LINEAR PREDICTION BASED DEREVERBERATION FOR SPHERICAL MICROPHONE ARRAYS

Alastair H. Moore and Patrick A. Naylor

Imperial College London, Dept. of Electrical and Electronic Engineering

ABSTRACT

Dereverberation is an important preprocessing step in many speech systems, both for human and machine listening. In many situations, including robot audition, the sound sources of interest can be incident from any direction. In such circumstances, a spherical microphone array allows direction of arrival estimation which is free of spatial aliasing and direction-independent beam patterns can be formed. This contribution formulates the Weighted Prediction Error algorithm in the spherical harmonic domain and compares the performance to a space domain implementation. Simulation results demonstrate that performing dereverberation in the spherical harmonic domain allows many more microphones to be used without increasing the computational cost. The benefit of using many microphones is particularly apparent at low signal to noise ratios, where for the conditions tested up to 71% improvement in speech-to-reverberation modulation ratio was achieved.

Index Terms— speech dereverberation, spherical microphone array, spherical harmonic domain, multichannel linear prediction, weighted prediction error

1. INTRODUCTION

Dereverberation processing can improve the accuracy of automatic speech recognition and, in some cases, the perceived quality of speech transmitted over a communications channel [1]. We are particularly interested in Spherical Harmonic (SH) domain processing of Spherical Microphone Array (SMA) signals due to the convenience with which beams can be steered in any direction without affecting their directivity pattern. This is important for applications such as robot audition where the positions of sound sources relative to the microphone array are unknown in advance. Dereverberation for SMAs has been presented for the case where detailed knowledge of the environment can be reliably estimated or is known in advance [2–4]. On the other hand, blind dereverberation attempts to perform such an enhancement without prior knowledge of the room or the positions of the source and microphone array.

In a recent comparative study of single and multichannel approaches, long-term linear prediction in subbands was shown to be particularly effective [5]. The basis of this approach has been developed over a number of years. The Weighted Prediction Error (WPE) approach [6], determines a multichannel linear prediction filter whose inverse estimates the desired speech signal at one of the input channels. The method processes each subband independently whereupon an iterative algorithm alternately estimates the time-varying variance of the desired signal and the linear prediction coefficients. To dereverberate multiple channels, the WPE algorithm can be applied independently for each channel. The Generalized WPE (GWPE) algorithm [7] extends WPE to the Multiple-Input-Multiple-Output (MIMO) case where the iterative approach alternately estimates the spatial covariance of the dereverberated signals and the linear prediction filter coefficients. It was further shown in [7] that GWPE and WPE are identical in the special case of a diagonal spatial covariance matrix. Although unlikely to be the case in practice, making such an assumption achieves a significant amount of dereverberation and extensive computational savings.

In the context of SMAs, the WPE could be applied in the space domain, that is directly to the microphone signals, before transforming the dereverberated signals to the SH domain for subsequent processing, such as Direction-of-Arrival (DOA) estimation and beamforming. However, we propose that WPE be applied after transforming the signals to the SH domain so that a large number of microphones can be used to improve robustness to noise without increasing the computational cost of the dereverberation process.

The contributions of this paper are to formulate the WPE algorithm in the SH domain and to evaluate its performance compared to WPE operating in the space domain over a range of Signal-to-Noise Ratios (SNRs).

2. TECHNICAL BACKGROUND

Working in the Short Time Fourier Transform (STFT) domain with frequency index ν and frame index ℓ, the signal received at the q-th channel of a microphone array is given by

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

978-1-5090-2007-2/16/$31.00

$$X_q(\nu,\ell) = D_q(\nu,\ell) + X_q^{(r)}(\nu,\ell) + V_q(\nu,\ell) \quad (1)$$

where $D_q(\nu,\ell)$ is the desired signal, which includes the direct path and some early reflections, $X_q^{(r)}(\nu,\ell)$ is the undesired reverberation which is correlated with $D_q(\nu,\ell)$ and $V_q(\nu,\ell)$ is additive noise which is not correlated with $D_q(\nu,\ell)$. Collecting the observation of $X_q(\nu,\ell)$ over N_ℓ frames into a vector gives

$$\mathbf{x}_q(\nu,\ell) = \left[X_q(\nu,\ell-N_\ell+1),\ldots,X_q(\nu,\ell)\right]^T \quad (2)$$

and $\mathbf{d}_q(\nu,\ell)$, $\mathbf{x}_q^{(r)}(\nu,\ell)$ and $\mathbf{v}_q(\nu,\ell)$ are similarly defined. Since each subband is processed independently, we consider only a single subband and drop the dependence on ν

$$\mathbf{x}_q(\ell) = \mathbf{d}_q(\ell) + \mathbf{x}_q^{(r)}(\ell) + \mathbf{v}_q(\ell). \quad (3)$$

