ADAPTIVE LISTENING ROOM EQUALIZATION USING A SCALABLE FILTERING STRUCTURE IN THE WAVE DOMAIN

Martin Schneider and Walter Kellermann

Multimedia Communications and Signal Processing*
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
mail: {schneider,wk} @LNT.de

ABSTRACT

Massive multichannel reproduction systems like wave field synthesis (WFS) are potentially well suited to be complemented by listening room equalization (LRE). However, their typically large number of reproduction channels makes this task challenging for both computational and algorithmic reasons. Wave-domain adaptive filtering (WDAF) was proposed earlier and is especially well-suited to adaptive filtering tasks in the context of WFS. In this paper, we propose to generalize the model originally used for WDAF to allow an adaptive LRE for a broader range of reproduction scenarios, while maintaining the advantages of the original approach. The proposed approach is evaluated for filtering structures of varying complexity along with considering the robustness to varying listener positions.

Index Terms— wave domain, adaptive filtering, listening room compensation, listening room equalization

1. INTRODUCTION

WFS [1] is used to achieve a highly detailed spatial reproduction of an acoustic scene overcoming the limitations of a sweet spot by using an array of typically several tens to hundreds of loudspeakers. The loudspeaker signals for WFS are usually determined assuming free-field conditions. As a consequence, an enclosing room must not exhibit significant wall reflections to avoid a distortion of the synthesized wave field. In many scenarios, the necessary acoustic treatment to achieve such room properties may be too expensive or impractical. An alternative to acoustical countermeasures is to compensate for the wall reflections by means of an listening room equalization (LRE), often termed listening room compensation. To this end, the reproduction signals are filtered to pre-equalize the MIMO room system response from the loudspeakers to the positions of multiple microphones, ideally achieving an equalization at any point in the listening area.

Although, the precise spatial control over the synthesized wave field makes a WFS system particularly suitable for LRE, its many reproduction channels constitute a major challenge for the development of such a system. As the MIMO loudspeaker-enclosure-microphone system (LEMS) must be expected to change over time, it has to be continuously identified by adaptive filtering. As known from acoustic echo cancellation (AEC), this problem may be underdetermined or at least ill-conditioned when using multiple reproduction channels [2]. Additionally, the inverse filtering problem underlying LRE must be expected to be ill-conditioned as well. Besides these algorithmic problems, the large number of reproduction channels also leads to a large computational effort for both the system identification and the determination of the equalizing prefilters. As the MIMO system response of the LEMS can only be measured for the microphone positions, while equalization should be achieved in the entire listening area, the spatial robustness of the solution for the equalizers has to be ensured additionally. The classical LRE aimed for an equalization at multiple points in the listening room [3]. Since this approach disregards the wave propagation, the obtained results suffered from a low spatial robustness. With WDAF, an approach was presented which considers the wave propagation and exhibits therefore an improved spatial robustness [4]. In [4], it was also shown that in the wave domain, the LEMS may be approximated so that a very simple equalizer structure results. Then system identification is never an underdetermined problem, but there may be a residual error due to model restrictions. In any case, this simplified model does not suffice for every reproduction scenario, as shown below.

In this paper, we generalize the approach in [4] using a more flexible LEMS model combined with a more flexible equalizer structure. The complexity of both can be chosen to allow a trade-off between the suitability for differently complex reproduction scenarios on one side and robustness and computational demands on the other side. The paper is organized as follows: In Sec. 2 the structure of the proposed LRE system is explained. The used wave-domain signal model is briefly presented in Sec. 3, the evaluation results and the conclusions follow in Sections 4 and 5, respectively.

