An Iterative Least-Squares Design Method for Filters with Constrained Magnitude Response in Sound Reproduction

Martin Schneider[1,2] and Emanuël A. P. Habets[1,2]
[1] *Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany*
[2] *International Audio Laboratories Erlangen*

Introduction

Linear time-invariant filters determined according to a least-squares criterion are frequently used in applications related to sound zones, sound-field reproduction, and adjustable directivity [1, 2, 3, 4]. Without further constraints, the filter's magnitude response can exhibit very large values whenever the underlying optimization problem is ill-conditioned. This can lead to distortion in the loudspeakers, which is unwanted for audio reproduction and implies a violation of the typically assumed linear system model. Determining magnitude-constrained least-squares optimal filters is a straightforward solution to this problem, but implies to solve a quadratically constrained quadratic program. This is an algorithmically non-trivial and computationally expensive task.

Although time-domain finite impulse response (FIR) filters are desired, the filters are often designed according to a discrete Fourier transform (DFT)-domain criterion, which reduces the computational complexity [5]. This implies that the obtained filter coefficients are not necessarily optimal according to the original time domain criterion. However, time-domain approaches imply rather large computational demands, which renders an application of those approaches unlikely when large numbers of loudspeakers or long filter lengths have to be considered.

In this contribution, a previously presented efficient iterative least-squares filter design algorithm [6] is modified such that it yields a filter with a constrained magnitude response. The algorithm is designed to maximize computational efficiency and can be straightforwardly implemented.

Problem Formulation

In the following, the problem of designing filters for sound reproduction is considered, independently of a specific application scenario.

The signal model is shown in Fig. 1, where the single-channel time-discrete source signal is denoted by $q(k)$ with k being the time index. This signal is fed to the reproduction filters such that N_L loudspeaker signals denoted by $x_l(k)$ are obtained by linear convolution, where l is the loudspeaker index. The reproduction filters are described by the length-L_G FIRs $g_l(k)$. The considered acoutic channel is described by the length-L_H FIRs $h_{m,l}(k)$ and yields N_M output signals described by $y_m(k)$, where m is the signal index. The desired output signals $z_m(k)$ are obtained by filtering $q(k)$ using the FIRs $d_m(k)$

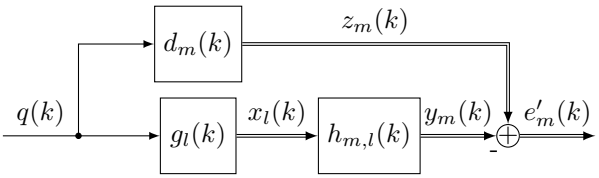

Figure 1: Signal model for filter design

of length L_C. The optimization of $g_l(k)$ is performed such the error signals $e'_m(k) = z_m(k) - y_m(k)$ are minimized according to a least-mean-squares criterion.

When ignoring the spectral characteristics of $q(k)$, the optimization of $g_l(k)$ and all following considerations can be restricted to the FIRs $g_l(k)$, $h_{m,l}(k)$, and $d_m(k)$. Then, the error to be minimized is given by

$$e_m(k) = d_m(k) - \sum_{l=1}^{N_\mathrm{L}} \sum_{\kappa=0}^{L_\mathrm{H}-1} g_l(k-\kappa) h_{m,l}(\kappa), \qquad (1)$$

which represents an impulse response instead of a signal. Since all considered impulse responses are of finite length, (1) can be represented using matrices and vectors:

$$\mathbf{e} = \mathbf{d} - \mathbf{H}\mathbf{g}, \qquad (2)$$

$$\mathbf{e} = (e_1(0), \dots e_1(L_\mathrm{C}-1), e_2(0), \dots e_{N_\mathrm{M}}(L_\mathrm{C}-1))^T,$$
$$\mathbf{d} = (d_1(0), \dots d_1(L_\mathrm{C}-1), d_2(0), \dots d_{N_\mathrm{M}}(L_\mathrm{C}-1))^T,$$
$$\mathbf{g} = (g_1(0), \dots g_1(L_\mathrm{G}-1), g_2(0), \dots g_{N_\mathrm{L}}(L_\mathrm{G}-1))^T,$$