The aim of WPE is to estimate the desired signal at the q-th microphone by predicting the late reverberation, $\hat{\mathbf{x}}_q^{(r)}(\ell)$, and subtracting it

$$\hat{\mathbf{d}}_q(\ell) = \mathbf{x}_q(\ell) - \hat{\mathbf{x}}_q^{(r)}(\ell) \quad (4)$$

$$= \mathbf{x}_q(\ell) - \mathbf{X}(\ell-\tau)\mathbf{g}_q \quad (5)$$

where $\mathbf{g}_q = \left[\mathbf{g}_{q,1}^T,\ldots,\mathbf{g}_{q,N_q}^T\right]^T$ is the multichannel prediction filter,

$$\mathbf{g}_{q,q'} = \left[G_{q,q'}(0),\ldots,G_{q,q'}(N_g-1)\right]^T,$$

$$\mathbf{X}(\ell-\tau) = \left[\mathbf{X}_1(\ell-\tau),\ldots,\mathbf{X}_{N_q}(\ell-\tau)\right]$$

and

$$\mathbf{X}_q(\ell-\tau) = \left[\mathbf{x}_q(\ell-\tau-N_g+1),\ldots,\mathbf{x}_q(\ell-\tau)\right]$$

is the $N_\ell \times N_g$ convolution matrix for the q-th channel including a delay of τ frames. The purpose of the delay is to avoid overwhitening, where short term correlations in the source signal are equalized out. The core of the WPE algorithm is an iterative approach to estimating $\mathbf{g}_q$ from which an estimate of the desired signal follows directly according to (5). Using $(\cdot)^{(n)}$ to denote the value of a variable on the n-th iteration, the estimated variance of the desired signal is updated as

$$\hat{\boldsymbol{\lambda}}_q^{(n)} = \max\left\{\left|\hat{\mathbf{d}}_q^{(n)}\right|^2, \epsilon\right\} \quad (6)$$

and the prediction filter is updated as

$$\hat{\mathbf{g}}_q^{(n)} = \left(\mathbf{X}_\tau{}^H \mathcal{D}^{-1}_{\hat{\boldsymbol{\lambda}}_q^{(n)}} \mathbf{X}_\tau\right)^{-1} \mathbf{X}_\tau{}^H \mathcal{D}^{-1}_{\hat{\boldsymbol{\lambda}}_q^{(n)}} \mathbf{x}_q \quad (7)$$

where $\mathcal{D}^{-1}_{\hat{\boldsymbol{\lambda}}_q^{(n)}} = \mathrm{diag}\left\{1/\hat{\boldsymbol{\lambda}}_q^{(n)}\right\}$, ϵ in (6) is a lower bound on the variance to avoid singularities and the dependence on the frame index has been omitted for clarity. The next iteration proceeds with (5) substituting $\hat{\mathbf{d}}_q = \hat{\mathbf{d}}_q^{(n+1)}$ and $\mathbf{g}_q = \hat{\mathbf{g}}_q^{(n)}$.

3. WPE IN THE SHD

The Spherical Fourier Transform (SFT) of a square-integrable function defined on the surface of a sphere, $A(\Omega)$, where $\Omega = (\theta,\phi)$ is the position on the sphere, can be approximated by sampling at discrete positions according to [8]

$$\underline{A}_{l,m} = \sum_q w_q A\left(\Omega_q\right)\left[Y_l^m(\Omega_q)\right]^* \quad (8)$$

where $\{\Omega_q\}_{q=1\ldots N_q}$ are the sample positions, $\{w_q\}_{q=1\ldots N_q}$ are the weights of the sampling scheme,

$$Y_l^m(\Omega) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}} P_l^m\left(\cos\phi\right) e^{im\theta} \quad (9)$$

is the SH function of order $l \in \mathbb{N}$ and degree $m \in \{-l\ldots l\}$ and $P_l^m(\cdot)$ is the associated Legendre function. The maximum order, L, for which the approximation in (8) holds depends on the spatial bandwidth of $A(\Omega)$, the number of sample points, N_q, and their distribution over the sphere. For clarity of notation, we refer to each of the $N_p = (L+1)^2$ SH coefficients using a single index $p = F(l,m) = l^2+l+m+1$.

