Large-scale Multiple Input/Multiple Output system identification in room acoustics

Martin Schneider* and Walter Kellermann

*Corresponding author's address: University Erlangen-Nuremberg, Erlangen, 91058, Bavaria, Germany, schneider@lnt.de

In audio reproduction scenarios, room acoustics may be described as a multiple-input/multiple-output system response from multiple loudspeaker to multiple microphones in the listening space. This system response may, e.g., be used for an equalization of the listening room and must be identified from observing the available loudspeaker and microphone signals in real-world systems. For few transducers this task is mostly solved, but massive multichannel reproduction with dozens to hundreds of loudspeakers left many research questions open. This contribution points out the fundamental challenges, previous solutions and recent advances. As a key issue, the so-called nonuniqueness problem for multiple-input/multiple-output system identification by adaptive filtering will be discussed along with decorrelation schemes for the loudspeaker signals to alleviate this problem. Successful adaptation algorithms suitable for these scenarios imply considerable computational demands and require additional measures to ensure robustness. Recently emerging system models in spatial transform domains allow for approximative models and seem to be promising for robust real-time implementations.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

The identification of acoustic systems plays a crucial role for various highly relevant applications such as, e.g., acoustic echo cancellation (AEC) and active noise control. In the context of room acoustics, an identified system may also be used to determine pre-equalization filters for reproduction systems in order to compensate for undesired room properties such as reverberation. To this end, microphones are placed within the listening room and the resulting loudspeaker–enclosure-microphone-system (LEMS) is then identified to acquire knowledge regarding the presumed room acoustics for the listener. The equalizers are then determined, such that the cascade of equalizing filters and the identified LEMS approximates a desired multiple-input/multiple-output (MIMO) response. The latter typically describes the free-field impulse response or the actual room impulse response with an attenuated reverberation tail.

Loudspeaker arrays used for wave field synthesis (WFS) [1] or Higher Order Ambisonics (HOA) [2] provide control over the reproduced acoustic scene within an extended listening area. A complementation of those systems with listening room equalization (LRE) aims at an equalization within this listening area [3]. Typically, WFS or HOA systems are equipped with several tens to several hundreds of reproduction channels ($N_L$ in the following). In order to fully exploit the control over the reproduced wave field provided by the loudspeakers, there must be an equalization filter path connecting every loudspeaker signal input with every loudspeaker of the LEMS. Consequently, a large number of microphones is required such that the identified LEMS provides enough information to uniquely determine the resulting $N_L^2$ equalization filters. As the room properties may change due to varying temperature or attendance of the audience, system identification must be accomplished by continuously adapted filters such that the equalizing filters may also be continuously updated. The task of MIMO system identification becomes more challenging with an increasing number of loudspeakers [4, 5].

In this paper the difficulties of MIMO system identification for LEMSs comprising a large number of input and output channels are discussed. As explained in the second section of the paper, the strong cross-correlation of the loudspeaker signals often hampers a unique identification of the LEMS which results in the so-called nonuniqueness problem. A state-of-the-art remedy against the nonuniqueness problem is a decorrelation of the loudspeaker signals. The application of according methods to scenarios with a large number of reproduction channels is treated in the third section of this contribution. The fourth section motivates the choice of the generalized frequency-domain adaptive filtering (GFDAF) algorithm for MIMO system identification. A way to reduce the computational effort for MIMO system identification besides the choice of the adaptation algorithm are approximate wave-domain LEMS models as discussed in the fifth section. Finally, the conclusions of this paper are presented in the last section.

THE NONUNIQUENESS PROBLEM

In this section, the properties of typical loudspeaker signals and the resulting non-uniqueness problem for MIMO system identification are discussed.

