Acoustic Echo Cancellation for Ambisonics-based Spatial Audio Systems: A Wave-Domain Approach

Marcel Nophut, Stephan Preihs, and Jürgen Peissig
Leibniz Universität Hannover, Institut für Kommunikationstechnik, 30167 Hannover, Deutschland,
Email: marcel.nophut@ikt.uni-hannover.de

Abstract
The technique of wave-domain adaptive filtering (WDAF) is a powerful tool in massive-multichannel acoustic system identification. By choosing fundamental solutions of the wave-equation for signal representation, the adaptive filter no longer models the acoustic paths from loudspeakers to microphones, but from wave-component to wave-component. In suitable setups the modeled paths in the wave-domain exhibit desirable properties, which can be exploited to improve the system identification task. Recently, the authors have successfully applied the WDAF method to three-dimensional acoustical setups and have proposed suitable transforms based on the Ambisonics technique. In this contribution WDAF and the previously proposed transforms are applied to acoustic echo cancellation in Ambisonics-based spatial audio systems. The performance of techniques exploiting the above-mentioned properties is examined and compared with regard to echo reduction and misalignment in different scenarios.

Introduction
Modern telepresence systems incorporating spatial audio may comprise a large number of both loudspeakers and microphones. However, acoustic echo cancellation (AEC) in massive-multichannel systems severely suffers from non-uniqueness of the underlying identification problem [1]. Wave-domain adaptive filtering (WDAF) [2] models the acoustic paths of an loudspeaker-enclosure-microphone-system (LEMS) from wave component to wave component, rather than from loudspeaker to microphone. Different methods exploiting the characteristics of the wave-domain (WD) LEMS have been proposed to improve the AEC performance and even reduce the non-uniqueness problem [1, 3]. However, these and other previous WDAF studies (e.g. [2, 4]) have been limited to two-dimensional setups. After they had evaluated the characteristics of the WD-LEMS in the context of the well-known higher-order Ambisonics (HOA) method and a practical three-dimensional loudspeaker layout in a previous study [5], the authors now investigate the actual AEC performance of known WDAF methods in this scenario.

Evaluation of Energy Couplings in the Wave-Domain
For investigating the properties of the WD-LEMS in the considered scenarios the following experiment was conducted. The contents of this section were previously published by the authors [5] and are only briefly reviewed in the following. Two different loudspeaker layouts were considered: First, a layout of 20 uniformly distributed over a sphere (Uni20) and the non-uniform, practically motivated layout of 32 loudspeakers of the Immersive Media Lab at IKT (IML32), depicted in Fig. 1. For the IML32 layout all channels were equalized regarding broadband gain and delay with respect to the center of the microphone array. So both setups can be considered concentrical layouts. The spherical microphone array used was an Eigenmike em32 with 32 microphone capsules on a rigid sphere.

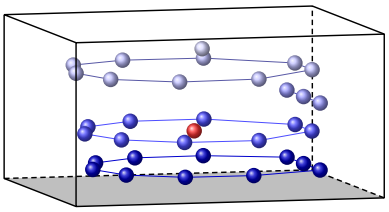

Figure 1: Loudspeaker layout IML32 (blue) and Eigenmike em32 (red).

Furthermore, two different acoustical environments were considered: simulated transfer functions from an anechoic environment (SimAnechoic) and measured transfer functions from the Immersive Media Lab (IML) at IKT (MeasIML). For the WD transform on the loudspeaker side ($\boldsymbol{\mathcal{T}}_1^{-1}$) two different decoding methods, the mode-matching Ambisonics decoder (MMD) [6] and the energy-preserving Ambisonics decoder (EPAD) [7] were used. The WD transform on the microphone side ($\boldsymbol{\mathcal{T}}_2$) was a microphone-array-to-Ambisonics transform matrix together with Tikhonov regularized radial filters [8]. The Ambisonics order was set to 3rd order, Ambisonics channel numbering ACN was used.

Figure 2 show the broadband energy $E_{i,j}$ (50 Hz to 9 kHz, in dB) of the transfer functions $H_{l,m}(j\omega)$ and $\underline{\underline{H}}_{\lambda,\mu}(j\omega)$ in the point-to-point-domain (PTP) and wave-domain (WD), respectively.

The overall result of the first experiment was that the EPAD method maintained a (WDAF-typical) dominant diagonal in the coupling matrix for all cases investigated and may be better suitable for exploitation by known WDAF methods.