Substituting $A\left(\Omega_q\right) = X_q(\nu,\ell)$ into (8), the SFT of the sampled sound field is given by $\{\underline{X}_p(\nu,\ell)\}_{p=1\ldots N_p}$. Similarly, when transformed to the SH domain, $\{D_q(\nu,\ell)\}_{q=1\ldots N_q}$, $\{X_q^{(r)}(\nu,\ell)\}_{q=1\ldots N_q}$ and $\{V_q(\nu,\ell)\}_{q=1\ldots N_q}$ become $\{\underline{D}_p(\nu,\ell)\}_{p=1\ldots N_p}$, $\{\underline{X}_p^{(r)}(\nu,\ell)\}_{p=1\ldots N_p}$ and $\{\underline{V}_p(\nu,\ell)\}_{p=1\ldots N_p}$, respectively. Following the exposition in Sec. 2, we consider a single subband and collate signals into vectors of N_ℓ frames as in (2) such that

$$\underline{\mathbf{x}}_p(\ell) = \underline{\mathbf{d}}_p(\ell) + \underline{\mathbf{x}}_p^{(r)}(\ell) + \underline{\mathbf{v}}_p(\ell) \quad (10)$$

and

$$\hat{\underline{\mathbf{d}}}_p(\ell) = \underline{\mathbf{x}}_p(\ell) - \underline{\mathbf{X}}(\ell-\tau)\underline{\mathbf{g}}_p \quad (11)$$

where $\underline{\mathbf{X}}(\ell-\tau)$ is the delayed multi-SH convolution matrix and $\underline{\mathbf{g}}_p$ is the multi-SH linear prediction filter which estimates the reverberant component of the p-th SH signal.

Following (6) and (7) the iterative updates for WPE in the SH domain can be found as

$$\hat{\underline{\boldsymbol{\lambda}}}_p^{(n)} = \max\left\{\left|\hat{\underline{\mathbf{d}}}_p^{(n)}\right|^2, \underline{\epsilon}\right\} \quad (12)$$

and

$$\hat{\underline{\mathbf{g}}}_p^{(n)} = \left(\underline{\mathbf{X}}_\tau{}^H \mathcal{D}^{-1}_{\hat{\underline{\boldsymbol{\lambda}}}_p^{(n)}} \underline{\mathbf{X}}_\tau\right)^{-1} \underline{\mathbf{X}}_\tau{}^H \mathcal{D}^{-1}_{\hat{\underline{\boldsymbol{\lambda}}}_p^{(n)}} \underline{\mathbf{x}}_p. \quad (13)$$

4. EVALUATION

The efficacy of WPE dereverberation in the SH domain is compared to WPE in the space domain as a function of the SNR, number of microphones, N_q, and linear prediction filter length, N_g, using computer simulations.

4.1. Method

The image method [9] was used to simulate the Acoustic Impulse Response (AIR) from a single source to each microphone in three, co-located open SMAs, each with radius 4.2 cm and with $N_q \in \{4, 8, 32\}$. An open configuration, where the microphones are placed in free space, was selected to ensure that the results are as general as possible; had a rigid configuration, where the microphones are placed on a baffle, been used, any benefit for SH domain processing might have been attributed to scattering effects having a negative impact on space domain processing. The room size was 4×6×3 m, the Reverberation Time (RT) was 0.5 s and the source-array distance was 1.5 m. On each of 10 trials per test condition, speech from the TIMIT database [10] was concatenated to form utterances of at least 6 s and convolved with the AIRs for a randomly selected SMA position and relative source direction. The level of the reverberant speech was set to normalize the active level of the direct path speech component according to [11] and independent white Gaussian noise added to each sensor to obtain Direct-to-Incoherent Noise Ratios (DINRs) of 0, 10 and 20 dB. Signals were sampled at 8 kHz and the STFT used 64 ms Hamming-windowed frames with 50% overlap.

For SMAs with $N_q = 4$ and 8, sensors were equally distributed over the sphere according to the vertices of the corresponding platonic solids and $\{w_q\}_{q=1\ldots N_q} = 4\pi/N_q$. For $N_q = 32$, the sensor angles were approximately equally distributed according to [12]. For each trial, the SMAs were rotated such that each had an arbitrarily selected reference sensor oriented towards the source direction, which ensured that all arrays had access to the best sensor and the direction of the source with respect to the SMAs was fixed and so known a priori. A number of methods for estimating DOAs are available [13–17] and so this process is beyond the scope of the current paper.

