim at identifying time-variant acoustic systems that slowly change over time. This offers the possibility to measure numerous spatial configurations between transmitter(s) and receiver(s) in a short time, and to simulate time-variant systems, as in [10], [11].

The quasi-continuous measurements, especially of HRTFs, have become popular due to a reduced measurement duration [5], [12], [13]. In this application, one or more loudspeakers reproduce a predefined signal while microphones in the ear canals of a subject capture signals that are filtered with the desired acoustic system's responses. To obtain a spatially dense grid of head-related impulse responses (HRIRs) the subject is either rotated on a turntable, as in [14], [15], [16], [17], [18], [19], or can move freely [20], [21], [22], [23], [24]. Adaptive filtering method, such as the normalized least-mean-square (NLMS) algorithm or variants thereof are often applied to estimate the IRs. The NLMS algorithm can also be related to deconvolution used in static measurements [25].

Some other adaptive filtering methods can be interpreted as state estimation techniques [26] to obtain IRs as state estimates, as, for example, applied in the time domain for ANC in [27] or in the discrete Fourier transform (DFT) domain for AEC in [26]. For the latter application, a variant of the expectation maximization (EM) algorithm has been adopted to learn the measurement noise covariance and process noise covariance for a DFT-domain adaptive filter [28]. While [28] focuses on online processing, which is required in AEC, in measurement applications, however, online filtering poses an unnecessarily strict requirement. In constrast to adaptive filtering, offline processing or algorithms that require a large lookahead can be applied without restrictions once recording the signals is completed, as in [29]. In [30], the EM algorithm has been applied to identify time-invariant multiple-input-multiple-output (MIMO) systems in the frequency domain. We proposed to apply the EM algorithm offline to estimate time-variant IRs in HRTF measurements, and to learn the parameters of the corresponding state-space model independently on overlapping sequences [31]. While the results indicate a large potential, the applicability of this approach to real-world measurements with higher sampling rates, longer IRs and more loudspeakers in parallel is limited due to the excessive computational complexity and memory demand.

In this contribution, we present a flexible framework which applies EM-based joint learning of state-space model parameters and IRs for application in quasi-continuous system identification of time-variant linear acoustic systems so that the limitations of [31] can be overcome. This flexible framework allows to estimate time-domain finite impulse response (FIR) coefficients, a DFT-domain state representation, or any other transformed state representation obtained from a linear invertible transform. Moreover, the proposed framework employs blockwise processing and allows to treat multiple microphones jointly in a MIMO system [32]. Imposing different coupling models for dependencies within the state vector can yield block diagonal matrix structures with various levels of computational complexity. As a result, a wider range of applications can be tackled, e.g., measuring time-variant IRs of ANC headphones, as conducted in [33] for a single reference microphone, could be improved and extended to multi-microphone setups, as discussed in [1].

This contribution is structured as follows: Section II presents the state-space system model while Section III describes the EM algorithm for the joint estimation of states and parameters for this model. In Sections IV and V, we present specialized model structures and derive variants of M-step update rules for various models. Examples are provided in Section VI, and Section VII concludes the paper.

II. STATE-SPACE MODEL

Quasi-continuous MIMO system identification aims at estimating IRs of an acoustic system with very high temporal resolution, resulting in either one estimate per sample or per every few milliseconds. We assume that T transmitters (loudspeakers) simultaneously play back signals $x_t(k)$, $t = 1, \ldots, T$, such that R receivers (microphones) receive signals $y_r(k)$, $r = 1, \ldots, R$. The IR between receiver r and transmitter t at time k is modeled as an FIR filter $h_{rt,k}(\ell)$.

To completely represent this linear acoustic MIMO system at time k, we define the state vector at time k, comprising the IRs valid at this time instance, as

$$\mathbf{z}_k = \begin{bmatrix} \mathbf{h}_{11,k}^{\mathrm{T}} & \ldots & \mathbf{h}_{1T,k}^{\mathrm{T}} & \ldots & \mathbf{h}_{RT,k}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{RTL}, \tag{1}$$

where $\mathbf{h}_{rt,k} = [h_{rt,k}(0), \ldots, h_{rt,k}(L-1)]^{\mathrm{T}}$ is a length-L coefficient vector for each IR $h_{rt,k}(\ell)$. The order of the state-space system is given by the number of states $N_z = RTL$. To describe the state-space model, we define the state equation, now using a frame index n, similarly to [31], as

$$\mathbf{z}_n = \mathbf{A}\mathbf{z}_{n-1} + \mathbf{q}_n. \tag{2}$$

(2) describes the evolution of the IRs over time. Here, $\mathbf{A}$ is the state transition matrix, and $\mathbf{q}_n$ is the process noise, which is assumed to be zero-mean Gaussian with covariance $\mathbf{\Gamma}$.

The observation equation describes how the current state, the IRs, relate to the observations, i.e., the recorded signal samples. We extend the observation model from the block time-domain Kalman filter in [34] such that multiple receivers are considered jointly in a multiple-output system. This leads to the observation equation

$$\mathbf{y}_n = \begin{bmatrix} \mathbf{y}_{1,n}^{\mathrm{T}} & \ldots & \mathbf{y}_{R,n}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} = \mathbf{C}_n\mathbf{z}_n + \boldsymbol{\nu}_n \in \mathbb{R}^{RN_o}, \tag{3}$$

where N_o is the number of samples that form an observation vector $\mathbf{y}_{r,n} = [y_r(nN_o - 1), \ldots, y_r(nN_o - N_o)]^{\mathrm{T}} \in \mathbb{R}^{N_o}$ for receiver r. $\boldsymbol{\nu}_n$ models additive zero-mean Gaussian measurement noise that is assumed to have covariance $\mathbf{\Sigma}$. The frame index n is a time index, similar to k, that is temporally down-sampled by a factor of N_o, and it ranges from $n = 1, \ldots, N$. The choice of N_o allows to control how many received signal samples form one observation, and hence defines the temporal resolution of the changes modeled by the state equation (2) and of the IR estimates. The observation matrix $\mathbf{C}_n$, relating IRs and the recorded signal samples, implements the convolution between the IRs and the playback signals. The vector $\mathbf{x}_{t,k} = [x_t(k), \ldots, x_t(k-L+1)]^{\mathrm{T}}$ contains the L most recent samples of x_t at time k. This allows to write the observation matrix as

$$\mathbf{C}_n = \begin{bmatrix} \mathbf{C}_{11,n} & \ldots & \mathbf{C}_{1T,n} & \mathbf{0} & \ldots & \mathbf{0} \\ & & & \ddots & & \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{C}_{R1,n} & \ldots & \mathbf{C}_{RT,n} \end{bmatrix} \in \mathbb{R}^{RN_o \times RTL},$$

where $\mathbf{C}_{rt,n} = \begin{bmatrix} \mathbf{x}_{t,n\cdot N_o - 1} & \ldots & \mathbf{x}_{t,n\cdot N_o - N_o} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{N_o \times L}$. Note that $\mathbf{C}_n$ is a block diagonal matrix with R blocks. (2) and (3) completely specify the state-space model with appropriate model parameters $\mathbf{A}$, $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$.

