SOURCE-SPECIFIC SYSTEM IDENTIFICATION

Christian Hofmann and Walter Kellermann

Friedrich-Alexander University Erlangen-Nürnberg (FAU),
Multimedia Communications and Signal Processing,
Cauerstr. 7, D-91058 Erlangen, Germany
{christian.hofmann,walter.kellermann}@FAU.de

ABSTRACT

Many applications in audio communication require the identification of Loudspeaker-Enclosure-Microphone Systems (LEMS) with multiple inputs and outputs. The according computational complexity typically grows at least proportionally along the number of acoustic paths, which is the product of the number of loudspeakers and the number of microphones. Furthermore, the typical, highly correlated loudspeaker signals preclude an exact identification of the LEMS. To this end, a novel system identification scheme employing prior information from an object-based rendering system, e.g., Ambisonics [1, 2] or Wave Field Synthesis (WFS) [3, 4], is proposed. In this scheme, only a source-specific system from each virtual source to each microphone is identified adaptively and uniquely. This estimate for a source-specific system can then be transformed into a statistically optimal estimate of the LEMS, which could have been found by a computationally expensive direct LEMS estimation as well. The basic concept is extended to time-varying acoustic scenes and simulations of a WFS application confirm the validity of this novel approach for system identification.

***Index Terms*—** AEC, system identification, reduced-complexity, subspace system identification, side information

1. INTRODUCTION

Applications such as Acoustic Echo Cancellation (AEC) or Listening Room Equalization (LRE) require the identification of acoustic Multiple-Input/Multiple-Output (MIMO) systems. In practice, multichannel acoustic system identification suffers severely from strongly cross-correlated loudspeaker signals typically occurring when rendering virtual acoustic scenes with more than one loudspeaker [5, 6]: the computational complexity grows with at least the number of acoustical paths through the MIMO system, which is $N_\mathrm{L} \cdot N_\mathrm{M}$ for N_L loudspeakers and N_M microphones. Robust fast-converging algorithms for multichannel filter adaptation, such as the Generalized Frequency-Domain Adaptive Filtering (GFDAF) algorithm [7] even have a complexity $\mathcal{O}\left(N_\mathrm{L}^3\right)$ when robustly solving the involved linear systems of equations for cross-correlated loudspeaker signals by a Cholesky decomposition [8]. Moreover, if the number of loudspeakers is larger than the number of virtual sources N_S (i.e. the number of spatially separated sources with independent signals), the acoustic paths from the loudspeakers to the microphones of the Loudspeaker-Enclosure-Microphone System (LEMS) cannot be determined uniquely. As this so-called non-uniqueness problem [6] is inevitable in practice, an infinitely large set of possible solutions for the LEMS exists, from which only one corresponds to the true LEMS.

In the past decades, decorrelation by additive noise or coding noise [5, 9], as well as nonlinear [10, 11] or time-variant [12–14] pre-processing of the loudspeaker signals has been proposed to address the non-uniqueness problem while at least slightly increasing the computational burden. On the other hand, the concept of Wave-Domain Adaptive Filtering (WDAF) [15,16] alleviates both the computational complexity and the non-uniqueness problem [17] and is optimum for uniform, concentric, circular loudspeaker and microphone arrays. Another approach known as Source-Domain Adaptive Filtering (SDAF) [18] performs a data-driven spatio-temporal transform on the loudspeaker and microphone signals in order to allow an effective modeling of acoustic echo paths in the resulting highly time-varying transform domain. Yet, the identified system does not represent the LEMS, but is a signal-dependent approximation. Another adaptation scheme is called Eigenspace Adaptive Filtering (EAF) [19], where an N^2-channel acoustic MIMO system with $N_\mathrm{L} = N_\mathrm{M} = N$ would correspond to exactly N paths after transformation of the signals into the system's eigenspace. On the other hand, this requires to estimate the eigenspace of the particular LEMS in the first place [20].

Different to all the aforementioned approaches, we propose a novel method for system identification which employs prior information from an object-based rendering system (statistically independent source signals and the corresponding rendering filters) in order to reduce the computational complexity and, although the LEMS cannot be determined uniquely, to allow for a unique solution of the involved adaptive filtering problem. The proposed method, denoted as Source-Specific System Identification (SSSysId), will be introduced in Sec. 2 and evaluated in Sec. 3, followed by a brief summary in Sec. 4.