AEC in the Wave-Domain
System Overview
The setup and transforms from the previous section were used for the actual AEC experiments in the following section without any changes. The general signal flow

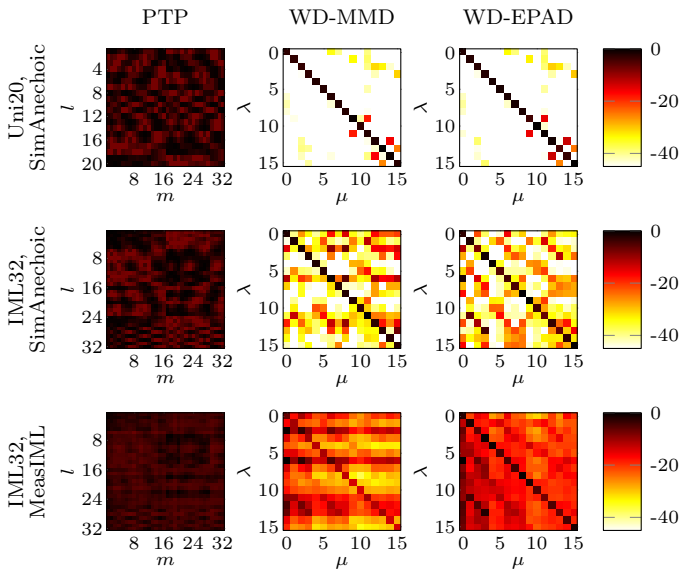

Figure 2: Broadband energy couplings $E_{i,j}$ (in dB) from loudspeaker channel l to microphone channel m (point-to-point, PTP) and between Ambisonics channels at input λ and output μ (wave-domain, WD) for three different cases.

chart for WD-AEC in Ambisonics-based setups is depicted in Fig. 3. The signal frames of the N_λ-channel WD input signal $\underset{\sim}{\mathbf{X}}(\tau)$, where τ is the frame-time instant, are received from the far-end/transmitting room as an Ambisonics-encoded signal. The signal frames of the N_m-channel frequency-domain microphone signal $\mathbf{Y}(\tau)$ do not only contain the echo signals $\mathbf{D}(\tau)$ (loudspeaker signals picked up by microphones in the near-end/receiving room), but also a white (microphone) noise signal $\mathbf{N}_{\mathrm{WGN}}(\tau)$ and a double-talk signal $\mathbf{N}_{\mathrm{DT}}(\tau)$ ($\mathbf{Y}(\tau) = \mathbf{D}(\tau) + \mathbf{N}_{\mathrm{WGN}}(\tau) + \mathbf{N}_{\mathrm{DT}}(\tau)$).

The depicted quantities are defined as follows: Matrix $\underset{\sim}{\mathbf{X}}(\tau) = \left[\underset{\sim}{\mathbf{X}}_0(\tau) \cdots \underset{\sim}{\mathbf{X}}_\lambda(\tau) \cdots \underset{\sim}{\mathbf{X}}_{N_\lambda-1}(\tau)\right]$ consists of N_λ submatrices with each of them containing the DFT representation $\underset{\sim}{\mathbf{X}}_\lambda(\tau) = \mathrm{diag}\left\{\mathbf{F}_M \underset{\sim}{\mathbf{x}}_\lambda(\tau)\right\}$ of the λ-th channel WD input signal time-frame of length M on its diagonal, where $\mathbf{F}_M$ is the M-point DFT-matrix. Matrix $\underset{\sim}{\mathbf{Y}}(\tau) = \left[\underset{\sim}{\mathbf{y}}'_0(\tau) \cdots \underset{\sim}{\mathbf{y}}'_\mu(\tau) \cdots \underset{\sim}{\mathbf{y}}'_{N_\mu-1}(\tau)\right]$ consists of N_μ column vectors with each of them containing the DFT representation $\underset{\sim}{\mathbf{y}}'_\mu(\tau) = \mathbf{F}_R \underset{\sim}{\mathbf{y}}_\mu(\tau)$ of the μ-th channel WD output signal time-frame of length R. Signal matrices $\hat{\underset{\sim}{\mathbf{D}}}(\tau)$, $\underset{\sim}{\mathbf{E}}(\tau)$ and $\mathbf{Y}(\tau)$ are defined accordingly. Matrix $\hat{\underset{\sim}{\mathbf{H}}}(\tau) = \left[\hat{\underset{\sim}{\mathbf{h}}}'_1(\tau) \cdots \hat{\underset{\sim}{\mathbf{h}}}'_\mu(\tau) \cdots \hat{\underset{\sim}{\mathbf{h}}}'_{N_\mu-1}(\tau)\right]$ consists of N_μ column vectors with each of them containing the estimated WD transfer functions for the N_λ input channels contributing to output channel μ. Vector $\hat{\underset{\sim}{\mathbf{h}}}'_\mu(\tau) = \left[\hat{\underset{\sim}{\mathbf{h}}}'^T_{0,\mu}(\tau) \cdots \hat{\underset{\sim}{\mathbf{h}}}'^T_{\lambda,\mu}(\tau) \cdots \hat{\underset{\sim}{\mathbf{h}}}'^T_{N_\lambda-1,\mu}(\tau)\right]^T$ contains the vertically concatenated transfer functions $\hat{\underset{\sim}{\mathbf{h}}}'_{\lambda,\mu}(\tau) = \mathbf{F}_L \hat{\underset{\sim}{\mathbf{h}}}_{\lambda,\mu}(\tau)$ of the estimated WD impulse response $\hat{\underset{\sim}{\mathbf{h}}}_{\lambda,\mu}(\tau)$ of length L.