III. EM-BASED STATE AND PARAMETER ESTIMATION

Given a sequence of observations $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, we want to jointly estimate the sequence of hidden states, i.e., the IRs, $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ and the set of state-space model parameters $\boldsymbol{\theta} = \{\mathbf{A}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \mathbf{P}_0\}$. Here $\boldsymbol{\mu}_0$ is the initial mean state vector and $\mathbf{P}_0$ is the associated initial *a priori* state covariance matrix. The EM algorithm for state-space models [35] conducts this joint optimization by maximizing the expected log-likelihood, which is given by

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}) = &-\frac{1}{2}\log\left(\det\left(\mathbf{P}_0\right)\right) - \frac{N-1}{2}\log\left(\det\left(\boldsymbol{\Gamma}\right)\right)\\ &-\mathbb{E}\left\{\frac{1}{2}\left(\mathbf{z}_1-\boldsymbol{\mu}_0\right)^{\mathrm{T}}\mathbf{P}_0^{-1}\left(\mathbf{z}_1-\boldsymbol{\mu}_0\right)\right\}\\ &-\mathbb{E}\left\{\frac{1}{2}\sum_{n=2}^{N}\left(\mathbf{z}_n-\mathbf{A}\mathbf{z}_{n-1}\right)^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\left(\mathbf{z}_n-\mathbf{A}\mathbf{z}_{n-1}\right)\right\}\\ &-\mathbb{E}\left\{\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{y}_n-\mathbf{C}_n\mathbf{z}_n\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}_n-\mathbf{C}_n\mathbf{z}_n\right)\right\}\\ &-\frac{N}{2}\log\left(\det\left(\boldsymbol{\Sigma}\right)\right)+\text{const.}\end{aligned} \tag{4}$$

for the above state-space model. Here $\det(\cdot)$ denotes the determinant of a matrix and $\mathbb{E}\{\cdot\}$ is the expectation operator. Note that, in contrast to [35], the observation matrix is time-dependent, as in [36], but defined by the playback signals.

The EM algorithm iterates between E-step and M-step to find a locally optimal solution for both state and parameter estimates. The E-step calculates the maximum-likelihood state estimates for a fixed set of parameters $\boldsymbol{\theta}$, and is given by the recursive Kalman filtering and Kalman smoothing equations, which involve the following quantities: the *a priori* state covariance matrix $\mathbf{P}_n$, the Kalman gain $\mathbf{K}_n$, the filtered state estimate $\boldsymbol{\mu}_n$, the *a posteriori* state covariance matrix $\mathbf{V}_n$, the Kalman smoother gain $\mathbf{J}_n$, the smoothed state estimate $\hat{\boldsymbol{\mu}}_n$, and the smoothed *a posteriori* state covariance matrix $\hat{\mathbf{V}}_n$. The Kalman filtering equations [35] are

$$\mathbf{P}_{n-1} = \begin{cases}\mathbf{P}_0 & \text{if } n=1,\\ \mathbf{A}\mathbf{V}_{n-1}\mathbf{A}^{\mathrm{T}}+\boldsymbol{\Gamma} & \text{otherwise,}\end{cases} \tag{5a}$$

$$\mathbf{K}_n = \mathbf{P}_{n-1}\mathbf{C}_n^{\mathrm{T}}\left(\mathbf{C}_n\mathbf{P}_{n-1}\mathbf{C}_n^{\mathrm{T}}+\boldsymbol{\Sigma}\right)^{-1}, \tag{5b}$$

$$\boldsymbol{\mu}_n = \begin{cases}\boldsymbol{\mu}_0+\mathbf{K}_1\left(\mathbf{y}_1-\mathbf{C}_1\boldsymbol{\mu}_0\right) & \text{if } n=1,\\ \mathbf{A}\boldsymbol{\mu}_{n-1}+\mathbf{K}_n\left(\mathbf{y}_n-\mathbf{C}_n\mathbf{A}\boldsymbol{\mu}_{n-1}\right) & \text{otherwise,}\end{cases} \tag{5c}$$

$$\mathbf{V}_n = \left(\mathbf{I}_{N_z}-\mathbf{K}_n\mathbf{C}_n\right)\mathbf{P}_{n-1}, \tag{5d}$$

with the $N_z \times N_z$ identity matrix $\mathbf{I}_{N_z}$, and they are evaluated recursively for $n = 1, \ldots, N$. The Kalman smoother equations [35]

$$\mathbf{J}_n = \mathbf{V}_n\mathbf{A}^{\mathrm{T}}\mathbf{P}_n^{-1}, \tag{6a}$$

$$\hat{\boldsymbol{\mu}}_n = \begin{cases}\boldsymbol{\mu}_n & \text{if } n=N,\\ \boldsymbol{\mu}_n+\mathbf{J}_n\left(\hat{\boldsymbol{\mu}}_{n+1}-\mathbf{A}\boldsymbol{\mu}_n\right) & \text{otherwise,}\end{cases} \tag{6b}$$

$$\hat{\mathbf{V}}_n = \begin{cases}\mathbf{V}_n & \text{if } n=N,\\ \mathbf{V}_n+\mathbf{J}_n\left(\hat{\mathbf{V}}_{n+1}-\mathbf{P}_n\right)\mathbf{J}_n^{\mathrm{T}} & \text{otherwise,}\end{cases} \tag{6c}$$

are evaluated for $n = N, \ldots, 1$.

Then, the M-step updates the parameters $\boldsymbol{\theta}$ for the fixed set of estimates (from the E-step). Assuming the above state-space model and that all matrices are fully populated, the M-step update equations yield the updated quantities, which are highlighted by $\star$, as [35]

$$\mathbf{A}^{\star} = \left(\sum_{n=2}^{N}\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right)\left(\sum_{n=2}^{N}\mathbb{E}\left\{\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right)^{-1}, \tag{7}$$

$$\boldsymbol{\Gamma}^{\star} = \frac{1}{N-1}\sum_{n=2}^{N}\mathcal{G}_n, \tag{8}$$

$$\boldsymbol{\Sigma}^{\star} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{M}_n, \tag{9}$$

$$\boldsymbol{\mu}_0^{\star} = \hat{\boldsymbol{\mu}}_1, \tag{10}$$

$$\mathbf{P}_0^{\star} = \hat{\mathbf{V}}_1 \tag{11}$$

with the auxiliary definitions

$$\begin{aligned}\mathcal{G}_n = &\,\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right\} - \mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\mathbf{A}^{\star\mathrm{T}}\\ &-\mathbf{A}^{\star}\mathbb{E}\left\{\mathbf{z}_{n-1}\mathbf{z}_n^{\mathrm{T}}\right\}+\mathbf{A}^{\star}\mathbb{E}\left\{\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\mathbf{A}^{\star\mathrm{T}},\end{aligned} \tag{12}$$

$$\begin{aligned}\mathcal{M}_n = &\,\mathbf{y}_n\mathbf{y}_n^{\mathrm{T}} - \mathbf{y}_n\mathbb{E}\left\{\mathbf{z}_n^{\mathrm{T}}\right\}\mathbf{C}_n^{\mathrm{T}}\\ &-\mathbf{C}_n\mathbb{E}\left\{\mathbf{z}_n\right\}\mathbf{y}_n^{\mathrm{T}}+\mathbf{C}_n\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right\}\mathbf{C}_n^{\mathrm{T}}.\end{aligned} \tag{13}$$

Evaluating (7) to (9) requires the following equations [35]:

$$\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right\} = \hat{\mathbf{V}}_n+\hat{\boldsymbol{\mu}}_n\hat{\boldsymbol{\mu}}_n^{\mathrm{T}}, \tag{14}$$

$$\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right\} = \hat{\mathbf{V}}_n\mathbf{J}_{n-1}^{\mathrm{T}}+\hat{\boldsymbol{\mu}}_n\hat{\boldsymbol{\mu}}_{n-1}^{\mathrm{T}}. \tag{15}$$

For each sequence $\mathbf{Y}$, $\mathcal{I}$ EM iterations are conducted. These sequences are overlapping signal segments of the entire recorded signal, and each sequence consists of a lookback part, a central part and lookahead part, as in [31]. To finally provide IR estimates for each N_o-th sample of the entire recorded signal, the estimates, obtained independently on each sequence, are combined [31].