A signal model for the system identification task is shown in Fig. 1 using frequency-dependent vectors and matrices to describe signals and systems, respectively. The loudspeaker signal vector $x(j\omega) = (x_1(j\omega), x_2(j\omega), \ldots, x_{N_L}(j\omega))^T$ may be decomposed into $N_L$ loudspeaker signals $x_l(j\omega)$, where $l$ is the loudspeaker index, $\omega$ denotes the angular frequency and $j$ is used as the imaginary unit. Similarly, the $N_M$ microphone signals $d_m(j\omega)$, indexed by $m$, are captured by $d(j\omega) = (d_1(j\omega), d_2(j\omega), \ldots, d_{N_M}(j\omega))^T$ and are...
obtained by MIMO filtering of the loudspeaker signals which is expressed as
\[ d(j\omega) = H(j\omega)x(j\omega), \tag{1} \]
where the \( N_M \times N_L \) matrix \( H(j\omega) \) captures the frequency responses of the LEMS from all loudspeakers to all microphones. The influence of local acoustic sources and noise present in the LEMS is neglected in this consideration. An estimate \( \hat{H}(j\omega) \) for \( H(j\omega) \) can be obtained as Wiener solution minimizing \( e(j\omega) \) in the mean square error (MSE) sense. For MIMO systems an optimal Wiener solution can be determined by solving
\[ \hat{H}(j\omega)S_{xx}(j\omega) = S_{dx}(j\omega), \tag{2} \]
where \( S_{xx}(j\omega) \) describes the auto power spectral density (PSD) matrix of the signals captured by \( x(j\omega) \) and \( S_{dx}(j\omega) \) the cross PSD matrix of the signals captured by \( d(j\omega) \) and \( x(j\omega) \). In (2), \( H(j\omega) \) may be written as
\[ H(j\omega)S_{xx}(j\omega) = H[j\omega]S_{xx}(j\omega). \tag{3} \]
The unique perfect solution for system identification can be obtained by inverting \( S_{xx}(j\omega) \) if \( S_{xx}(j\omega) \) is non-singular. This solution is given by \( \hat{H}(j\omega) = H(j\omega) \) and fulfills (2) and (3) independently of \( S_{xx}(j\omega) \). As the system must be identified during reproduction, the signals present at the loudspeakers and therefore \( S_{xx}(j\omega) \) are determined by the acoustic scene presented to the listeners. Typical reproduction signals describe natural acoustic scenes comprising, e.g., several sources of music or speech signals, here described by the vector \( p(j\omega) = (p_1(j\omega), \ldots, p_{N_S}(j\omega))^T \) capturing the signals of all \( N_S \) sources. The resulting loudspeaker signals may be described by \( x(j\omega) = G(j\omega)p(j\omega) \), where the frequency responses captured by \( G(j\omega) \) describe the reproduction system. For WFS, \( G(j\omega) \) describes filters such that the wave field of the virtual sources corresponding to the signals in \( p(j\omega) \) is synthesized. For a recorded signal, \( G(j\omega) \) describes the frequency responses from the physical sources to the recording microphones. Using the PSD matrix \( S_{pp}(j\omega) \) of the signals captured by \( p(j\omega) \), one obtains
\[ S_{xx}(j\omega) = G(j\omega)S_{pp}(j\omega)G^H(j\omega), \tag{4} \]
with \((\cdot)^H\) denoting the conjugate transpose. The rank of \( S_{xx}(j\omega) \) is limited by the dimensions of \( S_{pp}(j\omega) \). Consequently, whenever the number \( N_S \) of recorded physical sources or reproduced virtual sources is smaller than the number \( N_L \) of reproduction channels, \( S_{xx}(j\omega) \) becomes singular. This is referred to as nonuniqueness problem [4] and has been a topic to research for several years [6, 7]. In this case, infinitely many \( \hat{H}(j\omega) \) fulfill (2) and (3), but any obtained \( \hat{H}(j\omega) \) is only valid for the present correlation properties described by \( S_{xx}(j\omega) \). Whenever these change, as it usually occurs for nonstationary signals in practice, the previously determined \( \hat{H}(j\omega) \) will no longer be optimal. Consequently, any technique relying on an identified system \( \hat{H}(j\omega) \) suffers from a lack of robustness to changing correlation properties of the loudspeaker signals under
these conditions [8]. For systems with a large number $N_L$ of reproduction channels, the nonuniqueness problem becomes severe because the number of loudspeakers $N_L$ will often exceed the number of sources $N_S$. Moreover, even if there are more sources than loudspeakers, a large condition number of $S_{xx}(j\omega)$, inherited from $G(j\omega)$, may render the problem close to underdetermined.