2. STRUCTURE OF THE LRE SYSTEM

The structure of the proposed LRE system in the wave domain is depicted in Fig. 1. The upper part in Fig. 1 is dedicated to the identification of the acoustic MIMO system in the wave domain. The obtained knowledge is then used in the lower part to determine the equalizers accordingly. In contrast to [4], these steps are separated to allow the use of the generalized equalizer structure. Regarding system identification , the input signal of the system is given by the loudspeaker signal vector $\mathbf{x}(n)$ containing a block (indexed by n) of L_X time-domain samples of all N_L loudspeaker signals:

$$\begin{aligned}\mathbf{x}(n) = (&x_1(nL_F - L_X + 1), \ldots, x_1(nL_F), \\ &x_2(nL_F - L_X + 1), \ldots, x_2(nL_F), \ldots, x_{N_L}(nL_F)),\end{aligned} \tag{1}$$

where $x_\lambda(k)$ is a time-domain sample of the loudspeaker signal λ at time instant k and L_F is the frame shift. All considered signal vectors are structured in the same way, but may differ in their lengths

*The work of M. Schneider and W. Kellermann was supported by the Fraunhofer Institute for Digital Media Technology (IDMT) in Ilmenau, Germany.

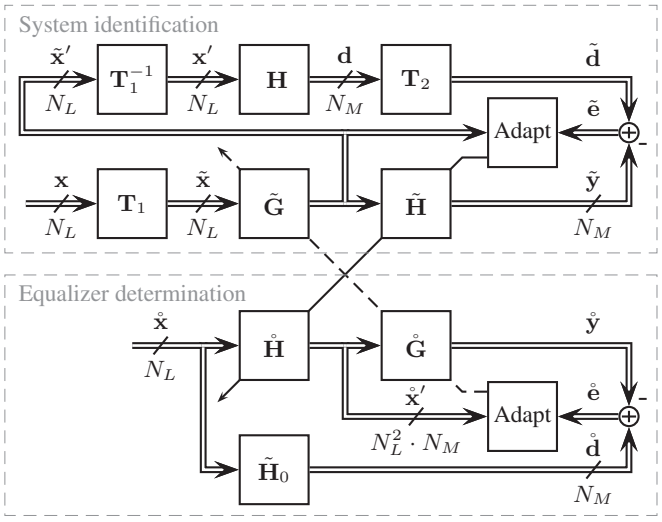

Fig. 1. Block diagram of an LRE system in the wave domain. $\mathbf{T}_1$, $\mathbf{T}_2$,$\mathbf{T}_1^{-1}$: Transforms to and from the wave domain, $\mathbf{H}$: system response of the LEMS, $\tilde{\mathbf{H}}$, $\mathring{\mathbf{H}}$: identified LEMS, $\mathring{\mathbf{G}}$, $\tilde{\mathbf{G}}$: equalizers, $\tilde{\mathbf{H}}_0$: desired free-field response. The dependence of the block index n of different quantities is omitted for convenience.

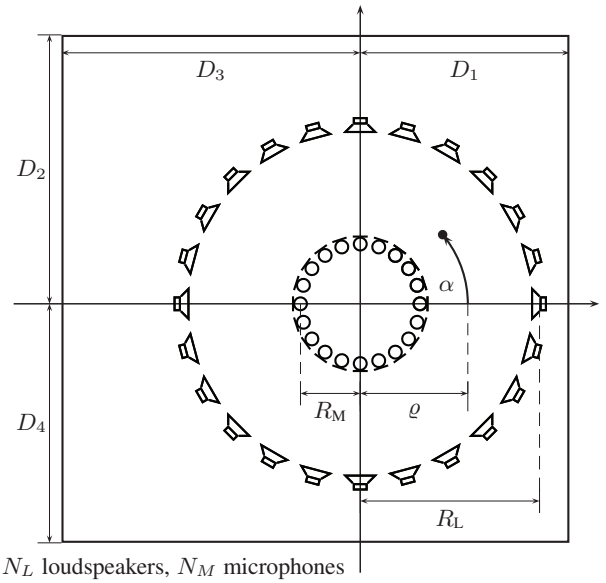

Fig. 2. Loudspeaker and microphone setup and in the LEMS

and numbers of components. Transform $\mathbf{T}_1$ is used to obtain the so-called free-field representation $\tilde{\mathbf{x}}(n) = \mathbf{T}_1\mathbf{x}(n)$ and will be explained in Sec. 3 together with $\mathbf{T}_2$. The equalizers in $\tilde{\mathbf{G}}(n)$ are copies of the filters in $\mathring{\mathbf{G}}(n)$ and are used to obtain the equalized loudspeaker signals $\tilde{\mathbf{x}}'(n) = \tilde{\mathbf{G}}(n)\tilde{\mathbf{x}}(n)$ in the wave domain. These are then transformed back and fed to the LEMS $\mathbf{H}$ from which we obtain the N_M microphone signals contained in $\mathbf{d}(n) = \mathbf{H}\mathbf{x}'(n)$. The matrix $\mathbf{H}$ is structured so that