2. SOURCE-SPECIFIC SYSTEM IDENTIFICATION

Consider an object-based rendering system, i.e. Wave Field Synthesis (WFS) [21], which renders N_S statistically independent virtual sound sources (e.g., point sources or plane-wave sources) employing an array of N_L loudspeakers. To allow for a voice control of an entertainment system or an additional use of the reproduction system as hands-free acoustic front-end in a communication scenario, a set of N_M microphones for sound acquisition and an AEC unit is required as well. The acoustic paths between the loudspeakers and N_M microphones of interest are described as linear systems with DTFT-domain transfer function matrices $\mathbf{H}(\mathrm{e}^{\mathrm{j}\Omega}) \in \mathbb{C}^{N_\mathrm{M} \times N_\mathrm{L}}$ with the normalized angular frequency Ω. For the conciseness of notation, the argument Ω will be neglected for all signal vectors and transfer function matrices, which means that $\mathbf{H}$ stands for $\mathbf{H}(\mathrm{e}^{\mathrm{j}\Omega})$. This notation is employed in Fig. 1, which depicts the vector of DTFT-domain source signals $\mathbf{s} \in \mathbb{C}^{N_\mathrm{S}}$, the rendering ('driving') filters' transfer function matrix $\mathbf{H}_\mathrm{D} \in \mathbb{C}^{N_\mathrm{L} \times N_\mathrm{S}}$, the loudspeaker signals $\mathbf{x}_\mathrm{LS} = \mathbf{H}_\mathrm{D}\mathbf{s} \in \mathbb{C}^{N_\mathrm{L}}$, the LEMS transfer function matrix $\mathbf{H}$, and the microphone signal vector

$$\mathbf{x}_\mathrm{Mic} = \mathbf{H}\mathbf{x}_\mathrm{LS} = \underbrace{\mathbf{H}\mathbf{H}_\mathrm{D}}_{\mathbf{H}_\mathrm{S}}\mathbf{s}, \tag{1}$$

The authors would like to thank the Fraunhofer Institute for Digital Media Technology (IDMT) in Ilmenau, Germany, for supporting this work.

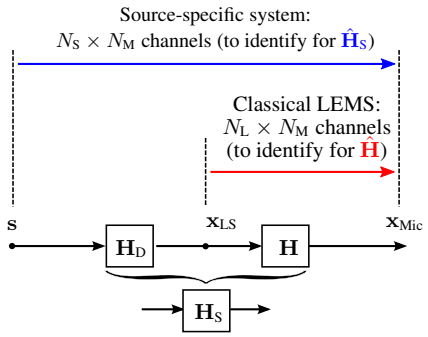


Fig. 1: Comparison of systems to be modeled by classical LEMS identification and by identifying a source-specific system.

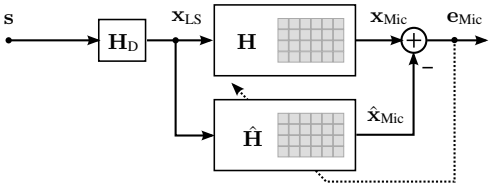


Fig. 2: Estimating the LEMS $\mathbf{H}$. The number of squares symbolizes the number of filter coefficients to estimate.

where the cascade of the rendering filters with the LEMS will be referred to as source-specific system

$$\mathbf{H}_S = \mathbf{H}\mathbf{H}_D \in \mathbb{C}^{N_M \times N_S}. \tag{2}$$

Both for recording near-end sources involving AEC, and for room equalization, the LEMS $\mathbf{H}$ has to be identified adaptively. This is typically done by minimizing a quadratic cost function derived from the difference $\mathbf{e}_{Mic}$ between the recorded microphone signals $\mathbf{x}_{Mic}$ and the microphone signal estimates $\hat{\mathbf{x}}_{Mic}$ obtained with the LEMS estimate $\hat{\mathbf{H}}$, which is depicted in Fig. 2. As mentioned earlier, multichannel acoustic system identification suffers from the strongly cross-correlated loudspeaker signals typically occurring when rendering acoustic scenes with more than one loudspeaker: for more loudspeakers than virtual sources ($N_L > N_S$), the acoustic paths of the LEMS $\mathbf{H}$ cannot be determined uniquely ('non-uniqueness problem' [6]). This means that an infinitely large set of possible solutions for $\hat{\mathbf{H}}$ exists, from which only one corresponds to the true LEMS $\mathbf{H}$.