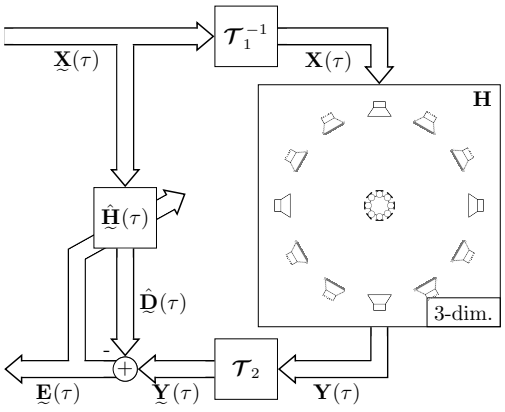

Figure 3: General signal flow chart for WDAF in Ambisonics-based spatial audio systems.

As it is common for MIMO adaptive filtering tasks, the $N_\lambda \times N_\mu$ MIMO problem is decomposed to N_μ separate $N_\lambda \times 1$ MISO problems [9]. The following formulations for deriving the echo cancelling adaptive filters are based on overlap-save convolution in the DFT-domain, where R is the signal frame size, L is the adaptive filter length and M is the length of a full overlap-save DFT-frame with $M = L + R$. Signal vectors and system responses (in any domain) of length R and L, respectively, are marked with a prime symbol $(\cdot)'$. Otherwise length M is assumed.

The μ-th channel WD adaptive filter output is computed as

$$\hat{\underset{\sim}{\mathbf{d}}}'_\mu(\tau) = \mathbf{W}_{01} \underset{\sim}{\mathbf{X}}(\tau) \mathbf{W}_{10} \hat{\underset{\sim}{\mathbf{h}}}'_\mu(\tau) \tag{1}$$

with the overlap-save constraint matrices

$$\mathbf{W}_{01} = \mathbf{F}_R \left[\mathbf{0}_{R\times M-R} \;\; \mathbf{I}_R\right] \mathbf{F}_M^{-1}, \tag{2}$$

$$\mathbf{W}_{10} = \mathbf{I}_{N_\lambda} \otimes \left[\mathbf{F}_M \left[\mathbf{I}_L \;\; \mathbf{0}_{L\times M-L}\right]^T \mathbf{F}_L^{-1}\right]. \tag{3}$$

Methods

Three AEC algorithms were used in the following experiments: the generalized frequency-domain adaptive filter (GFDAF) in its standard approximated form [9], the GFDAF estimating approximated models of the WD-LEMS (GFDAFapprxMdl) [3] and a constrained version of the GFDAF considering typical characteristics of the WD-LEMS (GFDAFconstr) [1].

The GFDAF can be considered a frame-based implementation in the frequency-domain of the well-known recursive least-squares (RLS) algorithm. The GFDAF's underlying cost function for WD output signal channel $\mu = 0, \ldots, N_\mu - 1$ exhibits a very similar form to the one of the RLS:

$$J_\mu(\tau) = (1 - \lambda_f) \sum_{i=0}^{\tau} \lambda_f^{\tau-i} \underset{\sim}{\mathbf{e}}'^H_{\tau,\mu}(i) \underset{\sim}{\mathbf{e}}'_{\tau,\mu}(i) \tag{4}$$

where λ_f is a forgetting factor ($0 \ll \lambda_f < 1$) and the error signal is obtained as

$$\underset{\sim}{\mathbf{e}}'_{\tau,\mu}(i) = \underset{\sim}{\mathbf{y}}'_\mu(i) - \mathbf{W}_{01} \underset{\sim}{\mathbf{X}}(i) \mathbf{W}_{10} \hat{\underset{\sim}{\mathbf{h}}}'_\mu(\tau). \tag{5}$$