As the state and parameter estimates resulting from the iterative EM algorithm depend on the choice of initial parameters [35], we provide a rule of thumb on how to choose an initial set of parameters $\boldsymbol{\theta}^{(0)}$. As a result of offline processing conducted in measurements, we can assume that preliminary estimates of the IRs (states) can be obtained for a given sequence, e.g., using the NLMS algorithm. These states shall be denoted as $\mathring{\mathbf{z}}_n$, $n = 1, \ldots, N$ corresponding to time samples $k = 1 \cdot N_{\mathrm{o}} - 1, \ldots, N \cdot N_{\mathrm{o}} - 1$. Then, we can

TABLE 1. Coupling Models for State Vector With N_{B_z} Blocks of Size B_z. The Indexing Order Can be Modified by Means of a Permutation Matrix $\mathcal{P}$. The Default Slow-to-Fast Indexing Order, as in (1), is Receiver Index r (Changes Slowest), Transmitter Index t, Coefficient Index ℓ (Changes Fastest), and It is Denoted as r-t-ℓ. For Complex-Valued Transforms There are $L/2$ Coefficients, Which are Also Represented by Index ℓ. Additionally, ζ Represents the Index Required to Differentiate Between Real and Imaginary Part of Complex Coefficients. Star $*$ Represents Arbitrary Indexing Order in the Remaining Indices

ID	Description	N_{B_z}	B_z	Slow-to-fast indexing order
1	full	1	RTL	any
2	independent receivers	R	TL	r-t-ℓ or r-ℓ-t
3	independent transmitters	T	RL	t-r-ℓ or t-ℓ-r
4	independent coefficients	L	RT	ℓ-t-r or ℓ-r-t
5	within-receiver coupling	TL	R	t-ℓ-r or ℓ-t-r
6	within-transmitter coupling	RL	T	r-ℓ-t or ℓ-r-t
7	independent IRs (within-coefficients coupling)	RT	L	r-t-ℓ or t-r-ℓ
8	fully independent	RTL	1	any
9	complex-valued: independent coefficients	$RTL/2$	2	$*$-ζ
10	complex-valued: within-receiver coupling	$TL/2$	$2R$	$*$-r-ζ or $*$-ζ-r
11	complex-valued: within-transmitter coupling	$RL/2$	$2T$	$*$-t-ζ or $*$-ζ-t
12	complex-valued: within-frequency coupling	$L/2$	$2RT$	ℓ-$*$

compute the initial mean state vector and the initial *a priori* state covariance matrix as the maximum-likelihood estimates of mean and covariance of these preliminary IRs as $\boldsymbol{\mu}_0^{(0)} = N^{-1}\sum_{n=1}^{N}\mathring{\mathbf{z}}_n$ and $\mathbf{P}_0^{(0)} = N^{-1}\sum_{n=1}^{N}(\mathring{\mathbf{z}}_n - \boldsymbol{\mu}_0^{(0)})(\mathring{\mathbf{z}}_n - \boldsymbol{\mu}_0^{(0)})^{\mathrm{T}}$, respectively. Similarly, with the initial assumption $\mathbf{A}^{(0)} = \mathbf{I}$, we can obtain a maximum-likelihood estimate of the process noise covariance from realizations of the process noise $\mathring{\mathbf{q}}_n = \mathring{\mathbf{z}}_n - \mathring{\mathbf{z}}_{n-1}$ corresponding to the preliminary state estimates, similar to (2), as $\boldsymbol{\Gamma}^{(0)} = (N-1)^{-1}\sum_{n=2}^{N}(\mathring{\mathbf{q}}_n - \bar{\mathring{\mathbf{q}}})(\mathring{\mathbf{q}}_n - \bar{\mathring{\mathbf{q}}})^{\mathrm{T}}$, with the mean process noise corresponding to the preliminary state estimates given by $\bar{\mathring{\mathbf{q}}} = (N-1)^{-1}\sum_{n=2}^{N}\mathring{\mathbf{z}}_n$. An initial measurement noise covariance matrix $\boldsymbol{\Sigma}^{(0)}$ can similarly be obtained from a background noise recording.

IV. SPECIALIZED MODEL STRUCTURES

The state and parameter estimation described above is similar to the one in [31], except that we consider a MIMO system instead of a multiple-input-single-output (MISO) system and that we use blockwise processing. If the IRs to be estimated are relatively long and/or the number of transmitters and/or receivers is high, the number of states N_z becomes large. As a result, the computational complexity and memory demand of the E-step and the M-step can become very large. To reduce the number of model parameters and the computational complexity, we propose to impose specific structures for the matrix-valued parameters $\mathbf{A}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Sigma}$, and $\mathbf{P}_0$. For example, imposing a (block) diagonal structure on the covariance matrices implies that the cross-covariances between particular states are zero, i.e., the states are assumed to be statistically independent.

A. MOTIVATIONAL EXAMPLES

The first example corresponds to combining two decoupled MISO systems: The state covariance matrices $\mathbf{P}_n$, $\mathbf{V}_n$, $\hat{\mathbf{V}}_n$, and the process noise covariance matrix $\boldsymbol{\Gamma}$ are then set up as block diagonal matrices with R blocks of size $TL \times TL$ such that the IRs to each receiver are modeled as independent, but the cross-covariances within the TL samples of the T IRs of length L are considered (cf. Table 1, ID 2). The state-transition matrix should also reflect this block diagonal structure.

Second example: Correlation between IRs to the different receivers can be assumed if they are physically close to each other, especially in freefield-like conditions. The direct path to a linear array with closely-spaced microphones for low frequencies can be expected to change in a similar fashion for adjacent microphones. In case of HRTF measurements, the two microphones in the subject's ear are physically connected through the head and jointly move. Then, dependencies between receivers for a specific frequency could be modeled by applying a frequency-domain transform and a permutation such that the coefficients corresponding to a given frequency appear as groups in the state vector. This requires to reorder the state vector. In (1) the coefficient index changes fastest and the receiver index changes slowest. To form the groups in this example that model within-receiver coupling for a fixed frequency, the receiver index should change fastest. This would result in $TL/2$ blocks of size $2R \times 2R$ (cf. Table 1, ID 10) for two coefficients (real and imaginary part) per DFT bin. By considering the cross-covariance between real and imaginary parts of a DFT bin, each DFT coefficient is modeled as an improper complex Gaussian random variable [37]. The DC and the Nyquist bin (for even L) are considered jointly in one 2×2 block.

