Multichannel equalisation for high-order spherical microphone arrays using beamformed channels

(Invited Paper)

Alastair H. Moore, Christine Evers and Patrick A. Naylor
Dept. of Electical and Electronic Engineering
Imperial College
London, UK
alastair.h.moore@imperial.ac.uk

***Abstract*—High-order spherical microphone arrays offer many practical benefits including relatively fine spatial resolution in all directions and rotation invariant processing using eigenbeams. Spatial filtering can reduce interference from noise and reverberation but in even moderately reverberant environments the beam pattern fails to suppress reverberation to a level adequate for typical applications. In this paper we investigate the feasibility of applying dereverberation by considering multiple beamformer outputs as channels to be dereverberated. In one realisation we process directly in the spherical harmonic domain where the beampatterns are mutually orthogonal. In a second realisation, which is not limited to spherical microphone arrays, beams are pointed in the direction of dominant reflections. Simulations demonstrate that in both cases reverberation is significantly reduced and, in the best case, clarity index is improved by 15 dB.**

***Index Terms*—speech dereverberation; spherical microphone array; beamforming**

I. Introduction

Reverberation degrades the quality of speech and in extreme cases can damage intelligibility for human listeners. Machine listening is less robust and so even moderate amounts of reverberation can severely reduce Automatic Speech Recognition (ASR) performance. Speech dereverberation is therefore important both for human communication systems and for human-machine interfaces [1].

Two distinct approaches to the enhancement of reverberant speech are common. Spatial filtering (or *beamforming*) improves the Direct-to-Reverberant Ratio (DRR) using directional selectivity to enhance sound with a particular Direction-of-Arrival (DOA). With this approach, since reverberation arriving from the 'look direction' is unaffected, there is generally still an audible reverberation tail. Alternatively, *dereverberation* refers to methods which attempt to explicitly model and suppress the reverberation. One promising approach is MultiChannel EQualisation (MCEQ) where Finite Impulse Response (FIR) models of the acoustic channels are estimated and used to design a set of deconvolution filters. Estimation of the channels remains a challenging problem leading to significant System Identification Errors (SIEs). The exact solution to the inverse problem [2] is sensitive to these SIEs and so robust solutions are preferred. Channel shortening [3]–[5] achieves this robustness by relaxing the requirement for perfect equalisation of the early reflections. In so doing the suppression of late reverberation is improved.

In this paper we are specifically concerned with the application of MCEQ methods to Spherical Microphone Arrays (SMAs), although some of the results may be generalised to arbitrary microphone arrays. SMAs typically have a large number of microphone elements and are of particular interest because they allow direction invariant beampatterns to be produced. As our starting point, we assume that estimates of the impulse responses from a source to each of the microphones in the array are available with some level of SIE. The question we wish to address is how can we best use these channel estimates to produce a dereverberated speech signal. We consider the application of MCEQ in three domains, namely (1) the spatial domain, that is, directly on the channel estimates; (2) the Spherical Harmonic Domain (SHD) and (3) the beamformed output domain. The first case is the baseline approach that does not depend on the microphone array being spherical. The second and third cases involve a transformation of the source-microphone channels into a smaller number of channels which are the inputs to MCEQ. They can both, therefore, be viewed as beamforming operations where the SHD is simply a special case in which all the beams are orthogonal. Transformation to the SHD representation depends on the array geometry being spherical. The beampatterns produced in the third case depend on the array geometry. For SMAs these will be direction invariant, but the principle of applying MCEQ to beamformed channels can equally be applied to arrays of arbitrary geometry.

The novel contribution of this paper is a two-step approach to dereverberation in which beamforming is applied prior to and independently of dereverberation processing. Alternative multi-stage approaches to dereverberation have used a single beamformer. For example, in [6] spatial filtering is used to achieve dereverberation and is followed by a noise reduction process whereas in [7] long-term linear prediction simultaneously dereverberates multiple channels which are subsequently

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609465.

combined using a beamformer.