with $(\cdot)^T$ being the transposition, $L_\mathrm{C} = L_\mathrm{G} + L_\mathrm{H} - 1$, and $\mathbf{H}$ describing a convolution matrix containing the impulse responses $h_{m,l}(k)$ such that (1) and (2) are equivalent. Using the introduced notation, the unconstrained optimization criterion is given by

$$\mathbf{g}_\mathrm{opt} = \arg\min_{\mathbf{g}} \left\{ \mathbf{e}^T \mathbf{e} \right\}. \qquad (3)$$

The constrained variant of (3) is given by

$$\mathbf{g}_\mathrm{opt} = \arg\min_{\mathbf{g}} \left\{ \mathbf{e}^T \mathbf{e} \right\} \text{ subject to } \|\mathbf{W}\mathbf{g}\|_\infty \leq c, \qquad (4)$$

where $\|\cdot\|_\infty$ is the maximum norm and c the chosen maximum DFT-domain amplitude of $\mathbf{g}$, which is given by $\mathbf{W}\mathbf{g}(n)$, where $\mathbf{W}$ is defined as

$$\mathbf{W} = \mathbf{I}_{N_\mathrm{L}} \otimes \left(\mathbf{F} \left(\mathbf{I}_{L_\mathrm{G}}, \mathbf{0} \right)^T \right), \qquad (5)$$

with $\mathbf{I}_N$ being the $N \times N$ identity matrix, $\mathbf{0}$ being an $L_\mathrm{G} \times (L_\mathrm{H} - 1)$ matrix with zero-valued entries and $\otimes$

denoting the Kronecker product. The DFT matrix $\mathbf{F}$ is defined by

$$[\mathbf{F}]_{p,q} = \frac{1}{\sqrt{L_\mathrm{C}}} e^{-j(q-1)(p-1)\frac{2\pi}{L_\mathrm{C}}}, \tag{6}$$

where $[\mathbf{F}]_{p,q}$ addresses the entry located at row p and column q in $\mathbf{F}$ and j is the imaginary unit ($j^2 = -1$).

It should be noted that (4) is a convex optimization problem and whenever a solution to (3) violates $\|\mathbf{W}\mathbf{g}\|_\infty \leq c$, the global optimum of (4) lies at one or more boundaries defined by its constraint.

Filter Design

In this section, the actual computation of optimal filters is described. The solution to (3) is given by

$$\mathbf{g}_\mathrm{opt} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{d}, \tag{7}$$

assuming $\mathbf{H}^T\mathbf{H}$ to be invertible, as in the following. If an optimal solution cannot be computed in closed form, an iterative method can be used to find an approximative solution. In such a scenario, the filter coefficient vector $\mathbf{g}(n)$ depends on the iteration index n and

$$\mathbf{e}'(n) = \mathbf{d} - \mathbf{H}\mathbf{g}(n-1), \tag{8}$$
$$\mathbf{e}(n) = \mathbf{d} - \mathbf{H}\mathbf{g}(n), \tag{9}$$

represent the *a priori* and *a posteriori* errors, respectively. Note that $\mathbf{e}'(n)$ is explicitly given, while $\mathbf{g}(n)$ is determined to reduce $\mathbf{e}(n)$ in each iteration. In (7), replacing $\mathbf{g}_\mathrm{opt}$ by $\mathbf{g}(n)$ and substituting $\mathbf{d}$ using (8) leads to the update equation

$$\mathbf{g}(n) = \mathbf{g}(n-1) + \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{e}'(n). \tag{10}$$