The third example considers distance changes between receiver(s) and transmitter(s). Cyclically shifting and scaling an IR can be represented as a complex-valued pointwise multiplication in the DFT domain. This can be achieved by applying the DFT to the state vector and by using a state transition matrix with 2×2 blocks to implement the complex-valued multiplications. Depending on the application, a meaningful coupling model can be chosen. See Table 1 for a list of suggested coupling models.

B. FURTHER APPROXIMATIONS AND STATE TRANSFORM

When independent blocks in the state are assumed, the block diagonal structure is preserved through the recursive Kalman filtering and Kalman smoothing equations, except in (5d), (14) and (15). Therefore, we suggest the following approximations that maintain the block diagonal structure:

$$\mathbf{V}_n \approx \left(\mathbf{I}_{N_z} - \text{blkdiag}\left\{\mathbf{K}_n\mathbf{C}_n\right\}\right)\mathbf{P}_{n-1}, \tag{16}$$

$$\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right\} \approx \hat{\mathbf{V}}_n + \text{blkdiag}\left\{\hat{\boldsymbol{\mu}}_n\hat{\boldsymbol{\mu}}_n^{\mathrm{T}}\right\}, \tag{17}$$

$$\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right\} \approx \hat{\mathbf{V}}_n\mathbf{J}_{n-1}^{\mathrm{T}} + \text{blkdiag}\left\{\hat{\boldsymbol{\mu}}_n\hat{\boldsymbol{\mu}}_{n-1}^{\mathrm{T}}\right\}. \tag{18}$$

Here, blkdiag$\{\cdot\}$ extracts a block diagonal matrix from a matrix. The block size should be clear from context. (16) is structurally similar to the covariance update in the subdiagonal DFT-domain MISO Kalman filter [38] and in the time-domain broadband Kalman filter [39]. It is worth noting that using these approximations it is no longer guaranteed that the expected log-likelihood does not decrease—in contrast to the regular EM algorithm, which provides this guarantee [36]. Yet, we found that these approximations yield useful results.

Instead of using a time-domain FIR coefficient representation of the IRs in the state vector, we can apply any linear transform to represent the IR coefficients $\mathbf{h}_{rt,k}$ in a different basis, such as real and imaginary parts of an L-point DFT. This transform, applied to each IR separately, shall be given by an invertible matrix $\mathbf{T}$ of size $L \times L$. To implement the reordering of the states according to the indexing order in Table 1, a permutation matrix $\mathcal{P}$ is introduced. The permuted and transformed state can then be described as

$$\tilde{\mathbf{z}}_n = \mathcal{P}\left(\mathbf{I}_{RT} \otimes \mathbf{T}\right)\mathbf{z}_n = \mathbf{W}\mathbf{z}_n. \tag{19}$$

Here, $\otimes$ denotes the Kronecker product. As the observation vectors remain in the time domain, the observation matrix that would be multiplied with the transformed state vector from the right is given by $\tilde{\mathbf{C}}_n = \mathbf{C}_n\mathbf{W}^{-1}$. When the state dimension N_z becomes large, it can be advantageous to exploit that $\mathbf{W}^{-1} = (\mathbf{I}_{RT} \otimes \mathbf{T}^{-1})\mathcal{P}^{\mathrm{T}}$ can be calculated more efficiently than simply inverting the large matrix $\mathbf{W}$.

If a transform-domain state representation is considered in the EM algorithm, the initial values can be transformed using the relations $\tilde{\mathbf{A}} = \mathbf{W}\mathbf{A}\mathbf{W}^{-1}$ and $\tilde{\mathbf{\Gamma}} = \mathbf{W}\mathbf{\Gamma}\mathbf{W}^{-1}$ and the IR estimates can be reconstructed using (19).

C. COMPUTATIONAL COMPLEXITY AND MEMORY REQUIREMENTS

The dominant term of the computational complexity of one EM iteration in Section III stems from the matrix-matrix multiplications of $N_z \times N_z$ matrices. For a signal sequence of length $N_x = N \cdot N_o$, the number of operations is in the order of $\mathcal{O}(\frac{N_x}{N_o}N_z^3)$. Assuming that the matrices $\mathbf{A}$, $\mathbf{\Gamma}$, $\mathbf{\Sigma}$, and $\mathbf{P}_0$ are block diagonal with N_{B_z} blocks of size $B_z \times B_z$, the number of operations for the above matrix-matrix multiplications reduces to $\mathcal{O}(\frac{N_x}{N_o}N_{B_z}B_z^3)$.[1] Further, it is required to store $\mathcal{O}(\frac{N_x}{N_o}N_z^2)$ elements of the *a posteriori* state covariance matrices in Section III. This number reduces to $\mathcal{O}(\frac{N_x}{N_o}N_{B_z}B_z^2)$ for block diagonal matrices.

Note that both the computational complexity terms above and the memory requirements are inversely proportional to the block size N_o. Therefore, assuming constant IRs for several time instances in the observation model in (3) can result in large savings compared to [31] where $N_o = 1$.

V. DERIVED M-STEP UPDATE RULES

To decrease the number of model parameters and/or to impose a block diagonal structure, we now derive the M-step update rules for numerous assumptions about the matrix-valued parameters' structure. Therefore, the derivative of the expected log-likelihood $\mathcal{Q}(\boldsymbol{\theta})$ in (4) w.r.t. the parameter is calculated and set to zero, analogously to the derivations in [35], [36]. Irrespective of the assumption about the structures, the update rules for the initial *a priori* state (10) and for the initial state covariance matrix (11) hold.

A. STATE TRANSITION UPDATE RULES

Assuming a *scaled identity* state transition matrix $\mathbf{A} = a\mathbf{I}$ yields

$$a^\star = \frac{\text{tr}\left\{\mathbf{\Gamma}^{-1}\sum_{n=2}^{N}\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right\}}{\text{tr}\left\{\mathbf{\Gamma}^{-1}\sum_{n=2}^{N}\mathbb{E}\left\{\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right\}}. \tag{20}$$

Here, tr$\{\cdot\}$ denotes the trace operator. This model corresponds to the scalar fading factor Markov model as found in many Kalman filtering approaches, e.g., in [26], [40], [41], [42].

Assuming a *diagonal* state transition matrix $\mathbf{A} = \text{diag}\{\mathbf{a}\}$ with $\mathbf{a} \in \mathbb{R}^{N_z}$ yields

$$\mathbf{a}^\star = \left(\mathbf{\Gamma}^{-1} \odot \sum_{n=2}^{N}\mathbb{E}\left\{\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right)^{-1} \cdot \text{diag}\left\{\mathbf{\Gamma}^{-1}\sum_{n=2}^{N}\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right\}, \tag{21}$$

where diag$\{\cdot\}$ converts a vector into a diagonal matrix or extracts the main diagonal from a matrix, and $\odot$ represents elementwise multiplication. In combination with a DFT-transformed state vector, this model allows for frequency-dependent fading factors. This could be assumed when the physical distance between transmitter(s) and receiver(s) is expected to change. This would result in slower changes

[1] For small block sizes, a different term could dominate the overall complexity. A comprehensive analysis of the exact computational complexity seems impractical here due to the variety of relationships between the parameters R, T, L, N_o, N_{B_z}, N_{B_y} and N that, together with the choice of M-step update equations, determine the block matrix dimensions and the number of computations per iteration.

in lower frequencies and faster changes in higher frequencies, i.e., different fading-factor time constants per frequency.