II. Formulation

A. Alternative channel transformations

Consider Q microphones positioned on the surface of a sphere with radius r where the angle of the q-th microphone in spherical coordinates, $\theta = \arccos(z/\sqrt{x^2+y^2+z^2})$ and $\phi = \arctan(y/x)$, is $\Omega_q = (\theta_q, \phi_q)$. Let the impulse response from a point in space to the output of the q-th microphone be $h_q(t, r, \Omega_q)$, where t denotes time. Taking the Fourier transform, $h_q(t, r, \Omega_q)$ can each be expressed in the frequency domain as

$$H_q(\omega, r, \Omega_q) = \int_{-\infty}^{\infty} h_q(t, r, \Omega_q) \exp(-j\omega t) dt \tag{1}$$

where ω is the angular frequency and j is the unit imaginary number.

The Spherical Fourier Transform (SFT) over the surface of the sphere can be approximated up to spherical harmonic order N as the weighted sum of the microphone signals

$$H_{lm}(\omega, r) \approx \sum_q w_q Y_{lm}^*(\Omega_q) H_q(\omega, r, \Omega_q),\ l \leq N, |m| \leq l \tag{2}$$

where Y_{lm} is the spherical harmonic of order l and degree m and $\{w_q\}_1^Q$ are the weights of the sampling scheme [8]. These spherical harmonic coefficients are often called eigenbeams. The approximation of (2) is valid provided $kr < N$ (where $k = \omega/c$ and c is the speed of sound), the $Q \geq (N+1)^2$ sensors are approximately equally distributed over the sphere and the sampling weights are chosen appropriately [9].

Transformation of each eigenbeam back into the time domain yields an impulse response

$$h_{lm}(t, r) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H_{lm}(\omega, r) \exp(j\omega t) d\omega \tag{3}$$

which is equivalent to that of a microphone at the centre of the sphere with directivity corresponding to that of the associated spherical harmonic.

Alternative directivity patterns can be obtained through weighted combinations of eigenbeams [10]

$$H_z(\omega, r) = \sum_l \sum_m W_{lm}^{(z)} H_{lm}(\omega, r) \tag{4}$$

where $H_z(\omega, r)$ is the frequency response of the z-th beamformer and $W_{lm}^{(z)}$ is the complex weight of the eigenbeam with order l and degree m for that beamformer. As in (3), the inverse Fourier transform of $H_z(\omega, r)$ gives the impulse response from the source to the output of the z-th beamformer, $h_z(t, r)$. Many approaches to selecting the beamformer weights are possible, any of which could be applied in this context. In this paper, we choose the Plane-Wave Decomposition (PWD) beamformer [11] because it maximises the Directivity Index (DI) of the beam pattern in the look direction and is independent of the received signal. Its weights are

$$W_{lm}^{(z)}(\omega, r) = \frac{1}{b_l(kr)} Y_{lm}(\Psi_z) \tag{5}$$

where Ψ_z is the look direction and the mode strength, $b_l(kr)$, depends on the sphere configuration (e.g. open or rigid baffle) [12].

B. Multichannel Equalisation

A particular Γ channel system can be represented by the set of L-tap impulse responses $\mathbf{h}_\gamma = \begin{bmatrix} h_\gamma(0) & h_\gamma(1) & \dots & h_\gamma(L-1) \end{bmatrix}^T$ for $\gamma = 1, 2, \dots, \Gamma$. Our aim is to design a set of L_i-tap inverse filters $\mathbf{g}_\gamma = \begin{bmatrix} g_\gamma(0) & g_\gamma(1) & \dots & g_\gamma(L_i-1) \end{bmatrix}^T$ for $\gamma = 1, 2, \dots, \Gamma$ such that the Equalized Impulse Response (EIR), $\mathbf{d}$, is given by

$$\mathbf{H}\mathbf{g} = \mathbf{d} \tag{6}$$

where

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_1 & \mathbf{H}_2 & \dots & \mathbf{H}_\Gamma \end{bmatrix}, \tag{7}$$

$$\mathbf{g} = \begin{bmatrix} \mathbf{g}_1^T & \mathbf{g}_2^T & \dots & \mathbf{g}_\Gamma^T \end{bmatrix}^T, \tag{8}$$

$$\mathbf{d} = \begin{bmatrix} d(0) & d(1) & \dots & d(L+L_i-2) \end{bmatrix}^T \tag{9}$$

and $\mathbf{H}_\gamma$ is the $(L+L_i-1) \times L_i$ convolution matrix of $\mathbf{h}_\gamma$.