$$d_\mu(k) = \sum_{\kappa=0}^{L_H-1} x'_\lambda(k-\kappa)h_{\mu,\lambda}(\kappa), \tag{2}$$

where $h_{\mu,\lambda}(k)$ describes the room impulse response of length L_H from loudspeaker λ to microphone μ. All other considered matrices are of similar structure. To identify the LEMS by $\tilde{\mathbf{H}}(n)$ in the wave domain, we transform the microphone signals to the measured wave field $\tilde{\mathbf{d}}(n) = \mathbf{T}_2\mathbf{d}(n)$ and determine the wave-domain error $\tilde{\mathbf{e}}(n)$ as the difference between $\tilde{\mathbf{d}}(n)$ and its estimate $\tilde{\mathbf{y}}(n) = \tilde{\mathbf{H}}(n)\mathbf{x}'(n)$. For the adaptation of $\tilde{\mathbf{H}}(n)$, the squared error $\tilde{\mathbf{e}}^H(n)\tilde{\mathbf{e}}(n)$ is minimized.

For the determination of the equalizers we use the free-field description of the loudspeaker signals as input $\mathring{\mathbf{x}}(n) = \tilde{\mathbf{x}}(n)$ (noise could also be used, see [5]). The signals are filtered by $\mathring{\mathbf{H}}(n)$ which contains the copied coefficients from $\tilde{\mathbf{H}}(n)$, although the output vector $\mathring{\mathbf{x}}'(n) = \mathring{\mathbf{H}}(n)\mathring{\mathbf{x}}(n)$ is structured differently: it contains all $N_L^2 \cdot N_M$ possible combinations of filtering the N_L signal components in $\mathring{\mathbf{x}}(n)$ with the $N_L \cdot N_M$ impulse responses contained in $\tilde{\mathbf{H}}(n)$. This is necessary for the multichannel filtered-X generalized frequency domain adaptive filtering (GFDAF) as described in [5] for conventional (not wave-domain) equalization. The N_L^2 filters in $\mathring{\mathbf{G}}(n)$ are then adapted so that $\mathring{\mathbf{y}}(n) = \mathring{\mathbf{G}}(n)\mathring{\mathbf{x}}'(n)$ approximates the desired signal $\mathring{\mathbf{d}}(n) = \tilde{\mathbf{H}}_0\mathring{\mathbf{x}}(n)$ which is obtained by filtering $\mathring{\mathbf{x}}(n)$ with the free-field response $\tilde{\mathbf{H}}_0$ in the wave domain. The error $\mathring{\mathbf{e}}(n) = \mathring{\mathbf{y}}(n) - \mathring{\mathbf{d}}(n)$ is squared and $\mathring{\mathbf{e}}^H(n)\mathring{\mathbf{e}}(n)$ is used as an optimization criterion for adapting $\mathring{\mathbf{G}}(n)$.

Regarding adaptation algorithms, the GFDAF algorithm (as described in [6] for AEC) has been used for the system identification in the wave domain i. e. the adaptation of $\tilde{\mathbf{H}}(n)$. For the adaptation of $\mathring{\mathbf{G}}(n)$, the filtered-X GFDAF was used with $\mathring{\mathbf{x}}'(n)$ as filter input, according to [5].