**DECORRELATION OF LOUDSPEAKER SIGNALS**

In this section, different approaches for the decorrelation of the loudspeaker signals in order to improve the conditioning of the matrix $S_{xx}(j\omega)$ are reviewed with respect to their applicability to a large number of reproduction channels and high quality audio reproduction. For decorrelation, the loudspeaker signals are altered differently for each reproduction channel, and the altered signals are then fed to the LEMS replacing the unaltered $x(j\omega)$. Essentially, there are three obvious options to modify a pair of signals in order to reduce correlation between them: The addition of uncorrelated noise signals, different nonlinear preprocessing, and different time-variant processing.

As a first straightforward solution, the addition of mutually uncorrelated noise signals to the loudspeaker channels is discussed [4, 9, 10]. When considering $S_{xx}(j\omega)$, the effect of this can simply be described by the addition of a diagonal matrix, which ensures invertibility. As mutually uncorrelated noise signals can be easily generated, even for a large number of signals, this approach may easily be extended for any reproduction system. However, noise signals are prone to be audible, which disqualifies this approach for high-quality audio reproduction.

Another approach is to apply different non-linear functions to each of the loudspeaker channels [6, 11]. This generates different additional components in each of the mutually correlated loudspeaker signals. These components are not statistically independent but uncorrelated, which is sufficient to improve system identification using second order statistics. As nonlinear functions must differ sufficiently for each channel to achieve good results, an extension to an arbitrary number of channels is difficult. Further, this approach was essentially proposed for speech signals, rather than for music signals because a non-linear distortion will generally not be acceptable for music signals.

Approaches better suited for high-quality audio reproduction are based on phase-modulation [12, 13] and resampling [14]. The phase modulation causes time-varying cross-correlation properties of loudspeaker signals, described by different $S_{xx}(j\omega)$. If $H(j\omega)$ fulfills (2) for a large number of different $S_{xx}(j\omega)$, $\hat{H}(j\omega)$ is close to the perfect solution $H(j\omega) = H(j\omega)$ which is valid independently of $S_{xx}(j\omega)$. Resampling-based approaches are closely related to phase modulation as the different time-scaling introduced there for the individual loudspeaker channels has a similar influence on the phase of the signals. For both, phase modulation and sampling-based approaches, the signals may only be altered to a certain extent in order to retain high audio quality. On the other hand, the variation of the loudspeaker signals necessary for an effective decorrelation will increase with an increasing number of loudspeaker channels. Consequently, the improvement in convergence speed for system identification is likely to decrease for a large number of channels.

In general, experimental evaluations for a large number of channels are still missing as most publications only address systems with two channels [4, 6, 9, 10, 11, 14] and only a few a larger number of channels as, e.g., five channels in [13]. Hence, an important dimension for comparison of the proposed approaches is disregarded: The extendability towards a large number of channels. Compared to others, a decorrelation method may achieve a lesser improvement of the convergence speed, but it might be more easily
Another difficulty arises for any time-varying filtering in the context of WFS or Ambisonics. There, the loudspeaker signals are analytically determined and any alteration of the phase, would distort the reproduced wave field. For those reproduction techniques, a model-based approach achieving a controlled alteration of the reproduced wave field can be better suited. A slight time-varying rotation of the reproduced wave field might, e.g., be perceptually acceptable and can be used for a decorrelation of the loudspeaker signals [15].