The Multiple-input/output INverse Theorem (MINT) solution [2] demands that the equalised impulse response be a delayed unit impulse

$$\mathbf{d}_{\text{MINT}} = \begin{bmatrix} \underbrace{0 \quad \dots \quad 0}_{\tau} & 1 & 0 & \dots & 0 \end{bmatrix}^T_{[(L+L_i-1)\times 1]} \tag{10}$$

where τ is the delay. In this case the inversion is perfect and so complete dereverberation is achieved. However, in blind estimation problems, one only has access to an estimate of the channel

$$\hat{\mathbf{h}}_\gamma = \mathbf{h}_\gamma + \boldsymbol{\varepsilon}_\gamma \tag{11}$$

where $\boldsymbol{\varepsilon}_\gamma$ is the estimation error and both $\hat{\mathbf{h}}_\gamma$ and $\boldsymbol{\varepsilon}_\gamma$ have the same dimensions as $\mathbf{h}_\gamma$. It has been found that equalisation filters which satisfy the MINT solution for the estimated channel are ineffective (and often counterproductive) in dereverberating the true acoustic channels. So, for robust MCEQ, the challenge is to use the estimated FIR channels, $\{\hat{\mathbf{h}}_\gamma\}_{\gamma=1}^\Gamma$, to design a set of equalising filters, $\mathbf{g}$, such that convolution with true channels, $\{\mathbf{h}_\gamma\}_{\gamma=1}^\Gamma$, as in (6), gives an EIR with specific properties. Rather than using prescribed values for $\mathbf{d}$, the channel shortening approach, as exemplified by Relaxed Multichannel Least Squares (RMCLS) [3], [13], defines a relaxation window, corresponding to the Times-of-Arrival (TOAs) of the direct path and early reflections, in which the response is unconstrained. The coefficients of the remainder of the EIR, which correspond to the reverberation tail, are constrained to be zero. These relaxed constraints are fulfilled by $\mathbf{g}$ which minimises the cost function [3]

$$J = ||\mathbf{W}(\hat{\mathbf{H}}\mathbf{g} - \mathbf{d}_{\text{MINT}})||_2^2 \tag{12}$$

where $\hat{\mathbf{H}} = \begin{bmatrix}\hat{\mathbf{H}}_1 & \hat{\mathbf{H}}_2 & \dots & \hat{\mathbf{H}}_\Gamma\end{bmatrix}$, $\hat{\mathbf{H}}_\gamma$ is the $(L+L_i-1)\times L_i$ convolution matrix of $\hat{\mathbf{h}}_\gamma$, $\mathbf{W} = \mathrm{diag}\{\mathbf{w}\}$ and

$$\mathbf{w} = \big[\underbrace{1 \quad 1 \quad \dots \quad 1}_{\tau} \quad \underbrace{1 \quad 0 \quad \dots \quad 0}_{L_w} \quad 1 \ \dots \ 1\big]^T_{[(L+L_i-1)\times 1]} \tag{13}$$

determines which taps in the EIR are constrained. The leading τ ones ensure that no energy in the EIR precedes the direct path time of arrival. The length of the relaxation window is denoted L_w and has a single one to avoid the trivial solution. Since multiple solutions achieve the minimum value (zero) in (12), the RMCLS solution is chosen to be that which yields the minimum ℓ_2-norm according to

$$\mathbf{g} = \left(\mathbf{W}\hat{\mathbf{H}}\right)^+ \mathbf{W}\mathbf{d}_{\text{MINT}} \tag{14}$$

where $(\cdot)^+$ represents the Moore-Penrose pseudo-inverse.