The GFDAFapprxMdl method uses the standard GFDAF algorithm, but models only a subset of all

transfer-functions contributing to each WD output channel μ:

$$\hat{\underset{\sim}{\mathbf{h}}}'_{\text{apprx},\mu}(\tau) = \left[\hat{\underset{\sim}{\mathbf{h}}}'^{T}_{\lambda_1,\mu}(\tau)\ \hat{\underset{\sim}{\mathbf{h}}}'^{T}_{\lambda_2,\mu}(\tau)\ \cdots\ \hat{\underset{\sim}{\mathbf{h}}}'^{T}_{\lambda_i,\mu}(\tau)\right]^T. \quad (6)$$

Thus, the input signal matrix (now individual for each output channel μ) changes accordingly to

$$\underset{\sim}{\mathbf{X}}_{\text{apprx},\mu}(\tau) = \left[\underset{\sim}{\mathbf{X}}_{\lambda_1,\mu}(\tau)\ \underset{\sim}{\mathbf{X}}_{\lambda_2,\mu}(\tau)\ \cdots\ \underset{\sim}{\mathbf{X}}_{\lambda_i,\mu}(\tau)\right]. \quad (7)$$

The GFDAFconstr method exploits the knowledge of a characteristic energy coupling pattern, as the dominant diagonal coupling pattern, by introducing a constraint matrix $\mathbf{C}_\mu(\tau)$ to the cost function of the GFDAF:

$$J_\mu^{constr}(\tau) = (1-\lambda_f)\sum_{i=0}^{\tau}\lambda_f^{\tau-i}\underset{\sim}{\mathbf{e}}'^{H}_{\tau,\mu}(i)\underset{\sim}{\mathbf{e}}'_{\tau,\mu}(i) + \hat{\underset{\sim}{\mathbf{h}}}'^{H}_{\mu}(\tau)\mathbf{W}_{10}^H\mathbf{C}_\mu(\tau)\mathbf{W}_{10}\hat{\underset{\sim}{\mathbf{h}}}'_{\mu}(\tau). \quad (8)$$

This constraint matrix is defined as

$$\mathbf{C}_\mu(\tau) = \beta_0\, w_C(\tau) \times \text{diag}\left\{c_{1,\mu}, c_{2,\mu}, \ldots, c_{N_\lambda,\mu}\right\} \otimes \mathbf{I}_M \quad (9)$$

with the constant coefficient β_0 and a weighting function $w_C(\tau)$ (cf. [1]) ensuring good performance during both convergence and steady-state phases. The channel-dependent coefficients $c_{1,\mu}$ can be considered penalty quantities and thus have to be chosen inversely proportional to the expected weight of the transfer paths $\underset{\sim}{\mathbf{h}}_{\lambda,\mu}(\tau)$.

Two common performance metrics were used in the experiments: the echo-return-loss-enhancement (ERLE) as a metric for echo reduction in the microphone signal and the normalized misalignment (NMA) as a metric for the correct estimation of the WD-LEMS. The ERLE is defined as

$$ERLE(\tau) = 10\log_{10}\left(\frac{||\underset{\sim}{\mathbf{D}}(\tau)||_F^2}{||\underset{\sim}{\mathbf{D}}(\tau) - \hat{\underset{\sim}{\mathbf{D}}}(\tau)||_F^2}\right) \quad (10)$$

where $||\cdot||_F$ is the Frobenius norm and $\underset{\sim}{\mathbf{D}}(\tau) = \underset{\sim}{\mathbf{Y}}(\tau) - \underset{\sim}{\mathbf{N}}_{\text{WGN}}(\tau) - \underset{\sim}{\mathbf{N}}_{\text{DT}}(\tau)$. The NMA is defined as

$$NMA(\tau) = 10\log_{10}\left(\frac{||\underset{\sim}{\mathbf{H}}(\tau) - \hat{\underset{\sim}{\mathbf{H}}}(\tau)||_F^2}{||\underset{\sim}{\mathbf{H}}(\tau)||_F^2}\right). \quad (11)$$

Experiments

Following the results from the previous section, the EPAD decoding method was chosen for the experiments in combination with the IML32 loudspeaker layout. Since the objective of this work was to compare different algorithms and methods for WDAF, rather than comparing WDAF against traditional PTP adaptive filtering, only WD-LEMS are considered in the following.