If $\left(\mathbf{H}^T\mathbf{H}\right)^{-1}$ can be computed exactly, $\mathbf{g}(n)$ is identical to $\mathbf{g}$ as given by (7) after each iteration. A previously presented algorithm avoids computing $\left(\mathbf{H}^T\mathbf{H}\right)^{-1}$ using a computationally efficient approximation in the DFT domain [6]. The derivation in [6] is based on the overlap-save fast convolution described by

$$\mathbf{H} = \mathbf{F}_M^H\,\underline{\mathbf{H}}\,\mathbf{W}, \tag{11}$$
$$\mathbf{F}_M = \mathbf{I}_{N_\mathrm{M}} \otimes \mathbf{F} \tag{12}$$

where $(\cdot)^H$ denotes the conjugate transpose and $\underline{\mathbf{H}}$ is a sparse matrix representing $\mathbf{H}$ in the DFT-domain.

The DFT-domain representation of (8) is given by

$$\underline{\mathbf{e}}'(n) = \mathbf{F}_M\mathbf{e}'(n) = \underline{\mathbf{d}} - \underline{\mathbf{H}}\underline{\mathbf{g}}(n-1), \tag{13}$$

using $\underline{\mathbf{g}}(n) = \mathbf{W}\mathbf{g}(n)$, which allows for the approximation

$$\underline{\mathbf{g}}(n) = \underline{\mathbf{g}}(n-1) + \mathbf{W}\mathbf{W}^H\left(\underline{\mathbf{H}}^H\underline{\mathbf{H}} + \delta\mathbf{I}_{L_\mathrm{C}N_\mathrm{L}}\right)^{-1}\underline{\mathbf{H}}^H\underline{\mathbf{e}}'(n), \tag{14}$$

where $\underline{\mathbf{H}}^H\underline{\mathbf{H}}$ constitutes a sparse matrix, which can be efficiently inverted [6], and δ is a regularization parameter.

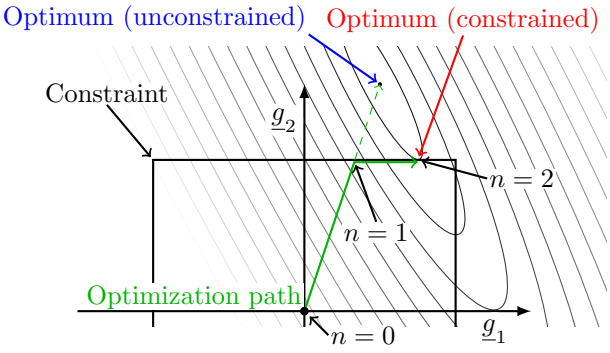

Figure 2: Illustration how an optimum solution is approached

This parameter was identified to be crucial for the convergence behavior of the algorithm. However, a deeper investigation of this effect exceeds the scope of this paper.

The proposed algorithm adjusts the unconstrained update term in (14) such that all constraints are satisfied. In further iterations, the DFT-domain coefficients touching a boundary are regarded as being fixed, while the optimization problem is solved only considering the remaining degrees of freedom. This procedure is illustrated in Fig. 2 considering the real part of two DFT-domain coefficients of $\underline{\mathbf{g}}(n)$ and two update steps. The algorithm can be derived using a redefinition of the *a priori* error according to:

$$\mathbf{e}(n) = \mathbf{d} - \mathbf{F}_M^H\underline{\mathbf{H}}\left(\mathbf{V}^2(n)\underline{\mathbf{g}}(n) - \left(\mathbf{I} - \mathbf{V}^2(n)\right)\underline{\mathbf{g}}(n-1)\right), \tag{15}$$

where $\mathbf{V}^2(n) = \mathbf{V}^H(n)\mathbf{V}(n)$ and

$$v_\eta(n) = \begin{cases} 1 & \text{if } \left|\left[\underline{\mathbf{g}}(n-1)\right]_\eta\right| < c - \epsilon, \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

$$\left[\mathbf{V}(n)\right]_{p,\eta} = \begin{cases} 1 & \text{if } v_\eta(n) = 1 \cap p = \sum_{q=1}^{\eta} v_q(n), \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