**Adaptation Algorithms for System Identification**

In this section, different adaptation algorithms are reviewed considering their suitability for MIMO system identification. To this end, the discrete-time representations of \( x_l(j\omega), \) \( d_m(j\omega), \) and \( e_m(j\omega) \) are considered, which are given by \( x_l(k), d_m(k), \) and \( e_m(k), \) where \( k \) represents the time instant. Those values are grouped into vectors according to

\[
x(n) = [x_1^T(n), x_2^T(n), \ldots, x_{N_L}^T(n)]^T
\]

\[
x_l(n) = [x_l(nL_F - L_X + 1), x_l(nL_F - L_X + 2), \ldots, x_l(nL_F)]^T,
\]

\[
d(n) = [d_1^T(n), d_2^T(n), \ldots, d_{N_M}^T(n)]^T
\]

\[
d_m(n) = [d_m(nL_F - L_D + 1), d_m(nL_F - L_D + 2), \ldots, d_m(nL_F)]^T,
\]

\[
e(n) = [e_1^T(n), e_2^T(n), \ldots, e_{N_M}^T(n)]^T
\]

\[
e_m(n) = [e_m(nL_F - L_D + 1), e_m(nL_F - L_D + 2), \ldots, e_m(nL_F)]^T,
\]

where \( L_X \) is length of the vector \( x_l(n) \) and \( L_D \) is the length of the vectors \( d_m(n) \) and \( e_m(n) \), \( L_F \) describes the frame shift for block processing and \( n \) denotes the block index.

A popular and simple adaptation algorithm for system identification is the normalized least-mean-squares (NLMS) algorithm, which can also be formulated for MIMO system identification. Applying the NLMS to each channel individually leads to slow convergence when the signals are strongly autocorrelated [16]. The situation worsens if the multichannel NLMS is applied to additionally strongly cross-correlated loudspeaker signals [7]. Finally, this algorithm is only suitable if the excitation signal for system identification can be chosen according a so-called perfect sequence [17].

The recursive least-squares (RLS) algorithm can achieve a fast convergence even for strongly cross- and autocorrelated signals. Considering \( L_X = L_H, L_D = 1, \) and \( n = k \) this algorithm is described by [16]:

\[
\hat{H}(k) = \hat{H}(k - 1) + R_{xx}^{-1}(k)x(k)e^T(k)
\]

\[
e(k) = d(k) - \hat{H}^T(k - 1)x(k)
\]

\[
R_{xx}(k) = \lambda R_{xx}(k - 1) + (1 - \lambda)x(k)x^H(k),
\]

where \( \lambda \) is an exponential forgetting factor. \( \hat{H}(k) \) is the length-\( L_H \) discrete-time representation of \( \hat{H}(j\omega) \). Even though the inverse of the auto-correlation matrix estimate \( R_{xx}(k) \) may be efficiently calculated in a recursive manner, its large dimensions (i.e. \( L_H \cdot N_L \times L_H \cdot N_L \)) make an application of the RLS impracticable for the considered scenarios.

There are various adaptation algorithms which could also be used for system identification, such as the proportionate NLMS or the affine projection algorithm [7]. Another option is the GFDAF algorithm, which exploits a DFT-domain approximation to
allow a very efficient implementation [18]. Considering \( L_X = 2L_H, L_F = L_D = L_H \), this algorithm reads

\[
\begin{align*}
X(n) &= \text{Diag} \{ Fx_1(n) \}, \text{Diag} \{ Fx_2(n) \}, \ldots, \text{Diag} \{ Fx_{N_L}(n) \} \\
e_m(n) &= d_n(n) - [0, I]^T X(n) \tilde{h}_m(n - 1), \\
\tilde{h}_m(n) &= \tilde{h}_m(n - 1) + \mu(1 - \lambda) S^{-1}(n) X^H(n) F [0, I]^T e_m(n) \\
S(n) &= \lambda S(n - 1) + (1 - \lambda) \frac{1}{2} X^H(n) X(n)
\end{align*}
\]