III. Experimental investigation

Ground truth Acoustic Impulse Responses (AIRs) were simulated for a 32-channel rigid spherical microphone array with r=4.2 cm using a modification of the image-source method [14], [15] for a room with dimensions 4×6×3 m and source-microphone distance of 1.5 m. The sample rate f_s was 8 kHz, N was 3 and wall absorption coefficients were set such that three different Reverberation Times (RTs) were simulated – 250, 400 and 550 ms. As in [3]–[5], [16], SIEs were added to the microphone channel AIRs in the form of White Gaussian Noise (WGN). These were weighted to achieve the desired Normalized Projection Misalignment (NPM) [17] of -30 dB using the algorithm in [18]. Inverse filters were designed according to (14) for three transformations of the estimated AIRs: (1) 'chan32' — The unmodified $Q = 32$ channel AIRs; (2) 'eig16' — the $(N+1)^2 = 16$ channel eigenbeam impulse responses; and (3) 'bf7' — the $Z = 7$ beamformer impulse responses found by pointing a PWD beamformer in the DOAs of the direct path and first order reflections. It is assumed that the DOAs of these early reflections are available using, for example, any of the algorithms reviewed in [19]. For the purposes of this investigation ground truth values were used to avoid introducing possible errors which are unrelated to the topic of interest.

To obtain spatially averaged results, 50 Monte Carlo trials were conducted for each test condition. In each trial the (x, y) coordinates of the microphone array were drawn from a uniform distribution while the height was fixed at 1.5 m. Similarly, the source position was drawn from a circle on the horizontal plane of the microphone defined by the source-microphone distance.

The effectiveness of MCEQ in each domain is evaluated visually, using the Energy Decay Curve (EDC) of the EIR averaged across all trials in each test condition, and quantitatively, using the clarity index, C_{50}, defined as [1]

$$C_{50} = 10\log_{10}\left(\frac{\sum_{\eta=0}^{\eta_e} d^2(\eta)}{\sum_{\eta=\eta_e+1}^{\infty} d^2(\eta)}\right) \text{ dB} \tag{15}$$

where η is the discrete time sample index and $\eta_e = 0.05 f_s$. C_{50} has been shown to be a good predictor of speech intelligibility and ASR performance [20], [21]. For the spatial domain and the SHD the baseline performance for each trial is taken as the average across all channels. For the beamformer domain the baseline performance is taken directly from the beamformer which is steered towards the direct path DOA since one would expect this to have the highest C_{50}.

IV. Results and Discussion

Figure 1 shows the EDCs for each test condition averaged across 50 trials. Note that the horizontal axis is adjusted according to the varying RT whereas the vertical scale is the same in all three plots. Considering first the results for a RT of 250 ms (Fig. 1(top)), the unprocessed 'chan32' and 'eig16' responses decay exponentially (which is seen as a linear decay on the logarithmic scale). On the other hand, the spatial selectivity of the PWD beamformer response gives a large initial decay at around 8 ms, because the direct path is relatively large compared to the early reflection arriving from other directions. However, following this initial drop, the decay rate follows that of the 'chan32' and 'eig16' responses. For all three processing domains the EDC is roughly equal or above the corresponding curves for the unprocessed impulse responses up until 50 ms. This is a direct result of the 50 ms relaxation window used in the RMCLS algorithm — these taps were unconstrained. After 50 ms there is significant drop in all the EDCs with 'bf7' giving the best performance (largest drop), followed by 'eig16'. The processed 'chan32' EDC is below that of the unprocessed 'chan32' but is only marginally better than the unprocessed 'bf7' curve. This suggests that it may be preferable to use a beamformer in isolation than to try to apply MCEQ in the spatial domain.

The overall trends observed for 250 ms RT can also be seen for 400 ms and 550 ms in Fig. 1(b) and Fig. 1(c), respectively. The main difference is that as the RT increases the extent of the benefit delivered by MCEQ reduces. For example, for 'bf7' the EIR EDC after 50 ms is -32 dB (RT: 250 ms), -20 dB (RT: 400 ms) and -16 dB (RT: 550 ms).