3. WAVE-DOMAIN SIGNAL MODEL

In this section the wave-domain representations of the involved signals and systems are introduced. For two concentric uniform circular arrays, i. e. a loudspeaker array enclosing a microphone array with smaller radius as depicted in Fig. 2. For this planar array setup, we use the so-called circular harmonics[6] as basis functions for the signal representations. This approach is similar to [7], but instead of a perfect steady state equalization we aim for a computationally efficient adaptive equalization. The spectrum of the sound pressure $P(\alpha, \varrho, j\omega)$ at any point $\vec{x} = (\alpha, \varrho)^T$ is then given by a sum of circular harmonics

$$P(\alpha,\varrho,j\omega) = \sum_{l=-\infty}^{\infty} \left(\tilde{P}_l^{(1)}(j\omega)\mathcal{H}_l^{(1)}\left(\frac{\omega}{c}\varrho\right) + \tilde{P}_l^{(2)}(j\omega)\mathcal{H}_l^{(2)}\left(\frac{\omega}{c}\varrho\right)\right) e^{jl\alpha}, \tag{3}$$

where $\mathcal{H}_l^{(1)}(x)$ and $\mathcal{H}_l^{(2)}(x)$ are Hankel functions of the first and second kind and order l, respectively. The angular frequency is denoted by ω, c is the speed of sound, and j is used as the imaginary unit. The quantities $\tilde{P}_l^{(1)}(j\omega)$ and $\tilde{P}_l^{(2)}(j\omega)$ may be interpreted as the spectra of incoming and outgoing waves, so we obtain $\tilde{P}_l^{(1)}(j\omega) = \tilde{P}_l^{(2)}(j\omega)$, when there are no acoustic sources inside the circumference of the array.

Transform $\mathbf{T}_1$ is used to obtain the so-called free-field description $\tilde{\mathbf{x}}(n)$, which describes N_L components of the wave field according to (3), as it would be ideally exited by the N_L loudspeakers when driven with the loudspeaker signals $\mathbf{x}(n)$ under free-field conditions. The obtained wave-field components are identified by their mode order l as they are related to the array as a whole. Equivalently, the components of the pre-equalized wave-domain loudspeaker signals $\tilde{\mathbf{x}}'(n)$ are indexed by l'.

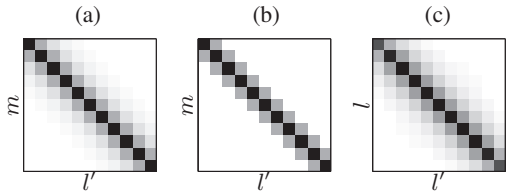

Fig. 3. Exemplary illustration of LEMS model and resulting equalizer weights. (a) Weights of couplings in $\mathbf{T}_2\mathbf{H}\mathbf{T}_1^{-1}$, (b) couplings modeled in $\tilde{\mathbf{H}}(n)$ with $|m-l'| < 2$ ($N_D = 3$), (c) resulting weights of the equalizers $\tilde{\mathbf{G}}(n)$ considering only $\tilde{\mathbf{H}}(n)$.

To obtain the N_M components of the measured wave field in $\tilde{\mathbf{d}}(n)$, $\mathbf{T}_2$ is applied to the N_M actually measured microphone signals in $\mathbf{d}(n)$. Like $\mathbf{T}_1$, $\mathbf{T}_2$ is chosen so that the components in $\tilde{\mathbf{d}}(n)$ are described according to (3), with a mode order denoted by m instead of l. For the considered array setup and basis functions, it was shown that the spatial DFT over the loudspeaker and microphone indices may be used for $\mathbf{T}_1$ and $\mathbf{T}_2$ [6], rendering the transform of (3) from the temporal frequency domain to the time domain unnecessary. However, these frequency-independent transforms do not correct the frequency responses of the considered signals according to (3). For the system shown in Fig. 1 this is acceptable as the adaptive filters will implicitly model the differences in the frequency responses and all descriptions remain consistent.

The advantage of the wave-domain description is the immediate spatial interpretation of all signal quantities and filter coefficients, which can be exploited in various ways. In [6] an approximative model for the LEMS model was successfully used for a computationally efficient AEC. This approach exploits the fact that the couplings of the wave field components described by $\tilde{\mathbf{x}}'(n)$ and $\tilde{\mathbf{d}}(n)$ are significantly stronger for components with a low difference $|m - l'|$ in the mode order [6]. For AEC it has been shown that modelling the coupling with $l' = m$ alone is sufficient for scenarios where a WFS system is synthesizing the wave field of a single source [8], while this model is not sufficient when multiple virtual sources are active [6]. In the latter case, a systematic correction of the system behavior as necessary for LRE is not possible, as the actual beviour is not sufficiently modeled. Therefore we propose to generalize the LEM model described in [4] to a structure as shown in Fig. 3(b), which constitutes an approximation of the model shown in Fig. 3(a). The resulting weights of the equalizers in $\tilde{\mathbf{G}}(n)$ are illustrated in Fig. 3(c). Again, we approximate the structure of $\tilde{\mathbf{G}}(n)$ shown in Fig. 3(c) by the most important equalizers resulting in a structure identical to the one shown in Fig. 3(b).