Next, a *block diagonal* state transition matrix

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{\Pi}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Pi}_{N_{B_z}} \end{bmatrix} = \text{mkblkdiag}\{\boldsymbol{\Pi}\} \tag{22}$$

with N_{B_z} blocks of size $B_z = N_z/N_{B_z}$ is assumed, and the auxiliary matrix

$$\boldsymbol{\Pi} = \begin{bmatrix} \boldsymbol{\Pi}_1^{\mathrm{T}} & \dots & \boldsymbol{\Pi}_{N_{B_z}}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{N_{B_z}B_z \times B_z} \tag{23}$$

allows to use the mkblkdiag{·} operator, which "makes" a block diagonal matrix and which can be understood as

$$\mathbf{A} = \sum_{b=1}^{N_{B_z}} \mathbf{E}_b \boldsymbol{\Pi} \boldsymbol{\mathcal{E}}_b^{\mathrm{T}} \tag{24}$$

with a block matrix

$$\mathbf{E}_i = \begin{bmatrix} \mathbf{0} & & & & \\ & \ddots & & & \\ & & \mathbf{I}_{B_z} & & \\ & & & \ddots & \\ & & & & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{N_{B_z}B_z \times N_{B_z}B_z} \tag{25}$$

that has an identity matrix of size $B_z \times B_z$ in block row i and block column i, and a unit vector-like matrix

$$\boldsymbol{\mathcal{E}}_i = \begin{bmatrix} \mathbf{0} & \dots & \mathbf{I}_{B_z} & \dots & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{B_z \times N_{B_z}B_z} \tag{26}$$

that has an identity matrix in block column i. To derive the update rule for $\boldsymbol{\Pi}$, which contains all the parameters describing $\mathbf{A}$, we set $\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\Pi}} \stackrel{!}{=} 0$, which yields the condition

$$\begin{aligned} &\text{blkdiag}\left\{\boldsymbol{\Gamma}^{-1} \sum_{n=2}^{N} \mathbb{E}\left\{\mathbf{z}_n \mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right\} \stackrel{!}{=} \\ &\text{blkdiag}\left\{\boldsymbol{\Gamma}^{-1}\, \text{mkblkdiag}\{\boldsymbol{\Pi}\} \sum_{n=2}^{N} \mathbb{E}\left\{\mathbf{z}_{n-1} \mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right\}. \end{aligned} \tag{27}$$

A system of the form

$$\text{blkdiag}\{\boldsymbol{\mathcal{A}}\} = \text{blkdiag}\{\boldsymbol{\mathcal{B}}\, \text{mkblkdiag}\{\boldsymbol{\mathcal{X}}\}\, \boldsymbol{\mathcal{C}}\} \tag{28}$$

for matrices $\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{C}} \in \mathbb{R}^{N_{B_z}B_z \times N_{B_z}B_z}$ and $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N_{B_z}B_z \times B_z}$ can be rewritten using a vectorized representation of the block matrices using vec{·}, i.e., stacking all matrix elements into a column vector, as

$$\text{vec}\{\boldsymbol{\mathcal{A}}\} = \mathcal{S}(\boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{C}})\, \text{vec}\{\boldsymbol{\mathcal{X}}\}, \tag{29}$$

where $\mathcal{S}(\boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{C}})$ is a matrix that implements (28). Eventually, to find a representation of $\boldsymbol{\Pi}$, we solve

$$\begin{aligned} \text{vec}\left\{\boldsymbol{\Pi}^{\star}\right\} = \mathcal{S}\left(\boldsymbol{\Gamma}^{-1}, \sum_{n=2}^{N} \mathbb{E}\left\{\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right)^{-1} \\ \cdot \text{vec}\left\{\text{blkdiag}\left\{\boldsymbol{\Gamma}^{-1} \sum_{n=2}^{N} \mathbb{E}\left\{\mathbf{z}_n \mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right\}\right\} \end{aligned} \tag{30}$$

and undo the vectorize operation. The update rule (30) requires solving a linear system of equations with $N_{B_z}B_z^2$ variables, which can become impractically large. Note that (20) and (21) also can involve large matrix multiplications.

To avoid them, we can, instead of full coupling between all states, assume *independent blocks* of B_z states in the state vector. Then, the condition in (27) simplifies, and we obtain N_{B_z} separate matrix-valued equations with B_z^2 variables each and obtain the update rule for the block with index b as

$$\begin{aligned} \boldsymbol{\Pi}_b^{\star} = \left(\sum_{n=2}^{N} \mathbb{E}\left\{\left[\mathbf{z}_n \mathbf{z}_{n-1}^{\mathrm{T}}\right]_{bb}\right\}\right) \\ \cdot \left(\sum_{n=2}^{N} \mathbb{E}\left\{\left[\mathbf{z}_{n-1} \mathbf{z}_{n-1}^{\mathrm{T}}\right]_{bb}\right\}\right)^{-1}, \end{aligned} \tag{31}$$

where $[\cdot]_{ij}$ denotes the block in the i-th block row and j-th block column. Instead of dealing with matrices of size $N_{B_z}B_z^2 \times N_{B_z}B_z^2$ and $N_z \times N_z$ in (30), the matrices in (31) are only of size $B_z \times B_z$.

This block structure can be applied to model that all IRs change independently, for instance, when the transmitters are spatially far apart from each other, or to model that complex-valued frequencies bins change independently (for $B_z = 2$).

B. PROCESS NOISE UPDATE RULES

Assuming a *scaled identity* matrix $\boldsymbol{\Gamma} = \gamma \mathbf{I}$ yields

$$\gamma^{\star} = \frac{1}{N_z\,(N-1)} \sum_{n=2}^{N} \text{tr}\{\boldsymbol{\mathcal{G}}_n\}. \tag{32}$$

This model resembles the so-called broadband Kalman filter in [39]. In combination with $\mathbf{T}$ as the real-valued DFT, the model could also be applied when assuming that all frequencies change by similar amounts.

Assuming a *diagonal* matrix $\boldsymbol{\Gamma} = \text{diag}\{\boldsymbol{\gamma}\}$ with $\boldsymbol{\gamma} \in \mathbb{R}^{N_z}$ yields

$$\boldsymbol{\gamma}^{\star} = \frac{1}{N-1} \sum_{n=2}^{N} \text{diag}\{\boldsymbol{\mathcal{G}}_n\}. \tag{33}$$

With $\mathbf{T}$ as the real-valued DFT, this model could be applied when assuming that all frequencies change by different amounts. This can be understood as similar to the diagonal process noise covariance matrix in the DFT-domain Kalman filter [26]. However, there the state vector results from transforming a zero-padded IR estimate into the DFT domain. In

the time domain, this structure corresponds to the structure found in the process noise estimation in [27].