To make a quantitative comparison between the effectiveness of applying MCEQ to different transformations of the acoustic channels we calculate the clarity index as defined in (15) for each EIR separately. Figure 2 shows the distribution of C_{50} for the 50 Monte Carlo trials in each test condition (i.e. each combination of RT, domain and processed/unprocessed). Clearly, for the unprocessed impulse responses, C_{50} decreases as RT increases, as one would expect. Also, in agreement with the EDCs in Fig. 1, MCEQ increases the C_{50} over the unprocessed impulse responses and the extent of the increase decreases with increasing RT. For MCEQ in the spatial domain ('chan32') the relative improvement in average C_{50} is 6.6 dB (RT: 250 ms), 4.2 dB (RT: 400 ms) and 3.1 dB (RT: 550 ms). In the SHD ('eig16') the relative improvements are 11.8 dB (RT: 250 ms), 8.2 dB (RT: 400 ms) and 6.45 dB (RT: 550 ms). Finally, for the beamformer output domain ('bf7') the improvements are 10.1 dB (RT: 250 ms), 8.3 dB (RT: 400 ms)

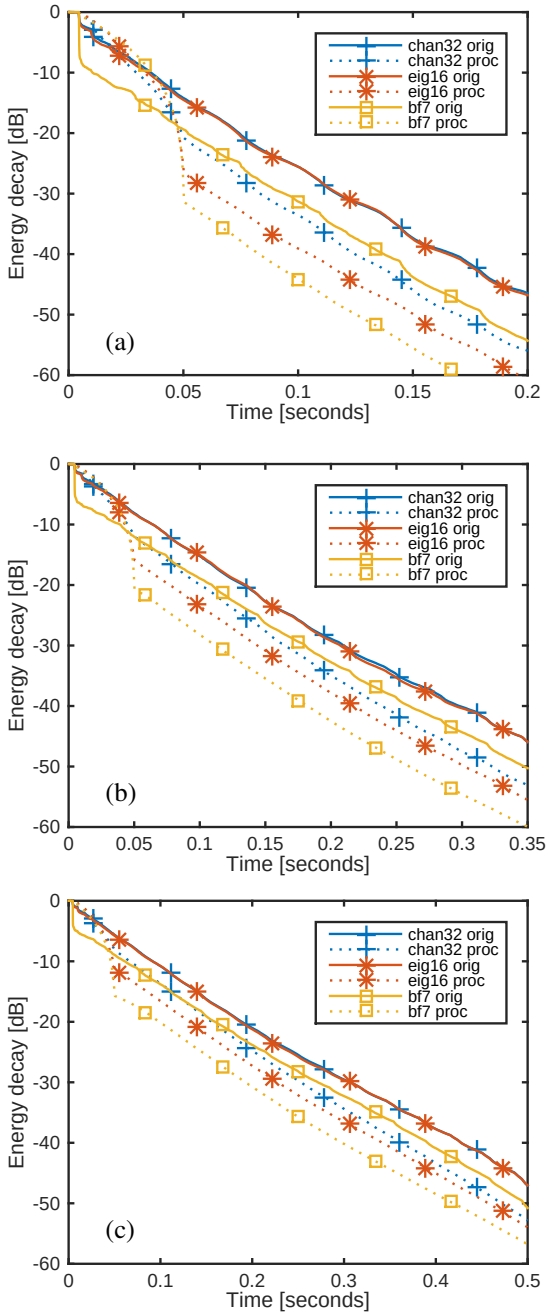

Fig. 1. EDCs for room with RT of 250 ms (a), 400 ms (b) and 550ms (c).

and 7.0 dB (RT: 550 ms). So, the relative improvement of MCEQ for 'eig16' and 'bf7' are roughly the same. The consistently higher C_{50} values of 'bf7' compared to 'eig16' therefore appear to be due to the fact that the unprocessed impulse responses are already less reverberant due to their spatial selectivity and their being targeted at directions of known strong early energy. The boxplots of Fig. 2 also suggest that there is more variation in C_{50} of dereverberated 'bf7' impulse responses than for the other transformations considered, especially for low RT. We suspect that this indicates that 'bf7' is more sensitive than the other transformations to the pattern