On the other hand, the paths from each virtual source to each microphone can be described as an $N_S \times N_M$ MIMO system $\mathbf{H}_S$ (see Fig. 1) which can be determined uniquely for the given set of statistically independent virtual sources. Due to the statistical independence of the virtual sources, the computational complexity of the system identification with a GFDAF algorithm increases only linearly with N_S instead of cubically with N_L, as the covariance matrices to be inverted become diagonal. Furthermore, the number of acoustic paths to be modeled is reduced by a factor of N_S/N_L. Hence, an estimate for $\hat{\mathbf{H}}_S$ can be obtained very accurately as depicted in Fig. 3 and with less effort than an estimate for $\hat{\mathbf{H}}$ according to Fig. 2. The systems to be identified and the respective estimates are indicated in Fig. 1 above the block diagrams.

Although $\hat{\mathbf{H}}$ is not determined uniquely by $\hat{\mathbf{H}}_S$ in general, the non-uniqueness of this mapping is exactly the same as the non-uniqueness problem for determining $\hat{\mathbf{H}}$ directly and finding one of the systems $\hat{\mathbf{H}}$ is easily possible by approximating a pseudo-inverse rendering system $\mathbf{H}_D^+$ and pre-filtering the source-specific system $\hat{\mathbf{H}}_S$ to obtain one particular

$$\hat{\mathbf{H}} = \hat{\mathbf{H}}_S \mathbf{H}_D^+. \tag{3}$$

Hence, a statistically optimal estimate $\hat{\mathbf{H}}$, which also could have been the result from adapting $\hat{\mathbf{H}}$ directly, can be obtained by identifying $\mathbf{H}_S$ by an $\hat{\mathbf{H}}_S$ with very low effort and without non-uniqueness problem and transforming $\hat{\mathbf{H}}_S$ into an estimate of $\hat{\mathbf{H}}$ in a systematic way. This can be seen as exploiting non-uniqueness rather than seeing it as a problem: if it is impossible to infer the true system anyway, the effort for finding one of the solutions should be minimized. As the number of paths to be identified reduces by a factor of N_L/N_S, its effort is reduced at least by the same ratio.

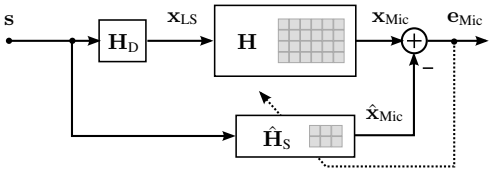


Fig. 3: Estimating the source-specific system $\mathbf{H}_S$. The number of squares symbolizes the number of filter coefficients to estimate.

Determining an LEMS estimate from a Source-Specific System Estimate: In the following, an efficient mapping from a source-specific system to an LEMS corresponding to the source-specific system will be described. For given source-specific transfer function estimates $\hat{\mathbf{H}}_S$, the concatenation of the driving filters $\mathbf{H}_D$ with the LEMS estimate $\hat{\mathbf{H}}$ should fulfill $\hat{\mathbf{H}}\mathbf{H}_D \overset{!}{=} \hat{\mathbf{H}}_S$, analogously to Eq. (2). For the typical case of $N_S < N_L$, an inverse matrix $\mathbf{H}_D^{-1}$ does not exist, but a minimum-norm solution is given by the Moore-Penrose pseudoinverse [22] $\mathbf{H}_D^+ = \left(\mathbf{H}_D^H \mathbf{H}_D\right)^{-1} \mathbf{H}_D^H$. Note that the rendering system's driving filters $\mathbf{H}_D$ and their pseudo inverses $\mathbf{H}_D^+$ can be pre-calculated during the production stage of the audio material. Hence, the LEMS estimate can then be computed from the source-specific transfer functions according to Eq. (3) by pre-filtering $\mathbf{H}_S$. For a driver matrix $\mathbf{H}_D$ with pseudoinverse $\mathbf{H}_D^+$,

$$\mathbf{P} = \mathbf{H}_D \mathbf{H}_D^+ \tag{4}$$

$$\mathbf{P}^\perp = (\mathbf{I} - \mathbf{P}) \tag{5}$$

are known as the projectors into the column space of $\mathbf{H}_D$ and into the left null space of $\mathbf{H}_D$, respectively [22], and decompose the N_L-dimensional space into two orthogonal subspaces. With this, the true LEMS $\mathbf{H}$ can be expressed as sum of two orthogonal components

$$\mathbf{H} = \overbrace{\mathbf{H}\mathbf{P}}^{\mathbf{H}^{||}} + \overbrace{\mathbf{H}(\mathbf{I}-\mathbf{P})}^{\mathbf{H}^{\perp}} \tag{6}$$