Assuming a *block diagonal* matrix $\mathbf{\Gamma}$ with N_{B_z} blocks of size $B_z \times B_z$, i.e., $\mathbf{\Gamma} = \mathrm{mkblkdiag}\{\mathbf{G}\}$ with $\mathbf{G} = \begin{bmatrix}\mathbf{G}_1^{\mathrm{T}} & \dots & \mathbf{G}_{N_{B_z}}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{N_{B_z}B_z \times B_z}$, containing all the parameters describing $\mathbf{\Gamma}$, yields

$$\mathbf{G}^{\star} = \frac{1}{N-1}\sum_{n=2}^{N} \mathrm{blkdiag}\left\{\mathcal{G}_n\right\}. \tag{34}$$

For a large number of states N_z, the update rules (32), (33) and (34) still require computing products of large dense matrices in (12) before extracting the relevant matrix entries. If it is instead assumed that *independent blocks* occur in the state vector and that the state transition matrix $\mathbf{A}$ is modeled as a block diagonal matrix as well, we can simplify to obtain the update rule for block b as follows:

$$\begin{aligned}\mathbf{G}_b^{\star} = \frac{1}{N-1}\sum_{n=2}^{N} &\left[\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right\}\right]_{bb} - \left[\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right]_{bb}\left[\mathbf{A}^{\mathrm{T}}\right]_{bb} \\ &- [\mathbf{A}]_{bb}\left[\mathbb{E}\left\{\mathbf{z}_{n-1}\mathbf{z}_n^{\mathrm{T}}\right\}\right]_{bb} \\ &+ [\mathbf{A}]_{bb}\left[\mathbb{E}\left\{\mathbf{z}_{n-1}\mathbf{z}_{n-1}^{\mathrm{T}}\right\}\right]_{bb}\left[\mathbf{A}^{\mathrm{T}}\right]_{bb}.\end{aligned} \tag{35}$$

Note that the matrices in (35) are only of size $B_z \times B_z$. This model is conceptually similar to the so-called submatrix-diagonal form for MISO systems in [38], where dependencies between transmitters for a fixed frequency bin are considered.

C. MEASUREMENT NOISE UPDATE RULES

Assuming a *scaled identity* measurement noise covariance matrix $\mathbf{\Sigma} = s\mathbf{I}_{N_\mathrm{o}}$ yields

$$s^{\star} = \frac{1}{NN_\mathrm{o}}\sum_{n=1}^{N} \mathrm{tr}\left\{\mathcal{M}_n\right\}. \tag{36}$$

This corresponds to modeling additive white noise with identical variance for all receivers.

Assuming a *diagonal* measurement noise covariance matrix $\mathbf{\Sigma} = \mathrm{diag}\{\mathbf{s}\}$ with $\mathbf{s} \in \mathbb{R}^{N_\mathrm{o}}$ yields

$$\mathbf{s}^{\star} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{diag}\left\{\mathcal{M}_n\right\}. \tag{37}$$

With samplewise processing, i.e., $N_\mathrm{o} = 1$, a different amount of additive white noise in each receiver could be modeled.

Assuming a *block diagonal* measurement noise covariance with N_{B_y} blocks of size $B_y \times B_y$ with $B_y = RN_o/N_{B_y}$, i.e., $\mathbf{\Sigma} = \mathrm{mkblkdiag}\{\mathbf{S}\}$ with $\mathbf{S} = [\mathbf{S}_1^{\mathrm{T}} \quad \dots \quad \mathbf{S}_{N_{B_y}}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{N_{B_y}B_y \times B_y}$ yields

$$\mathbf{S}^{\star} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{blkdiag}\left\{\mathcal{M}_n\right\}. \tag{38}$$

For $N_\mathrm{o} > 1$, this allows to model additive noise with correlation between noise samples, e.g., differently colored noise at each receiver. The larger N_o, the longer temporal correlations in the measurement noise can be modeled.

Assuming *independent blocks* in the observations $\mathbf{y}_n$, (38) can further be simplified as follows:

$$\begin{aligned}\mathbf{S}_b^{\star} = \frac{1}{N}\sum_{n=1}^{N} &\left[\mathbf{y}_n\mathbf{y}_n^{\mathrm{T}}\right]_{bb} - \left[\mathbf{y}_n\right]_{b1}\mathbb{E}\left\{\mathbf{z}_n^{\mathrm{T}}\right\}[\mathbf{C}_n]_{b:}^{\mathrm{T}} \\ &- [\mathbf{C}_n]_{b:}\mathbb{E}\left\{\mathbf{z}_n\right\}\left[\mathbf{y}_n\right]_{b1}^{\mathrm{T}} + [\mathbf{C}_n]_{b:}\mathbb{E}\left\{\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}\right\}[\mathbf{C}_n]_{b:}^{\mathrm{T}},\end{aligned} \tag{39}$$

where $[\mathbf{C}_n]_{b:}$ denotes the b-th block row and all (block) columns of $\mathbf{C}_n$. In contrast to the fully populated measurement noise covariance matrix in (9), which would also try to jointly model measurement noise at different receivers—a reasonable assumption if there is an external noise source that affects all receivers—(39) allows to model colored measurement noise independently at each receivers, for instance, microphone self-noise.

VI. EXAMPLES

To demonstrate the potential of the proposed framework for the identification of time-variant linear acoustic systems, two examples are presented.

To judge the IR estimate quality, we evaluate the relative system distance (also called normalized misalignment) at frame n between the IR estimates and those of a reference measurement for one specific position that is also contained in the continuous measurement, represented by a corresponding state vector $\mathbf{z}^{(\mathrm{ref})}$, as

$$\mathrm{SD}_n = 20\log_{10}\left(\left\|\hat{\boldsymbol{\mu}}_n - \mathbf{z}^{(\mathrm{ref})}\right\| / \left\|\mathbf{z}_n^{(\mathrm{ref})}\right\|\right)\mathrm{dB}. \tag{40}$$

Here, $\hat{\boldsymbol{\mu}}_n$ could also represent the IR estimates of a baseline algorithm, such as the NLMS algorithm.

A. HRTF MEASUREMENT WITH 37 CHANNELS

In [43], HRTFs were measured using a continuous system identification approach with $T = 37$ loudspeakers at different elevations in an anechoic chamber. Linear sweeps were played back while a turntable rotated the subject (or dummy head). The rotation speed was $1.5^{\circ}\mathrm{s}^{-1}$. Each impulse response was $L = 1024$ samples long, resulting in a sweep period length of $TL = 37\,888$. A reference measurement of the frontal HRTF for all 37 channels is available as the dummy head in the example measurement remained motionless for multiple sweep periods before the rotation started. Due to the optimal convergence properties of perfect periodic sequences [44] these IRs can be considered a valid reference set of HRIRs for this position.

We estimated the HRIRs using the NLMS algorithm with a step size of 0.5, as used in the original HUTUBS database [43], and with a step size of 1.0. Then, a comparison to the results using the proposed framework with the following settings was conducted: The segment length was chosen to be $6 \cdot 37\,888$ (5.15 s at a sampling rate of 44.1 kHz) with equal lengths of lookback part, central part and lookahead part, as in [31]. We assumed that the IRs are constant for about

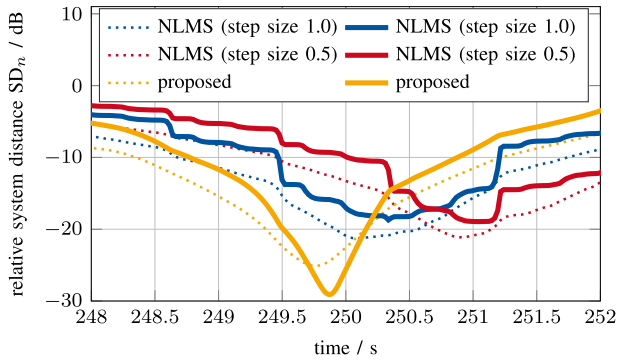

FIGURE 1. Comparison of time-dependent relative system distances for rotation through initial position for all channels (dashed) and 0°-elevation channel (solid).