4. EVALUATION

For evaluation of the proposed scheme, room impulse responses for $\mathbf{H}$ where calculated using a first order image source model for the setup depicted in Fig. 2 with $R_L = 1.5$m, $R_M = 0.5$m, $D_1 = D_4 = 2$m, $D_2 = D_3 = 3$m, $N_L = N_M = 48$ and a reflection factor of 0.9. The radii of the arrays were chosen so that the wave field in between the microphone and loudspeaker array circles may also be observed over a broad area. Operating at a sampling rate of $f_s = 2$kHz, we implicitly limit our considerations to frequencies where WFS effectively allows to control the synthesized wave field. Reproduction at higher frequencies is beyond the scope of this paper. The obtained impulse responses had a length of less than 64 samples, although the adaptive filters in $\tilde{\mathbf{H}}(n)$ were able to model a length of $L_H = 129$ samples. This choice for L_H accounts for an artificial

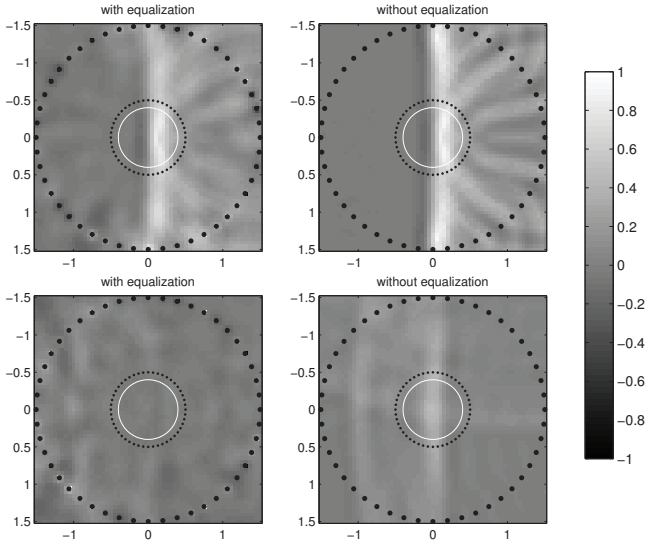

Fig. 4. Normalized sound pressure of synthesized plane wave within a room. The result with and without LRE is shown in the left and right column, respectively. The figures in the upper row show the direct component emitted by the loudspeakers, the figures in the lower row the portions reflected by the walls. The scale is meters.

delay of 40 samples introduced in $\tilde{\mathbf{H}}_0 = \mathbf{T}_2\mathbf{H}_0\mathbf{T}_1^{-1}$ to improve convergence (with $\mathbf{H}_0$ is describing the free-field response for the setup). The length of the equalizer impulse responses was chosen to $L_G = 256$ samples. For both GFDAF algorithms a forgetting factor of 0.95 and a frame shift of $L_F = 129$ samples were used. The normalized step size for the filtered-X GFDAF was 0.2.

To assess the achieved LRE, we calculated the difference of the actually measured wave field to the wave field under free-field conditions. The resulting value is then normalized to the value which would be obtained without equalization:

$$e_{\mathrm{MA}}(n) = 10\log_{10}\left(\frac{\left\|\left(\mathbf{T}_2\mathbf{H}\mathbf{T}_1^{-1}\tilde{\mathbf{G}}(n) - \tilde{\mathbf{H}}_0\right)\tilde{\mathbf{x}}(n)\right\|_2^2}{\left\|\left(\mathbf{T}_2\mathbf{H}\mathbf{T}_1^{-1}\tilde{\mathbf{I}} - \tilde{\mathbf{H}}_0\right)\tilde{\mathbf{x}}(n)\right\|_2^2}\right)\ \mathrm{dB}, \tag{4}$$

where $\tilde{\mathbf{I}}$ does not alter the signal, but ensures consistent vector lengths and $\|\cdot\|_2$ is the Euclidean norm. To assess the spatial robustness of the approach, we measure the error $e_{\mathrm{LA}}(n)$ within the listening area which is the area enclosed by the microphone array. The LRE error in the listening area $e_{\mathrm{LA}}(n)$ is determined in the same way as $e_{\mathrm{MA}}(n)$, but with a microphone array of a radius of $R_M = 0.4$m as shown by the white circle in Fig. 4.