$$= \mathbf{H}\mathbf{H}_D\mathbf{H}_D^+ + \mathbf{H}\mathbf{P}^\perp \tag{7}$$

$$= \mathbf{H}_S\mathbf{H}_D^+ + \mathbf{H}^\perp, \tag{8}$$

where $\mathbf{H}^{||} = \mathbf{H}_S\mathbf{H}_D^+$ is a filtered version of the source-specific system $\mathbf{H}_S$ (and can thus be estimated) and where $\mathbf{H}^\perp$ lies in the left null space of $\mathbf{H}_D$ and is not excited by the latter. Therefore, $\mathbf{H}^\perp$ is not observable at the microphones and represents the ambiguity of the solutions for $\hat{\mathbf{H}}$ (non-uniqueness problem). Whenever $\mathbf{H}_D^+$ is employed to map a source-specific system back to an LEMS estimate, the estimate's rows will lie in the column space of $\mathbf{H}_D$ and all components in the left null space of $\mathbf{H}_D$, namely $\mathbf{H}^\perp$, are implied to be $\mathbf{0}$.

Hence, only the LEMS components sensitive to the column space of $\mathbf{H}_D$ can and should be estimated from a particular $\mathbf{H}_S$. This idea will be employed in the next section to extend source-specific system identification for time-varying virtual acoustic scenes.

2.1. Time-varying Virtual Acoustic Scenes

In practice, the number and the positions of virtual acoustic sources may change over time. Thus, the rendering task will be divided into a sequence of intervals with different source configurations that remain constant within each interval. At the beginning of an interval κ ($\kappa \in \mathbb{Z}$), an initial source-specific system estimate

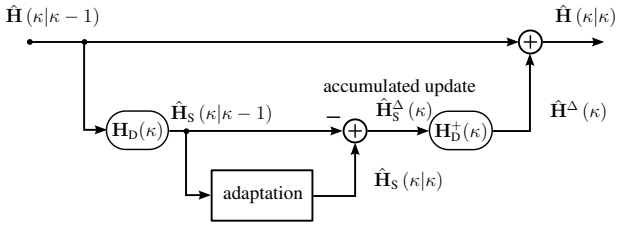


Fig. 4: Average-load-optimized system identification by SSSysId. The lines represent coefficients of MIMO systems and rounded boxes symbolize pre-filtering the connected incoming coefficients with the MIMO system in the box.

$$\hat{\mathbf{H}}_\mathrm{S}(\kappa|\kappa-1) = \hat{\mathbf{H}}(\kappa|\kappa-1)\,\mathbf{H}_\mathrm{D}(\kappa) \tag{9}$$

is computed from the information available from observing the interval $\kappa-1$, namely the initial LEMS estimate $\hat{\mathbf{H}}(\kappa|\kappa-1) = \hat{\mathbf{H}}(\kappa-1|\kappa-1)$ obtained during interval $\kappa-1$, and the current interval's rendering filters $\mathbf{H}_\mathrm{D}(\kappa)$. After adapting only the source-specific system $\hat{\mathbf{H}}_\mathrm{S}$ during interval κ, a final source-specific system estimate $\hat{\mathbf{H}}_\mathrm{S}(\kappa|\kappa)$ is available at the end of interval κ. Embodying the idea to update only $\mathbf{H}^{||}$ and to keep $\hat{\mathbf{H}}^{\perp}(\kappa\,|\,\kappa-1) = \hat{\mathbf{H}}(\kappa|\kappa-1)\left(\mathbf{I}-\mathbf{H}_\mathrm{D}(\kappa)\mathbf{H}_\mathrm{D}^{+}(\kappa)\right)$ constant during a particular interval κ, this can be formulated as

$$\begin{aligned}\hat{\mathbf{H}}(\kappa|\kappa) &= \hat{\mathbf{H}}^{\perp}(\kappa\,|\,\kappa-1) + \hat{\mathbf{H}}_\mathrm{S}(\kappa|\kappa)\,\mathbf{H}_\mathrm{D}^{+}(\kappa) &(10)\\ &= \hat{\mathbf{H}}(\kappa|\kappa-1) \\ &\quad + \underbrace{\left(\hat{\mathbf{H}}_\mathrm{S}(\kappa|\kappa) - \hat{\mathbf{H}}_\mathrm{S}(\kappa|\kappa-1)\right)\mathbf{H}_\mathrm{D}^{+}(\kappa)}_{\hat{\mathbf{H}}^{\Delta}(\kappa)}. &(11)\end{aligned}$$