The input signal $\underset{\sim}{\mathbf{X}}(\tau)$ was synthesized from measured room impulse responses of the Eigenmike em32 with white noise coming from two different incidence directions dir_i. From $t = 0\,\text{s}$ the sound was coming from 15° azimuth, 15° elevation (dir_1), from $t = 5\,\text{s}$ the sound was coming from −70° azimuth, 0° elevation (dir_2) and from $t = 15\,\text{s}$ the sound was coming from dir_1 and dir_2. As the mic-array-to-Ambisonics transform the same transform from the previous section ($\boldsymbol{\mathcal{T}}_2$) was used. As mentioned above only the EPAD method was used as transform $\boldsymbol{\mathcal{T}}_1^{-1}$. Moreover, the IML32 layout was considered in both the SimAnechoic and the MeasIML environment. White gaussian noise signals ($\mathbf{N}_{\text{WGN}}(\tau)$) were added to the microphone signals with $SNR = 60\,\text{dB}$ and a noise burst of 50 ms was present at $t = 10\,\text{s}$ with a $SIR = 0\,\text{dB}$ as a double-talk signal ($\mathbf{N}_{\text{DT}}(\tau)$).

The GFDAF was running at a sample rate of 16 kHz with frame size R and filter length L of identical size ($R = L = \frac{M}{2} = 1024$) with an overlap factor of $\alpha = 4$. The impulse responses of the WD-LEMS were truncated to $L_{LEMS} = 1024$. Other algorithm parameters were $\lambda_f = 0.95$ and $S_{init} = 0.01$. For the GFDAFapprxMdl two variants of approximations were used: First, a variant where only the main diagonal is modeled by the echo cancelling adaptive filter (diagOnly), which renders the N_μ MISO system identification problems N_μ SISO system identification problems (but without changing the MISO characteristic of the actual LEMS). Second, a variant where all coupling products down to −20 dB (top20dB) and −10 dB (top10dB) below the maximum coupling product are modeled. For the GFDAFconstr parameter β_0 was chosen $\beta_0 = 5 \times 10^{-4}$ and coefficents $c_{\lambda,\mu}$ were chosen as the inverse of the broadband magnitude of $\underset{\sim}{\mathbf{h}}_{\lambda,\mu}$ relative to the broadband magnitude of $\underset{\sim}{\mathbf{h}}_{0,0}$ (oracleMag).

Figure 4 shows the results of the SimAnechoic case. The disturbances at $t = 5\,\text{s}$, 10 s and 15 s cause the standard GFDAF to diverge and ERLE is more and more decreasing. A weak performance in terms of NMA with a comparatively high echo reduction means a strongly pronounced non-uniqueness for the GFDAF. The GFDAFapprxMdl variants show a rather robust convergence in terms of NMA. Its reduced echo reduction must be seen in the context of a (significantly) reduced computational complexity. The GFDAFconstr shows a very good echo reduction even after the disturbances, its convergence behavior in terms of NMA is very robust against the disturbances and it suffers much less from non-uniqueness than the standard GFDAF.

Figure 5 shows the results of the MeasIML case. Again the standard GFDAF diverges after the disturbances, while achieving a good, but decreasing echo reduction performance. The performance of the two GFDAFapprx variants regarding ERLE is further diminished and shows a hardly stable performance for the diagOnly variant and an instable behavior for the top10dB variant. Again, the GFDAFconstr shows a good echo reduction performance and a robust convergence behavior.

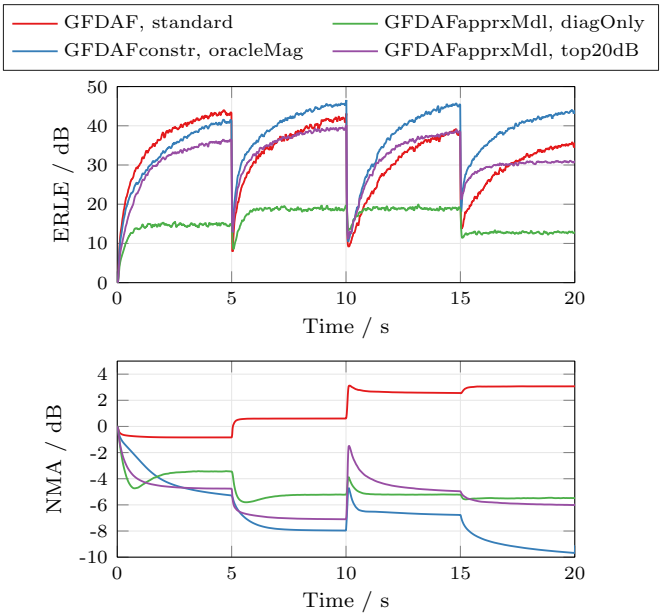

Figure 4: ERLE and NMA over time for environment SimAnechoic.