with $\mathbf{V}(n)$ being $L_\mathrm{C}N_\mathrm{L} \times \sum_{\eta=1}^{L_\mathrm{C}N_\mathrm{L}} v_\eta(n)$. Depending on $\mathbf{V}(n)$, individual coefficients of $\underline{\mathbf{g}}(n-1)$ will not be altered in the update. Equation (15) can be minimized in a least-squares sense following a derivation similar to that presented in [7], where

$$\left(\mathbf{I} - \mathbf{V}^2\right)\underline{\mathbf{g}}(n-1) = \left(\mathbf{I} - \mathbf{V}^2\right)\underline{\mathbf{g}}(n) \tag{18}$$

must hold to avoid ambiguity. This results in the update rule

$$\underline{\mathbf{g}}'(n) = \underline{\mathbf{g}}(n-1) + \mathbf{W}\mathbf{W}^H\mathbf{M}(n)\underline{\mathbf{u}}(n), \tag{19}$$

$$\underline{\mathbf{u}}(n) = \mathbf{V}^H(n)\left(\mathbf{V}(n)\left(\underline{\mathbf{H}}^H\underline{\mathbf{H}} + \delta\mathbf{I}_{L_\mathrm{C}N_\mathrm{L}}\right)\mathbf{V}^H(n)\right)^{-1} \times \mathbf{V}(n)\underline{\mathbf{H}}^H\underline{\mathbf{e}}'(n). \tag{20}$$

Here, $\underline{\mathbf{g}}'(n)$ replaces $\underline{\mathbf{g}}(n)$ because of reasons apparent later. The diagonal matrix $\mathbf{M}(n)$ is chosen such that each coefficient of $\underline{\mathbf{g}}'(n)$ meets the constraints:

$$p_\eta(n) = \mathfrak{Re}\left(\left[\underline{\mathbf{g}}^*(n-1)\right]_\eta\left[\underline{\mathbf{u}}(n)\right]_\eta\right) / \left|\left[\underline{\mathbf{u}}(n)\right]_\eta\right|^2, \tag{21}$$

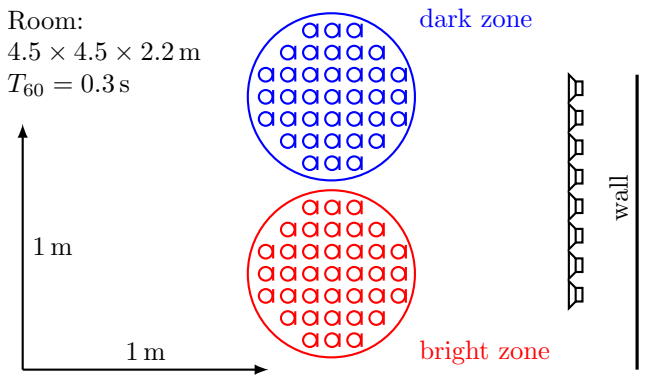

Figure 3: Loudspeaker-enclosure-microphone system used for evaluation (to scale)

$$q_\eta(n) = \left(\left| \left[\underline{\mathbf{g}}(n-1) \right]_\eta \right|^2 - c^2 \right) / \left| \left[\underline{\mathbf{u}}(n) \right]_\eta \right|^2, \tag{22}$$

$$m'_\eta(n) = \left(-1 - p_\eta(n) + \sqrt{|p_\eta(n)|^2 - q_\eta(n)} \right) v_\eta(n) + 1, \tag{23}$$

$$m_\eta(n) = \begin{cases} 1 & \text{if } m'_\eta(n) > 1, \\ 0 & \text{if } m'_\eta(n) < 0, \\ m'_\eta(n) & \text{otherwise,} \end{cases} \tag{24}$$

$$\left[\underline{\mathbf{m}}(n) \right]_\nu = \min_l \left\{ m_{\nu + L_\mathrm{C}(l-1)} \right\}, \tag{25}$$

$$\mathbf{M}(n) = \mathrm{Diag}\left(\mathbf{1} \otimes \underline{\mathbf{m}}(n) \right), \tag{26}$$