Due to the pseudoinverse, this corresponds to a minimum-norm update $\hat{\mathbf{H}}^{\Delta}(\kappa)$, the smallest update which leads to $\hat{\mathbf{H}}_\mathrm{S}(\kappa|\kappa)$. As this procedure leaves $\mathbf{H}^{\perp}$ unaltered ($\mathbf{H}^{\perp}(\kappa\,|\,\kappa) = \mathbf{H}^{\perp}(\kappa\,|\,\kappa-1)$), information about the true LEMS can accumulate over all intervals, allowing a continuous refinement of $\hat{\mathbf{H}}$ in case of time-varying acoustic scenes. For two alternating scenes, this means that the initial source-specific system $\hat{\mathbf{H}}_\mathrm{S}(\kappa+2|\kappa+1)$ is much closer to the optimum than $\hat{\mathbf{H}}_\mathrm{S}(\kappa|\kappa-1)$. The operations performed on an LEMS estimate during an interval κ are summarized in Fig. 4, where the 'adaptation' block has very low complexity and the pre-filtering operations with $\mathbf{H}_\mathrm{D}(\kappa)$ and $\mathbf{H}_\mathrm{D}^{+}(\kappa)$ lead to additional computations at a scene change (or during the frames before a scene change). A visualization of the chronological order of the proposed adaptation scheme is given in Fig. 5, where the time line is given at the top, for two subsequent time intervals Interval 1 and Interval 2.

Interval 1: At the beginning of Interval 1 ("Start" in Fig. 5), the estimate $\hat{\mathbf{H}}$ for the LEMS $\mathbf{H}$ is still all zero (indicated by white squares) and it remains like this for the entire interval. After obtaining an initial source-specific system $\hat{\mathbf{H}}_\mathrm{S}(1|0)$ via Eq. (9), the source-specific system $\hat{\mathbf{H}}_\mathrm{S}$ is continuously adapted during this interval, leading to the final estimate $\hat{\mathbf{H}}_\mathrm{S}(1|1)$.

Transition between Intervals 1 and 2: At the transition between Intervals 1 and 2 (center part of Fig. 5), the virtual source configuration changes. Thus, the driving system is exchanged to allow rendering a different virtual scene ($\mathbf{H}_\mathrm{D}(1)$ is replaced by $\mathbf{H}_\mathrm{D}(2)$) and information from $\hat{\mathbf{H}}_\mathrm{S}$ is incorporated into $\hat{\mathbf{H}}$. For this knowledge incorporation, the pseudoinverse $\mathbf{H}_\mathrm{D}^{+}(1)$ of the driving system $\mathbf{H}_\mathrm{D}(1)$ is employed. From the updated LEMS estimate $\hat{\mathbf{H}}(2|1) = \hat{\mathbf{H}}(1|1)$ and the new driving filters $\mathbf{H}_\mathrm{D}(2)$, an initialization $\hat{\mathbf{H}}_\mathrm{S}(2|1)$ for $\hat{\mathbf{H}}_\mathrm{S}$ for the Interval 2 is obtained via Eq. (9).

Interval 2: Analogously to Interval 1, only a small source-specific system is adapted within Interval 2 (bottom). Yet, an estimate $\hat{\mathbf{H}}$ is available in the background (system components contributed by Interval 1 are gray now). In case of another scene change (exceeds time line in Fig. 5), $\hat{\mathbf{H}}_\mathrm{S}(2|2)$ can then refine the LEMS estimate $\hat{\mathbf{H}}$ again, leading to an even better initialization for the subsequent interval's source-specific system. The effort for the adaptive filtering is thereby reduced by at least a factor of $\frac{N_\mathrm{L}}{N_\mathrm{S}}$, even when employing only a frequency-domain Normalized Least Mean Squares (NLMS) algorithm. Furthermore, the overhead due to a scene change can be distributed over several frames.