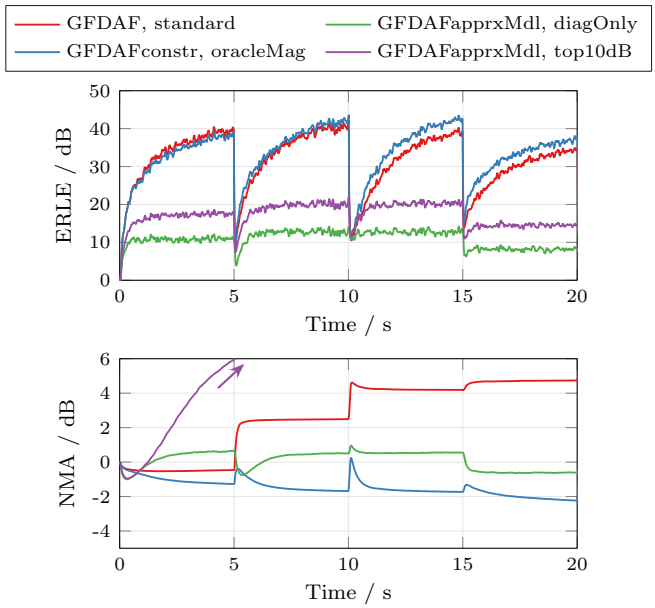

Figure 5: ERLE and NMA over time for environment MeasIML.

Conclusion

The performance of different WDAF methods applied to AEC in an Ambisonics-based spatial audio setup has been investigated. The study was focused on practical non-uniform loudspeaker layouts in anechoic and non-anechoic environments. The modelling of an approximated WD-LEMS (GFDAFapprxMdl) lead to a rather stable convergence behavior in the anechoic environment. For stronger approximations the echo reduction performance was lower compared to the standard GFDAF, but with higher savings regarding computational complexity. However, the GFDAFapprxMdl technique seems to be not suitable in non-anechoic environments, where the WD-LEMS coupling matrix is not sparse: Echo reduction performance is rather low and convergence behavior is hardly stable or instable. Introducing suitable constraints to the GFDAF (GFDAFconstr) improved the performance regarding ERLE and NMA, even in the presence of disturbances as double-talk or far-end speaker changes, in both considered cases. At the cost of a higher computational complexity, the standard GFDAF was clearly outperformed by the GFDAFconstr in terms of echo reduction and robustness of the system identification.

References

[1] M. Schneider and W. Kellermann, "Multichannel acoustic echo cancellation in the wave domain with increased robustness to nonuniqueness," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 518–529, 2016.

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

[3] M. Schneider and W. Kellermann, "A wave-domain model for acoustic mimo systems with reduced complexity," in *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2011)*, (Edinburgh, United Kingdom), pp. 133–138, IEEE, 2011.

[4] S. Emura, Y. Hiwasaki, and H. Ohmuro, "Wave-domain echo-path model with aliasing for echo cancellation," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2013)*, pp. 1–4, IEEE, 2013.

[5] M. Nophut, R. Hupke, S. Preihs, and J. Peissig, "Towards wave-domain adaptive filtering for multichannel acoustic echo cancellation in higher-order ambisonics systems," in *29th European Signal Processing Conference (EUSIPCO 2021)*, (Dublin, Ireland), pp. 161–165, IEEE, 2021.

[6] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc*, vol. 53, no. 11, pp. 1004–1025, 2005.

[7] F. Zotter, H. Pomberger, and M. Noisternig, "Energy-preserving ambisonic decoding," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 37–47, 2012.

[8] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, vol. 19 of *Springer Topics in Signal Processing*. Cham: Springer International Publishing, 2019.

[9] H. Buchner, J. Benesty, and W. Kellermann, "Multichannel frequency-domain adaptive filtering with application to acoustic echo cancellation," in *Adaptive Signal Processing* (J. Benesty and Y. Huang, eds.), Signals and Communication Technology, pp. 95–129, Berlin and Heidelberg: Springer, 2003.