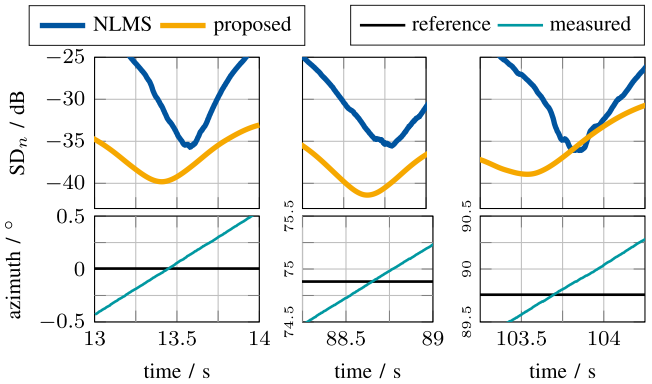

FIGURE 2. Comparisons of relative system distances when rotating through three different example reference measurement positions.

5.8 ms, corresponding to very small spatial angles, and hence chose $N_\mathrm{o} = 256$. Independent time-domain coefficients were assumed (cf. Table 1, ID 8) due to the high state dimension of $N_z = 37\,888$, and (31), (35), (10), and (11) were used in the M-step. $\boldsymbol{\theta}^{(0)}$ was chosen based on preliminary estimates obtained with step size 1.0 in the NLMS algorithm, and $\mathcal{I} = 2$ iterations were conducted.

Following Section IV-C, the dominant complexity term for the formulation in [31] with $N_o = 1$ would require about $1.2 \cdot 10^{19}$ operations, which is reduced to roughly $3.4 \cdot 10^7$ here due to $B_z = 1$. The memory requirement decreases from storing $3.3 \cdot 10^{14}$ to $3.4 \cdot 10^7$ elements. This highlights that the blockwise processing and the assumption of independent time-domain coefficients make it feasible to compute a solution with a reasonable amount of resources.

Fig. 1 shows the relative system distance SD$_n$ for the time when the dummy head rotates through the initial position again after having rotated 360°. When the dummy head approaches the initial position, SD$_n$ is expected to decrease until the initial orientation is reached. There SD$_n$ is expected to reach a minimum—the continuous measurement and the reference measurement match closest. When the dummy head moves away from this orientation, SD$_n$ is expected to increase as the HRIRs begin to deviate again.

The time-dependent relative system distance SD$_n$, comparing all 37 IRs, is shown with thin dashed lines and exhibits the lowest minimum for the proposed method, corresponding to an improvement of about 4 dB compared to the NLMS algorithm's results. Additionally, the relative system distance for the single IR between the 0°-elevation loudspeaker and the left ear, as shown by the thick solid line, improves by about 10 dB. The staircase-like shapes for the relative system distances for the NLMS algorithm are a result of the linear sweep excitation signal. The distance between the jumps matches one period length of 37 888 samples (0.86 s). The curve for step size 0.5 appears to be slightly delayed as a result of the implicit temporal smoothing of the NLMS algorithm for step sizes less than one. Both of these observations match the analyses in [45]. Overall, this result demonstrates that a significant improvement can be achieved with the proposed framework despite the simplifying assumptions about the matrix-valued parameters.

B. TWO-LOUDSPEAKER HRTF MEASUREMENT

We conducted a continuous measurement of HRTFs from $T = 2$ loudspeakers spanning a 60° stereo setup at a distance of 1.75 m in a semi-anechoic chamber with a reflective floor. A noise-like perfect periodic sequence of period length $2L = 9600$ was used as excitation signal in accordance with [44] to identify two IRs of length $L = 4800$ (100 ms at a sampling rate of 48 kHz) in parallel. We consider the system as a MIMO setup with $R = 2$ coupled receivers, represented by the microphones of the HEAD acoustics HMS II.3 dummy head. It was rotated on a turntable with a rotational velocity of $1°\mathrm{s}^{-1}$. Only coupling between receivers in the DFT domain (cf. Table 1, ID 10) was assumed as the two microphones of the dummy head rotate jointly. The M-step updates (31), (35), (10), and (11) were applied with $B_z = 4$. We assumed white measurement noise and hence also chose (36) in the M-step. It was assumed that the IRs are constant for durations of 1ms and thus $N_\mathrm{o} = 48$ was set. A segment length of 57 600 samples was chosen, and $\mathcal{I} = 2$ iterations were conducted. The reference IRs at azimuth angle near 0°, 75° and 90° were measured using exponential sweeps.

Similarly to above, the dominant complexity term for the MIMO formulation in Section III with $N_o = 1$ would require about $4.1 \cdot 10^{17}$ operations, which is reduced to roughly $3.7 \cdot 10^8$ here due to block size $B_z = 4$. The memory requirement decreases from storing $2.1 \cdot 10^{13}$ to $9.2 \cdot 10^7$ elements.

Fig. 2 shows the time-dependent relative system distances SD$_n$ for the times when the azimuth angle corresponding to the continuous rotation passed through the reference azimuth angles, as indicated by the lower plot that displays the azimuth angle recorded at the reference position and during the measurement. The minimum relative system distances, comparing all four IRs, are improved by about 3 dB to 5 dB compared to the result of applying the NLMS algorithm with step size 1.0. The minima occur temporally close to but slightly before the expected position, which is suspected to be a consequence of

the head-tracking system's latency and/or mechanical imperfections. The delay in attaining the minimum SD_n with the NLMS algorithm is a consequence of the systematic delay of half a period length, i.e., 4800 samples (0.1 s), as also analyzed in [45].

VII. CONCLUSION

We have presented a framework for EM-based identification of time-variant linear acoustic systems. This framework combines the time-domain block observation model for a MIMO system with a state vector transform, as well as a variety of coupling models that can yield a block diagonal matrix structure, for which we have derived M-step update rules. The choice of model structure and block sizes determines the computational complexity. This way, tasks that were previously considered computationally infeasible, such as HRTF measurements involving numerous channels, can be successfully addressed with the joint state and parameter estimation. Our examples illustrate that this framework can improve the quality of the quasi-continuous IR estimates by up to 10 dB in relative system distance when comparing to a reference measurement. The proposed framework hence enables improved quasi-continuous MIMO system identification.

ACKNOWLEDGMENT

The authors would like to thank Fabian Brinkmann from TU Berlin for sharing exemplary raw measurement data of the HUTUBS HRTF database and simulations were performed with computing resources granted by RWTH Aachen University under project rwth0827.