3. EVALUATION

This section provides a verification and evaluation of the basic properties of the SSSysId adaptation scheme by simulating a WFS scenario with a linear sound bar of $N_\mathrm{L} = 48$ loudspeakers in front of a single microphone under free-field conditions, as depicted at the right of Fig. 6a. Note that the use of just a single microphone is sufficient for general analyses of the behavior of the adaptation concept, as filter adaptation is performed independently for each microphone. The WFS system synthesizes at a sampling rate of 8 kHz one or more simultaneously active virtual point sources radiating statistically independent white noise signals. Besides, high-quality microphones and amplifiers are assumed by adding white Gaussian noise at a level of -60 dB to the microphones. The system identification is performed by a GFDAF algorithm. The rendering systems' inverses are approximated in the Discrete Fourier Transform (DFT) domain and a causal time-domain inverse system is obtained by applying a linear phase shift, an inverse DFT, and subsequent windowing. For numerical stability, the pseudoinverse is approximated in the DFT domain by a Tikhonov regularized inverse $\mathbf{H}_\mathrm{D}^{+\mathrm{Tik}} = \left(\mathbf{H}_\mathrm{D}^{\mathrm{H}}\mathbf{H}_\mathrm{D} + \lambda\mathbf{I}\right)^{-1}\mathbf{H}_\mathrm{D}^{\mathrm{H}}$ with a regularization constant $\lambda = 0.005$, thereby offering a trade-off between the accuracy of the inversion (small λ) and the filter coefficient norm for ill-conditioned $\mathbf{H}_\mathrm{D}$. To evaluate the simulations, we use the normalized residual error signal

$$\Delta_e(k) = 10\log_{10}\left(\frac{\mathbf{e}_\mathrm{Mic}(k)^{\mathrm{H}}\mathbf{e}_\mathrm{Mic}(k)}{\mathbf{x}_\mathrm{Mic}(k)^{\mathrm{H}}\mathbf{x}_\mathrm{Mic}(k)}\right)\mathrm{dB}, \tag{12}$$

where $\mathbf{x}_\mathrm{Mic}(k) \in \mathbb{R}^{N_\mathrm{M}}$ denotes the vector of microphone samples for the discrete-time sample index k and $\mathbf{e}_\mathrm{Mic}(k) \in \mathbb{R}^{N_\mathrm{M}}$ denotes the corresponding error signal vector (becoming a scalar for $N_\mathrm{M} = 1$). This corresponds to the inverse of the commonly used Echo-Return Loss Enhancement (ERLE) measure in AEC. In order to measure how well the LEMS is identified, we employ the normalized system error norm

$$\Delta_h(\kappa) = 10\log_{10}\left(\frac{\sum_{\mu=0}^{L-1}\left\|\hat{\mathbf{H}}_\mu(\kappa\,|\,\kappa) - \mathbf{H}_\mu\right\|_\mathrm{F}^2}{\sum_{\mu=0}^{L-1}\left\|\mathbf{H}_\mu\right\|_\mathrm{F}^2}\right)\mathrm{dB}, \tag{13}$$

where $\mathbf{H}_\mu$ and $\hat{\mathbf{H}}_\mu(\kappa\,|\,\kappa)$ are DFT-domain transfer function matrices of the true and the estimated LEMS, respectively, $\mu \in \{0,\ldots,L-1\}$ is the DFT bin index, and L is the DFT order.

Experiment 1: In this experiment, $N_\mathrm{S} = 4$ virtual sources are synthesized. This allows a reduction of the effort for the adaptive filtering of about $N_\mathrm{L}/N_\mathrm{S} = 12$. In total, three intervals of length 8 s with different virtual source configurations which do not change within each interval are simulated. The three interval's groups of virtual sources are depicted in Fig. 6a: each virtual source is marked by a filled circle and the sources belonging to the same interval of constant source configuration share the same color and are connected by correspondingly colored lines. As the source signals are uncorrelated and the non-uniqueness problem does not exist, SSSysId leads to a uniform decay of the residual error up to the noise floor, as can be seen in Fig. 6b. Furthermore, both SSSysId and a direct LEMS update reveal a very similar robustness against scene changes. This confirms the applicability of SSSysId for AEC.