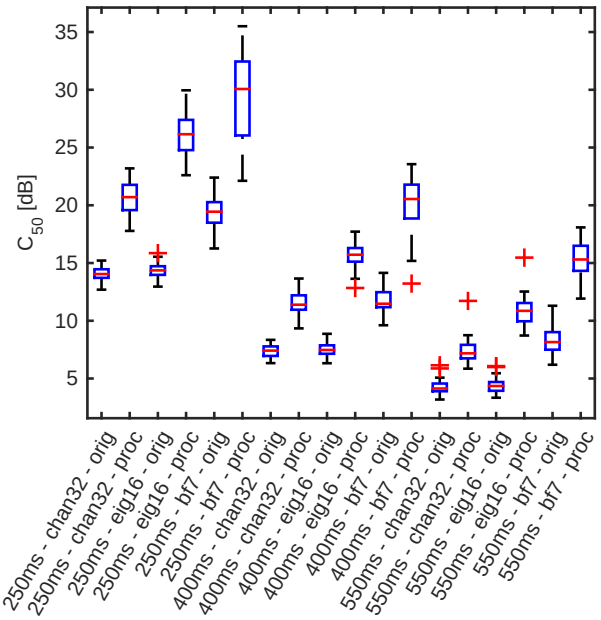

Fig. 2. Distribution of C_{50} before and after MCEQ for 50 Monte Carlo trials for each of 3 different transformations and 3 different RTs.

of early reflections encountered in each trial. This seems plausible, since for some source-microphone configurations first order reflections may arrive from more closely spaced directions than others. In cases where multiple reflections arrive from similar directions, their associated beamformers will have less spatial diversity than if the reflections were well distributed and so less dereverberation will be achieved.

Comparing 'bf7 - proc' to 'chan32 - orig' indicates the total extent of dereverberation which could be obtained using the two-step approach where beamforming ('bf7') is followed by RMCLS MCEQ. For these experiments the C_{50} improvements were 15.4 dB (RT: 250 ms), 12.7 dB (RT: 400 ms) and 11.1 dB (RT: 550 ms).

V. Conclusion

A novel two-step approach to dereverberation has been presented in which estimated acoustic channels containing system identification errors are transformed from the spatial domain into a beamformed domain before applying MCEQ. For high-order spherical microphone arrays, the SHD is a natural choice of beamformed domain. Alternatively, knowledge of the DOAs of strong reflections can be used to choose the look directions for a bank of conventional beamformers. In simulations both transformations led to improved dereverberation with the RMCLS algorithm. Best case performance of up to 15 dB improvement in C_{50} was achieved when equalising seven PWD beamformer channels directed at the direct path and first order reflection DOAs.

REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Springer, 2010.

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[3] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010.

[4] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[5] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 9, pp. 1379–1390, Sep. 2014.

[6] E. A. P. Habets and J. Benesty, "A two-stage beamforming approach for noise reduction and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 945–958, May 2013.

[7] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *Proc. REVERB Challenge Workshop*, vol. 1, Florence, Italy, 2014, pp. 1–8.

[8] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 135–143, Jan. 2005.

[9] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1003–1010, Mar. 2007.

[10] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, May 2002, pp. 1781–1784.

[11] B. Rafaely, "Plane-wave decomposition of the pressure on a sphere by spherical convolution," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2149–2157, Oct. 2004.

[12] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*, 1st ed. London: Academic Press, 1999.

[13] W. Zhang, E. A. P. Habets, and P. A. Naylor, "A system-identification-error-robust method for equalization of multichannel acoustic systems," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, USA, Mar. 2010.

[14] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[15] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid sphere room impulse response simulation: algorithm and applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, Sep. 2012.

[16] R. Rashobh and A. W. H. Khong, "A multichannel time-domain subspace approach exploiting multiple ttime-delay for acoustic channel equalization," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 418–422.

[17] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[18] W. Zhang and P. A. Naylor, "An algorithm to generate representations of system identification errors," *Research Letters in Signal Processing*, vol. 2008, pp. 13:1–13:4, Jan. 2008.

[19] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *J. Acoust. Soc. Am.*, vol. 131, pp. 2828–2840, 2012.

[20] A. Tsilfidis, I. Mporas, J. Mourjopoulos, and N. Fakotakis, "Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing," *Computer Speech and Language*, vol. 27, no. 1, pp. 380–395, 2013.

[21] P. Parada, D. Sharma, and P. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4718–4722.