The loudspeaker signals $\mathbf{x}$ were determined according to the theory of WFS, for simultaneously synthesizing three plane waves with the incidence angles $\varphi_1 = 0$, $\varphi_2 = \pi/2$, and $\varphi_3 = \pi$, where mutually uncorrelated white noise signals were used for the sources.

The evaluated structures differ in the number of modeled mode couplings in $\tilde{\mathbf{H}}(n)$ and corresponding equalizers in $\tilde{\mathbf{G}}(n)$. For each wave field component in $\tilde{\mathbf{x}}'(n)$ we modeled the couplings to N_D component in $\tilde{\mathbf{d}}(n)$ through $\tilde{\mathbf{H}}(n)$ according to $|m - l| < \mathrm{ceil}(N_D/2)$. The structure of the equalizers in $\tilde{\mathbf{G}}$ was chosen in the same way: for each mode in $\tilde{\mathbf{x}}(n)$ we determined equalizers to the N_D modes in $\tilde{\mathbf{x}}'(n)$ with $|l' - l| < \mathrm{ceil}(N_D/2)$.

In Fig.5 we can see the LRE errors over time for a system with $N_D = 3$. We can see that after a short phase of divergence the system stabilizes and converges towards an error of approximately

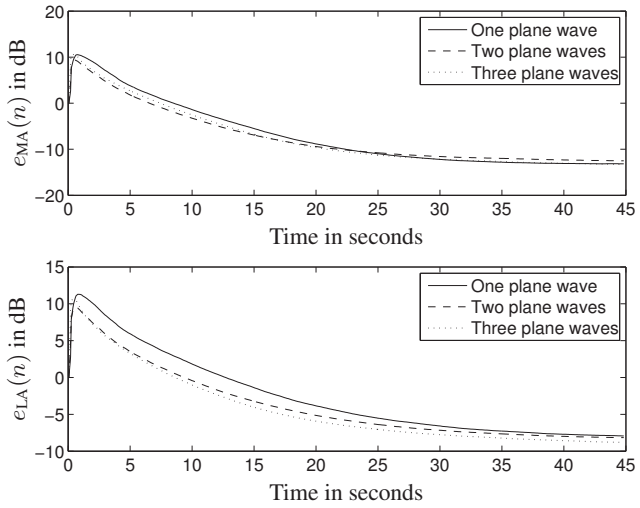

Fig. 5. Convergence over time for an LRE system with $N_D = 3$ for different scenarios. The upper plot shows the LRE performance at the microphone array, the lower plot within the listening area. Error at the microphone array: $e_{\mathrm{MA}}(n)$, error in the listening area $e_{\mathrm{LA}}(n)$

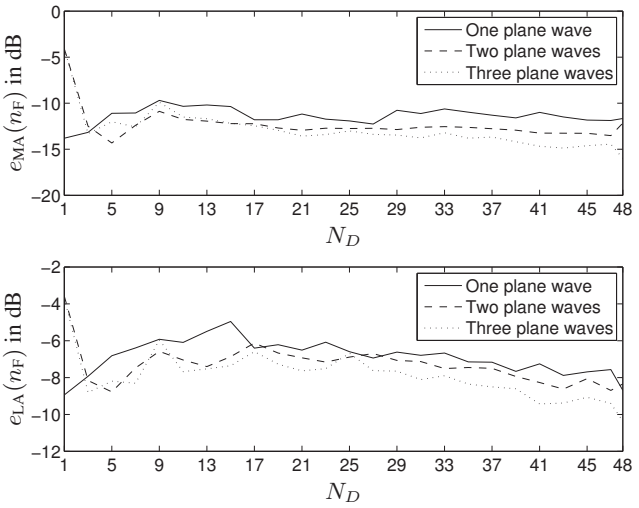

Fig. 6. LRE error after convergence for different equalizer structures. The index n_{F} equals n after 45 seconds.