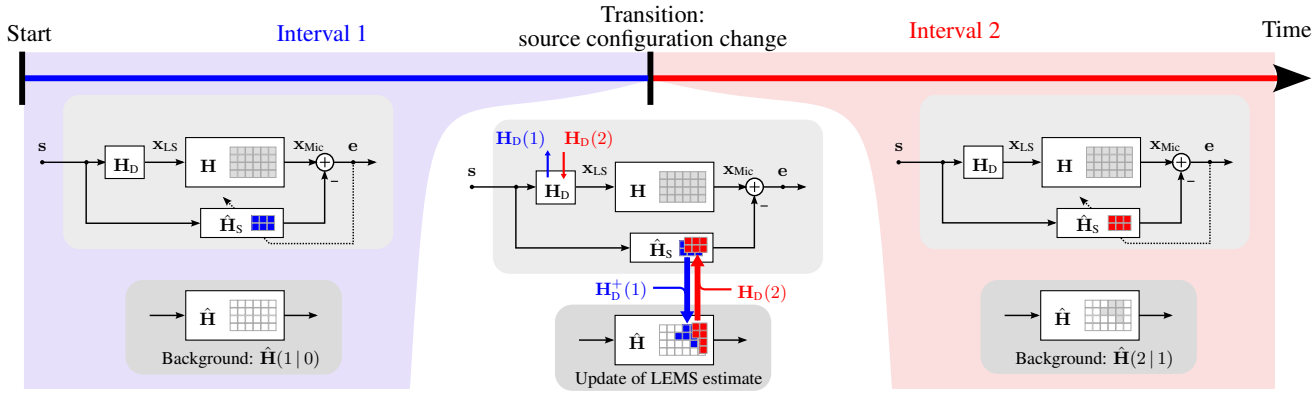
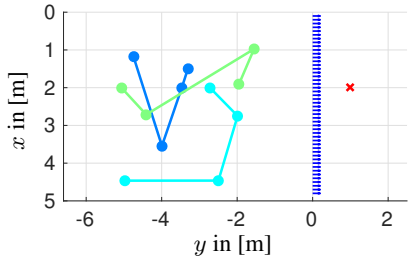
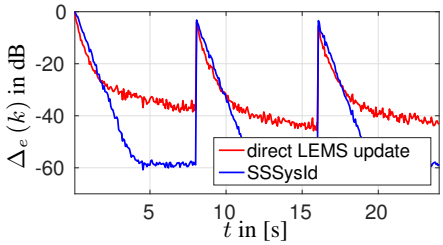


Fig. 5: Example for efficient identification of an LEMS by identifying source-specific systems $\mathbf{H}_\mathrm{S}$ during intervals of constant source configuration (constant $\mathbf{H}_\mathrm{D}$) and knowledge transfer between different intervals by means of a background model of the LEMS, $\hat{\mathbf{H}}$, where the identified system components accumulate.

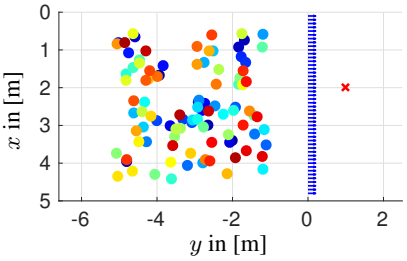


(a) Setup: $N_\mathrm{L} = 48$ loudspeakers (**blue** arrows), $N_\mathrm{M} = 1$ microphone (**red** cross), and 3 randomly chosen groups of 4 virtual sources. Their positions are marked by dots and are connected by a line to symbolize their simultaneous activity.

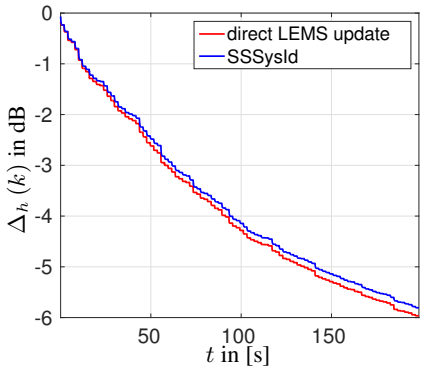


(b) Normalized residual error signal at the microphone resulting during Experiment 1.

Fig. 6: Geometrical setup, virtual scenes, and normalized residual error of system identification for Experiment 1.



(a) Setup: $N_\mathrm{L} = 48$ loudspeakers (**blue** arrows), $N_\mathrm{M} = 1$ microphone (**red** cross), and 100 randomly chosen virtual source positions.



(b) System error norm achievable during Experiment 2.

Fig. 7: Geometrical setup, virtual scenes, and normalized system error norm for Experiment 2.

Experiment 2: In this experiment, the long-term stability of SSSysId is studied. To this end, 100 different virtual source positions are drawn with coordinates $\vec{x}_\mathrm{S} = [x, y, 0]^\mathrm{T}$, $x \in [0.5, 4.5]$, $y \in [-5.1, -1.1]$ (depicted in Fig. 7a) and each source is exclusively active in its own interval of length 2 s. This corresponds to 99 source configuration changes. The adaptation of source-specific systems and the direct adaptation of the LEMS is compared in terms of the normalized system error norms. These are depicted in Fig. 7b for each of the 100 intervals (determined at the respective intervals' ends). Obviously, the less complex source-specific updates (**blue** curve) lead to a completely stable adaptation and similar performance as updating the LEMS directly (**red** curve), also in case of repeatedly changing virtual source configurations and for excitation with just a single virtual source. However, a slightly increased normalized system error norm (after 99 scene changes: -5.81 dB vs. -5.96 dB) is the result of the repeated transforms with regularized inverse rendering filters and the truncation of the convolution results to the modeled filter lengths - a small cost for a complexity reduction of about $N_\mathrm{L}/N_\mathrm{S} = 48$ for the adaptive filtering.