REFERENCES

[1] S. M. Kuo and D. R. Morgan, "Active noise control: A tutorial review," *Proc. IEEE*, vol. 87, no. 6, pp. 943–973, Jun. 1999.
[2] T. Betlehem, W. Zhang, M. A. Poletti, and T. D. Abhayapala, "Personal sound zones: Delivering interface-free audio to multiple listeners," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 81–91, Mar. 2015.
[3] O. Kirkeby and P. A. Nelson, "Digital filter design for inversion problems in sound reproduction," *J. Audio Eng. Soc.*, vol. 47, no. 7/8, pp. 583–595, 1999.
[4] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*. Plantation, FL, USA: J. Ross Publishing, 2013.
[5] S. Li and J. Peissig, "Measurement of head-related transfer functions: A review," *Appl. Sci.*, vol. 10, no. 14, 2020, Art. no. 5014.
[6] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *J. Acoust. Soc. Amer.*, vol. 49, no. 6, pp. 443–471, 2001.
[7] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Proc. Audio Eng. Soc. Conv.*, 2007, pp. 1–21.
[8] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *J. Audio Eng. Soc.*, vol. 55, no. 7/8, pp. 623–637, 2007.
[9] P. Dietrich, B. Masiero, and M. Vorländer, "On the optimization of the multiple exponential sweep method," *J. Audio Eng. Soc.*, vol. 61, no. 3, pp. 113–124, 2013.
[10] C. Antweiler and H. G. Symanzik, "Simulation of Time Variant Room Impulse Responses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 3031–3034.
[11] C. Urbanietz and G. Enzner, "Binaural rendering of dynamic head and sound source orientation using high-resolution HRTF and retarded time," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 566–570.
[12] G. Enzner, C. Antweiler, and S. Spors, "Trends in acquisition of individual head-related transfer functions," in *The Technology of Binaural Listening*, J. Blauert, Ed. Berlin, Germany: Springer, 2013, Ch. 3, pp. 57–92.
[13] V. Pulkki, M.-V. Laitinen, and V. Sivonen, "HRTF measurements with a continuously moving loudspeaker and swept sines," in *Proc. Audio Eng. Soc. Conv.*, 2010, pp. 1–9.
[14] K. Fukudome, T. Suetsugu, T. Ueshin, R. Idegami, and K. Takeya, "The fast measurement of head related impulse responses for all azimuthal directions using the continuous measurement method with a servo-swiveled chair," *Appl. Acoust.*, vol. 68, no. 8, pp. 864–884, 2007.
[15] G. Enzner, "Analysis and optimal control of LMS-type adaptive filtering for continuous-azimuth acquisition of head related impulse responses," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 393–396.
[16] G. Enzner, "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust*, 2009, pp. 325–328.
[17] C. Antweiler and G. Enzner, "Perfect sequence LMS for rapid acquisition of continuous-azimuth head related impulse responses," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust*, 2009, pp. 281–284.
[18] J.-G. Richter and J. Fels, "On the influence of continuous subject rotation during high-resolution head-related transfer function measurements," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 730–741, Apr. 2019.
[19] E.-L. Tan, S. Peksi, and W.-S. Gan, "Implementing continuous HRTF measurement in near-field," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
[20] J. He, R. Ranjan, and W.-S. Gan, "Fast continuous HRTF acquisition with unconstrained movements of human subjects," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 321–325.
[21] S. Li and J. Peissig, "Fast estimation of 2D individual HRTFs with arbitrary head movements," in *Proc. IEEE 22nd Int. Conf. Digit. Signal Process.*, 2017, pp. 1–5.
[22] J. He, R. Ranjan, W.-S. Gan, N. K. Chaudhary, N. D. Hai, and R. Gupta, "Fast continuous measurement of HRTFs with unconstrained head movements for 3D audio," *J. Audio Eng. Soc.*, vol. 66, no. 11, pp. 884–900, 2018.
[23] S. Nagel, T. Kabzinski, S. Kühl, C. Antweiler, and P. Jax, "Acoustic head-tracking for acquisition of head-related transfer functions with unconstrained subject movement," in *Proc. AES Int. Conf. Audio Virtual Augmented Reality. Audio Eng. Soc.*, 2018, pp. 1–10.
[24] J. Reijniers, B. Partoens, J. Steckel, and H. Peremans, "HRTF measurement by means of unsupervised head movements with respect to a single fixed speaker," *IEEE Access*, vol. 8, pp. 92287–92300, 2020.
[25] S. Kühl, S. Nagel, T. Kabzinski, C. Antweiler, and P. Jax, "A joint perspective of periodically excited efficient NLMS algorithm and inverse cyclic convolution," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 406–410.
[26] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Process.*, vol. 86, no. 6, pp. 1140–1156, 2006.
[27] S. Liebich, J. Fabry, P. Jax, and P. Vary, "Time-domain Kalman filter for active noise cancellation headphones," in *Proc. IEEE 25th Eur. Signal Process. Conf.*, 2017, pp. 593–597.
[28] S. Malik and G. Enzner, "Online maximum-likelihood learning of time-varying dynamical models in block-frequency-domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 3822–3825.
[29] C. Urbanietz and G. Enzner, "Spatial-fourier retrieval of head-related impulse responses from fast continuous-azimuth recordings in the time-domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 950–954.
[30] J. C. Agüero, J. I. Yuz, and G. C. Goodwin, "Frequency domain identification of MIMO state space models using the EM algorithm," in *Proc. IEEE Eur. Control Conf.*, 2007, pp. 5686–5693.
[31] T. Kabzinski and P. Jax, "Towards faster continuous multi-channel HRTF measurements based on learning system models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 436–440.
[32] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Berlin, Germany: Springer, 2006.
[33] J. Fabry, D. Hilkert, S. Liebich, and P. Jax, "Time-variant acoustic front-end measurements of active noise cancellation headphones," in *Proc. 23rd Int. Congr. Acoust.*, 2019, pp. 4326–4333.

[34] T. Kabzinski and P. Jax, "A unified perspective on time-domain and frequency-domain Kalman filters for acoustic system identification," in *Proc. IEEE 30th Eur. Signal Process. Conf.*, 2022, pp. 90–94.
[35] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[36] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Ser. Anal.*, vol. 3, no. 4, pp. 253–264, 1982.
[37] T. Adali, P. J. Schreier, and L. L. Scharf, "Complex-valued signal processing: The proper way to deal with impropriety," *Trans. Signal Process.*, vol. 59, no. 11, pp. 5101–5125, 2011.
[38] S. Malik and G. Enzner, "Recursive Bayesian control of multichannel acoustic echo cancellation," *Signal Process Lett.*, vol. 18, no. 11, pp. 619–622, 2011.
[39] G. Enzner, "Bayesian inference model for applications of time-varying acoustic system identification," in *Proc. IEEE 18th Eur. Signal Process. Conf.*, 2010, pp. 2126–2130.
[40] F. Kuech, E. Mabande, and G. Enzner, "State-space architecture of the partitioned-block-based acoustic echo controller," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1295–1299.
[41] J. Fabry, S. Liebich, P. Vary, and P. Jax, "Active noise control with reduced-complexity Kalman filter," in *Proc. IEEE 16th Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 166–170.
[42] J. Fabry, S. Kühl, and P. Jax, "On the steady state performance of the Kalman filter applied to acoustical systems," *Signal Process. Lett.*, vol. 27, pp. 1854–1858, 2020.
[43] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718, 2019.
[44] C. Antweiler, "Multi-Channel System Identification with Perfect Sequences–Theory and Applications," in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds. West Sussex, U.K.: Wiley, 2008, Ch. 7, pp. 171–198.
[45] S. Kühl, S. Nagel, T. Kabzinski, C. Antweiler, and P. Jax, "Tracking of time-variant linear systems: Influence of group delay for different excitation signals," in *Proc. IEEE 16th Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 131–135.