$e_{\mathrm{MA}}(n) = -13$dB. The initial divergence is due to a poorly identified system $\mathbf{H}$ in the beginning In practical systems one would wait with determining $\tilde{\mathbf{G}}(n)$ until $\tilde{\mathbf{H}}(n)$ has been sufficiently well identified. A slightly better convergence for the examples with two or three plane waves can also be explained through a better identification of $\mathbf{H}$, as the loudspeaker signals are less correlated for an increased number of synthesized plane waves. It can be seen that the error in the listening area shows the same behavior as the error at the position of the microphone array, although the remaining error is about 5dB larger. This shows that for the chosen array setup a solution for the circumference of the microphone array may be interpolated towards the center of the microphone array, i. e., the listening area.

Fig. 4 shows an example for an impulse-like plane wave with an incidence angle of $\varphi_1 = 0$ for the converged equalizers. It can be seen that the equalizers preserve the wave shape (upper left plot) and compensate for reflections within the listening area (lower left plot), while the wave field outside the listening area is somewhat distorted. This is not surprising as the wave field outside the listening area is not enclosed by the microphone array and is therefore not optimized. This effect is stronger for larger values of N_D, suggesting to apply additional constraints on the equalizer coefficients to suppress it.

In Fig. 6 we can see the errors $e_{\mathrm{MA}}(n_{\mathrm{F}})$ and $e_{\mathrm{LA}}(n_{\mathrm{F}})$ measured after convergence for structures with different N_D, where n_{F} is equal to n after 45 seconds. For the scenario with one synthesized plane wave denoted by the solid line, we can see that actually the simplest structure with $N_D = 1$ shows the best performance. Although the other structures with $N_D > 1$ have more degrees of freedom, they cannot take advantage of it, because the underlying inverse filtering problem is ill-conditioned. On the other hand, for the more complex scenarios with two or three synthesized plane waves, denoted by the dashed and the dotted line, respectively, the structure with $N_D = 1$ does not have sufficient degrees of freedom and the more complex structures perform significantly better.

5. CONCLUSIONS

In this paper we presented a generalization of the originally proposed structure for an adaptive LRE in the wave domain, by also considering the relations between wave-field components of different orders. It has been shown that the necessary complexity and optimum performance of the LRE structure is dependent on the complexity of the reproduced scene. Moreover, the underlying inverse filtering problem is strongly ill-conditioned suggesting to chose the number of degrees of freedom as low as possible. Due to the scalable complexity, the proposed system exhibits lower computational demands and a higher robustness compared to conventional systems, while it is also suitable for a broader range of reproduction scenarios compared to the first proposal of such a system operating in the wave domain. Future work will include an analysis of the residual error over the temporal frequency and explore options for extending LRE to the outside of the microphone array.

6. REFERENCES

[1] A.J. Berkhout, D. De Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, pp. 2764 – 2778, May 1993.

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

[3] P.A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Inverse filter design and equalization zones in multichannel sound reproduction," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 3, pp. 185 – 192, May 1995.

[4] 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 *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2004, vol. 4, pp. IV–29 – IV–32.

[5] S. Goetze, M. Kallinger, A. Mertins, and K.D. Kammeyer, "Multi-channel listening-room compensation using a decoupled filtered-X LMS algorithm," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Oct. 2008, pp. 811 – 815.

[6] M. Schneider and W. Kellermann, "A wave-domain model for acoustic MIMO systems with reduced complexity," in *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Edinburgh, UK, May 2011.

[7] T. Betlehem and T.D. Abhayapala, "Theory and design of sound field reproduction in reverberant rooms," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2100 – 2111, April 2005.

[8] H. Buchner, S. Spors, and W. Kellermann, "Wave-domain adaptive filtering: acoustic echo cancellation for full-duplex systems based on wave-field synthesis," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2004, vol. 4, pp. IV–117 – IV–120.