where \( 0 < \mu < 1 \) is a stepsize parameter, \( F \) is the \( 2L_H \times 2L_H \) unitary DFT matrix, \( I \) is the \( L_H \times L_H \) unit matrix, \( 0 \) describes an \( L_H \times L_H \) matrix with zero-valued entries and \( \tilde{h}_m(n) \) is the DFT-domain representation of \( H(j\omega) \) with respect to the microphone \( m \). Different to \( R_{xx}(k) \), the DFT-domain estimate of \( S_{xx}(j\omega) \), \( S(n) \) is a sparse matrix and its inversion may be separated into \( 2L_H \) inversions of \( N_L \times N_L \) matrices. While reducing the computational effort drastically, this property also allows a straightforward massive parallelization as, e.g., advantageous for an implementation on a graphics processing unit [19]. Additionally, the numerical stability is significantly increased in comparison to the RLS algorithm due to the inversion of smaller matrices. Still, as \( S_{xx}(j\omega) \) may be singular or ill-conditioned, \( S(n) \) inherits this property and an \( N_L \times N_L \) regularization must be applied to maintain robustness of this algorithm. A loudspeaker signal decorrelation, as discussed in the previous section also increases robustness in this regard. Overall, the GFDAF provides a good trade-off between computational complexity and sufficiently fast convergence which qualifies the algorithm as a candidate for a real-time implementation of MIMO system identification. Still, known implementations of the GFDAF have a computational complexity proportional to \( N_L^3 \) and a reduction below a value proportional to \( N_L^2 \) is not possible. The sampling rate and filter length \( L_H \) necessary for an effective LRE finally lead to a computational effort beyond the capabilities of a workstation, which is a potential platform for an LRE system. This motivates the use of approximative models for the LEMS allowing a further reduction of the computational complexity, as represented in the following section.

**APPROMMATING MODELS IN THE WAVE DOMAIN**

In this section, approximative wave-domain models for the LEMS are discussed. Conventional LEMS models describe the frequency responses between each loudspeaker and microphone (as captured by \( H(j\omega) \)) represented by discrete-time impulse responses. For LEMSs with a large number of transducers, the relative position of the transducers is often known, although this knowledge is not exploited in conventional models. In contrast, wave-domain models for the LEMS exploit this knowledge by using fundamental solutions of the wave equation as signal representations. Those basis functions may, e.g., be plane waves [20], spherical harmonics, or circular harmonics [21]. Using the latter, the sound pressure \( P(\alpha, \varrho, j\omega) \) at a point described by the angle \( \alpha \) and the radius \( \varrho \) in polar coordinates may be described by

\[
P(\alpha, \varrho, j\omega) = \sum_{\tilde{m} = -\infty}^{\infty} \left( \tilde{P}_{\tilde{m}}^{(1)}(j\omega) H_{\tilde{m}}^{(1)} \left( \frac{\omega}{c} \varrho \right) + \tilde{P}_{\tilde{m}}^{(2)}(j\omega) H_{\tilde{m}}^{(2)} \left( \frac{\omega}{c} \varrho \right) \right) e^{j\tilde{m}\alpha},
\]

where \( H_{\tilde{m}}^{(1)}(x) \) and \( H_{\tilde{m}}^{(2)}(x) \) are Hankel functions of the first and second kind and order \( \tilde{m} \), respectively and \( c \) is the speed of sound.

The signal model for wave-domain system identification is shown in Fig. 2. There, the loudspeaker signals \( x(j\omega) \) are transformed to the wave domain by describing the wave

M. Schneider and W. Kellermann

The problems of MIMO system identification, mainly due to the necessary computational effort and the nonuniqueness problem, magnify with an increasing...
number of loudspeakers channels. Solutions, were often proposed considering only a few reproduction channels and can not be straightforwardly applied to scenarios with many channels. Decorrelation methods, as a remedy for the nonuniqueness problem, have mostly been investigated for only two channels so their applicability to a large number of reproduction channels remains a topic to future research. Efficient adaptation algorithms tailored for MIMO system identification are already available and may be complemented by transform-domain LEMS models for a further improved computational efficiency. Thus, the first steps towards realizable real-time implementations are already taken. This could fuel future work exploring the number of reproduction channels as a further dimension in the development of algorithms.

ACKNOWLEDGMENTS

The work of M. Schneider and W. Kellermann was performed and funded as part of a joint research project with the Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany.

REFERENCES