where $\mathfrak{Re}\{\cdot\}$ is the real part of a number and $\mathbf{1}$ is a N_L-element column vector of ones. Note that (10), (14), and (19) exploit absolutely no assumptions on $\mathbf{g}(n-1)$ or $\underline{\mathbf{g}}(n-1)$, which implies total freedom in obtaining its value. This is important for multiple reasons: First, update steps can be scaled arbitrarily and independently for each DFT bin using $\mathbf{M}(n)$. Second, excluding particular DFT bins from updates using $\mathbf{V}(n)$, as well as computing the updates independently for each DFT bin, are approximations. This precludes exploiting any properties of $\mathbf{g}(n-1)$ that are only guaranteed if exact computations are possible. Third, the filter's transfer function may be modified to reduce the influence of the well-known Gibbs phenomenon resulting from windowing.

The Gibbs phenomenon can be mitigated by scaling the coefficients of $\underline{\mathbf{g}}'(n)$, where all coefficients corresponding to one DFT bin are scaled by the same factor:

$$\left[\underline{\mathbf{c}}(n) \right]_\nu = \begin{cases} 1 \text{ if } \left[\underline{\mathbf{g}}(n) \right]_{\nu + L_\mathrm{C}(l-1)} \leq c \;\forall\; l, \\ c / \max_l \left\{ \left[\underline{\mathbf{g}}(n) \right]_{\nu + L_\mathrm{C}(l-1)} \right\} \text{ otherwise,} \end{cases} \tag{27}$$

$$\underline{\mathbf{g}}(n) = \mathrm{Diag}\left(\mathbf{1} \otimes \underline{\mathbf{c}}(n) \right) \underline{\mathbf{g}}'(n) \tag{28}$$

where ν indexes the DFT bin.

The proposed algorithm can be summarized as follows:

1. Initialize $\underline{\mathbf{g}}(0)$ with zeros and $\mathbf{V}(0)$ with an identity matrix of appropriate dimensions.
2. Compute the error $\underline{\mathbf{e}}'(n)$ using (13).
3. Compute $\underline{\mathbf{u}}(n)$ using (20).
4. Determine $\mathbf{M}(n)$ using (21) to (26).
5. Determine $\mathbf{V}(n+1)$ using (17).
6. Scale $\underline{\mathbf{g}}'(n)$ using (28) to reduce consequences of Gibbs phenomenon.
7. Repeat from step 2 until error improvement is below a chosen threshold.