4. CONCLUSION

A novel method has been proposed for identifying a MIMO system employing side information (statistically independent virtual source signals, rendering filters) from an object-based rendering system (e.g., WFS or hands-free communication using a multi-loudspeaker front-end). This method does not impose any constraints on the positions of the transducers (N_L loudspeakers and N_M microphones). As opposed to state-of-the-art methods, this approach has predictably low computational complexity, independent of the spectral or spatial characteristics of the N_S virtual sources. For long intervals of constant virtual source configuration, a reduction of the average complexity by a factor of about $N_\mathrm{L}/N_\mathrm{S}$ is possible. A prototype has been simulated in order to verify the concept exemplarily for the identification of an LEMS for WFS with a linear sound bar. In future, a further reduction of the computational load peaks at the scene change may be pursued.

5. REFERENCES

[1] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[2] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society*, vol. 33, no. 11, pp. 859–871, 1985.

[3] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *Journal of the Acoustical Society of America (JASA)*, vol. 93, no. 5, pp. 2764–2778, 1993.

[4] E. N. G. Verheijen, "Sound reproduction by wave field synthesis," Ph.D. dissertation, Delft University of Technology (TU Delft), 1998.

[5] M. Sondhi, D. Morgan, and J. Hall, "Stereophonic acoustic echo cancellation-an overview of the fundamental problem," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 148–151, August 1995.

[6] J. Benesty, D. Morgan, and M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 156–165, 1998.

[7] H. Buchner, J. Benesty, and W. Kellermann, "Generalized multichannel frequency-domain adaptive filtering: Efficient realization and application to hands-free speech communication," *Signal Processing*, vol. 85, no. 3, pp. 549–570, March 2005.

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Johns Hopkins University Press, 1996.

[9] T. Gaensler and P. Eneroth, "Influence of audio coding on stereophonic acoustic echo cancellation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 6, Seattle, USA, May 1998, pp. 3649–3652.

[10] S. Shimauchi *et al.*, "A stereo echo canceller implemented using a stereo shaker and a duo-filter control system," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Phoenix, USA, March 1999, pp. 857–860.

[11] D. Morgan, J. Hall, and J. Benesty, "Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 686–696, Sep 2001.

[12] J. Herre, H. Buchner, and W. Kellermann, "Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, USA, April 2007.

[13] M. Schneider, C. Huemmer, and W. Kellermann, "Wave-domain loudspeaker signal decorrelation for system identification in multichannel audio reproduction scenarios," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 605–609.

[14] J. Wung, T. Wada, M. Souden, and B.-H. Juang, "Inter-channel decorrelation by sub-band resampling for multi-channel acoustic echo cancellation," *IEEE Transactions Signal Processing*, vol. 62, no. 8, pp. 2127–2142, April 2014.

[15] H. Buchner, S. Spors, and W. Kellermann, "Wave-domain adaptive filtering: Acoustic echo cancellation for full-duplex systems based on wave-field synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, Montreal, Canada, May 2004, pp. iv–117 – iv–120.

[16] S. Spors, H. Buchner, and R. Rabenstein, "A novel approach to active listening room compensation for wave field synthesis using wave-domain adaptive filtering," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, Montreal, Canada, May 2004, pp. iv–29 – iv–32.

[17] M. Schneider and W. Kellermann, "Apparatus and method for providing a loudspeaker-enclosure-microphone system description," Patent Application WO 2014/015 914 A1, January 30, 2014.

[18] K. Helwani, H. Buchner, and S. Spors, "Source-domain adaptive filtering for MIMO systems with application to acoustic echo cancellation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA, 2010, pp. 321–324.

[19] S. Spors, H. Buchner, and R. Rabenstein, "Eigenspace adaptive filtering for efficient pre-equalization of acoustic MIMO systems," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, vol. 6, Florence, Italy, 2006.

[20] K. Helwani and H. Buchner, "On the eigenspace estimation for supervised multichannel system identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013, pp. 630–634.

[21] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *Audio Engineering Society Convention 124*, Amsterdam, Netherlands, 2008, pp. 17–20.

[22] G. Strang, *Introduction to Linear Algebra*, 4th ed. Wellesley - Cambridge, 2009.