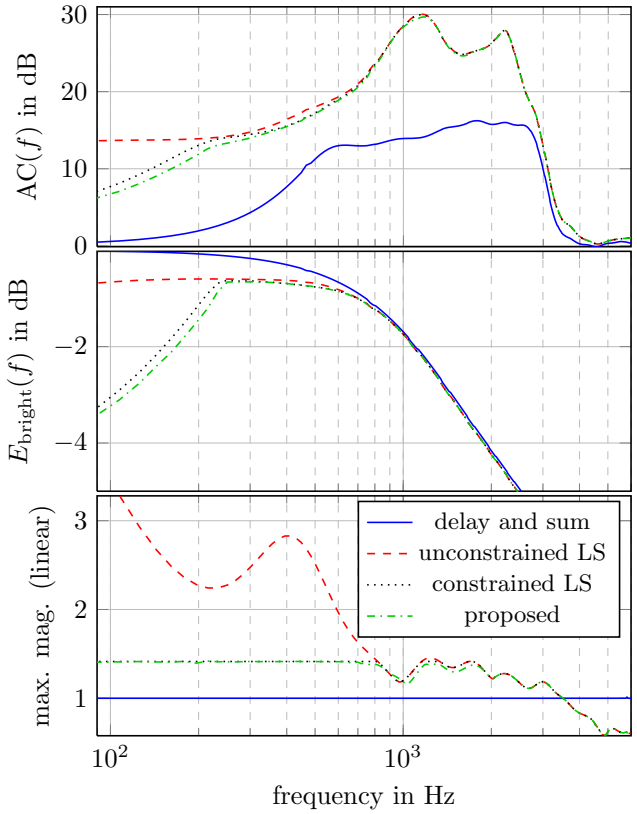

Figure 4: Evaluation results for the free-field case

Evaluation

In this section, evaluation results for the proposed approach are presented. The proposed algorithm was evaluated considering a sound-zones scenario as shown in Fig. 3 with $N_\mathrm{M} = 74$ and $N_\mathrm{L} = 8$. The desired impulse response $\mathbf{d}$ is chosen as if a sound field of a delay and sum beamformer radiating towards the center of the bright zone is reproduced in the bright zone, while no sound is radiated to the dark zone. This problem is ill-conditioned, even in the free-field case. Hence, the magnitude-squared sum of coefficients in $\mathbf{g}$ was penalized for regularization by adding virtual couplings in $\mathbf{H}$, where each loudspeaker was coupled to an individual microphone in the dark zone using a unit response scaled by $\frac{0.1}{N_\mathrm{L}} \cdot \sqrt{\sum_{k=0}^{L_\mathrm{H}} \sum_{l=1}^{N_\mathrm{L}} \sum_{m=1}^{N_\mathrm{M}} h_{m,l}^2(k)}$. The maximum DFT-domain amplitude c for the constrained approaches was chosen to be $\sqrt{2}$. A filter length L_G of 512 samples was chosen at a sampling frequency f_s of 12 kHz.

Two measures relevant for evaluation are the cascade impulse response energies to the bright and the dark zone

at a given frequency f:

$$E_{\text{bright}}(f) = \underline{\mathbf{g}}^H(n)\underline{\mathbf{H}}^H\mathbf{B}_{\frac{fL_C}{f_s}}\underline{\mathbf{H}}\underline{\mathbf{g}}(n), \tag{29}$$

$$E_{\text{dark}}(f) = \underline{\mathbf{g}}^H(n)\underline{\mathbf{H}}^H\mathbf{D}_{\frac{fL_C}{f_s}}\underline{\mathbf{H}}\underline{\mathbf{g}}(n), \tag{30}$$

where the diagonal matrices $\mathbf{B}_\nu$ and $\mathbf{D}_\nu$ exhibit one-valued entries only at the positions of DFT bin ν and for the microphones in bright and dark zone, respectively. These values can be used to determine the acoustic contrast given by

$$\text{AC}(f) = \frac{E_{\text{bright}}(f)}{E_{\text{dark}}(f)}. \tag{31}$$

In the following, the acoustic contrast, the coupling energy to the bright zone (equivalent to sound pressure in real-world applications), and the maximum magnitude of the DFT bins are considered to compare four approaches realizing sound zones: a delay-and-sum beamformer, unconstrained least-squares filters as given by (7), constrained least-squares filters obtained by solving (4) with the well-known CVX toolbox and the proposed algorithm. The iterations were stopped when $\mathbf{e}^T(n)\mathbf{e}(n) > \mathbf{e}^T(n-1)\mathbf{e}(n-1)(1-10^{-6})$ and δ was chosen to be $\frac{0.1}{N_L L_C} \cdot \sqrt{\sum_{k=0}^{L_H}\sum_{l=1}^{N_L}\sum_{m=1}^{N_M} h_{m,l}^2(k)}$.

For the results shown in Fig. 4, $\mathbf{H}$ was computed as it would occur in the ideal free-field case, which implied $L_H = 159$. It can be seen that the delay-and-sum beamformer achieves a limited acoustic contrast in a limited frequency range, while all least-squares approaches achieve a higher contrast in a broader frequency range. Eventhough the problem was regularized, the unconstrained least-squares filters exhibit a large magnitude, which could be successfully limited by the constrained approaches. This limitation was paid by a lower contrast and a lower coupling energy to the bright zone for frequencies below 200 Hz. It can be seen that the proposed approach almost achieves the same performance as the exact solution by CVX, which is remarkable since the computation time for the proposed algorithm was less than one second, compared to 550 seconds for the CVX solution. The exact unconstrained least-squares filters took around two seconds to compute.

The results shown in Fig. 5 were obtained for the same scenario as described above, but with measured impulse responses. For the filter design, only 300 time samples of $h_{m,l}(k)$ were considered due to computational limitations. For the evaluation 2000 samples were considered. It can be seen that the achieved acoustic contrast is much lower than for the free-field case, which can be accounted to the more difficult scenario. Like in the free-field case, the proposed algorithm approaches the performance of the exactly calculated solution closely, while the computations took around one second, compared to around 10^4 seconds for the exact solution.

Conclusions

An algorithm to obtain filters with a constrained magnitude response was presented. Although this algorithm only obtains an approximate solution, it has been shown that the obtained filters are suitable for real-word applications. Future research can be focused on the regularization of this algorithm, which was identified as a crucial aspect and is not yet fully understood.

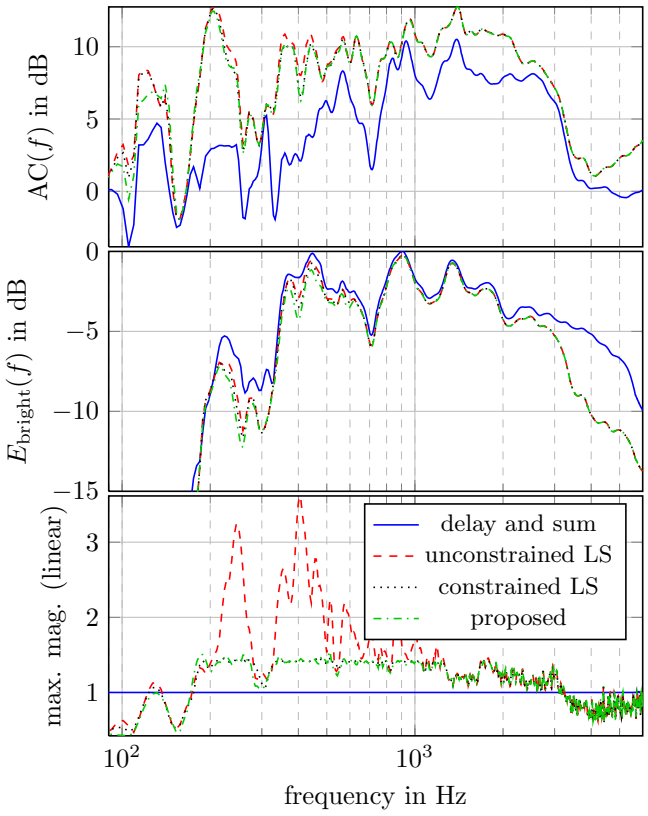

Figure 5: Evaluation results for measured impulse responses

References

[1] O. Kirkeby and P.A. Nelson, "Reproduction of plane wave sound fields," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2992, 1993.

[2] M. Poletti, "An investigation of 2-d multizone surround sound systems," in *Proceedings of the Convention of the Audio Engineering Society*, Oct. 2008.

[3] Y.J. Wu and T.D. Abhayapala, "Spatial multizone soundfield reproduction: Theory and design," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1711–1720, 2011.

[4] L. Bianchi, R. Magalotti, F. Antonacci, A. Sarti, and S. Tubaro, "Robust beamforming under uncertainties in the loudspeakers directivity pattern," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4448–4452.

[5] P.D. Teal, T. Betlehem, and M.A. Poletti, "An algorithm for power constrained holographic reproduction of sound," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 101–104.

[6] M. Schneider and W. Kellermann, "Iterative DFT-domain inverse filter determination for adaptive listening room equalization," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aachen, Germany, Sep. 2012, pp. 1 – 4.

[7] M. Schneider and W. Kellermann, "The generalized frequency-domain adaptive filtering algorithm as an approximation of the block recursive least-squares algorithm," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 6, 2016.