For each SMA three dereverberation approaches were evaluated. 1) No processing: The microphone signals, $\{X_q(\nu, \ell)\}_{q=1\ldots N_q}$, were transformed directly to the SH domain, $\{\underline{X}_p(\nu, \ell)\}_{p=1\ldots N_p}$, according to (8) with $L = 1$ ($N_p = 4$) and a maximum directivity beamformer [18] was steered towards the source. 2) Space domain: WPE was applied to the space domain signals, as described in Sec. 2, before taking the discrete SFT and beamforming, as in condition 1. 3) SH domain: The microphone signals were transformed to the SH domain, dereverberated as proposed in Sec. 3, and then beamformed. Note that the number of SHs, N_p, was 4, regardless of N_q to ensure that the directivity pattern of the SH domain beamforming was the same in all test conditions.

A range of dereverberation filter lengths were tested, depending on the number of channels being processed. For space domain WPE with $N_q = 4$, $N_g \in \{1, 2, 4, 8, 16, 32\}$; with $N_q = 8$, $N_g \in \{1, 2, 4, 8, 16\}$ and with $N_q = 32$, $N_g \in \{1, 2, 3\}$. For SH domain WPE $N_g \in \{1, 2, 4, 8, 16, 32\}$, regardless of N_q. In all cases, we set $\tau = 2$, as in [7].

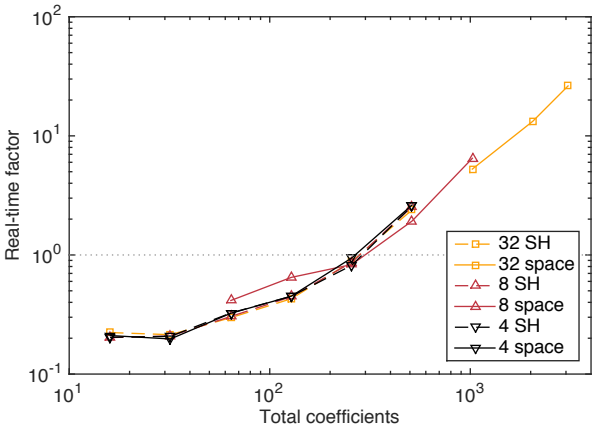


Fig. 1. Effect of the number of filter coefficients on algorithm run time as indicated by RTF for DINR of 0 dB. Legend indicates the number of microphones and the processing domain.

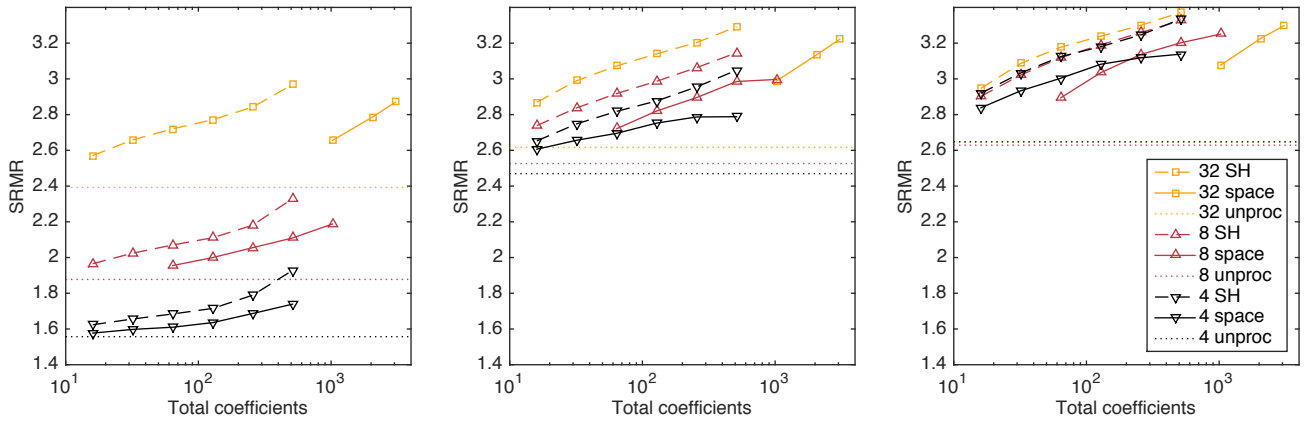
Other algorithm parameters were set empirically, leading to $\epsilon = \underline{\epsilon} = 0.01$ and the number of iterations was 5.

The results are evaluated in terms of the Speech-to-Reverberation Modulation Energy Ratio (SRMR) [19, 20], which gives an indication of how much the reverberation has been reduced, and Log Spectral Distortion (LSD) [1], which gives an indication of how similar to the clean, direct path speech the processed signals are.

4.2. Results

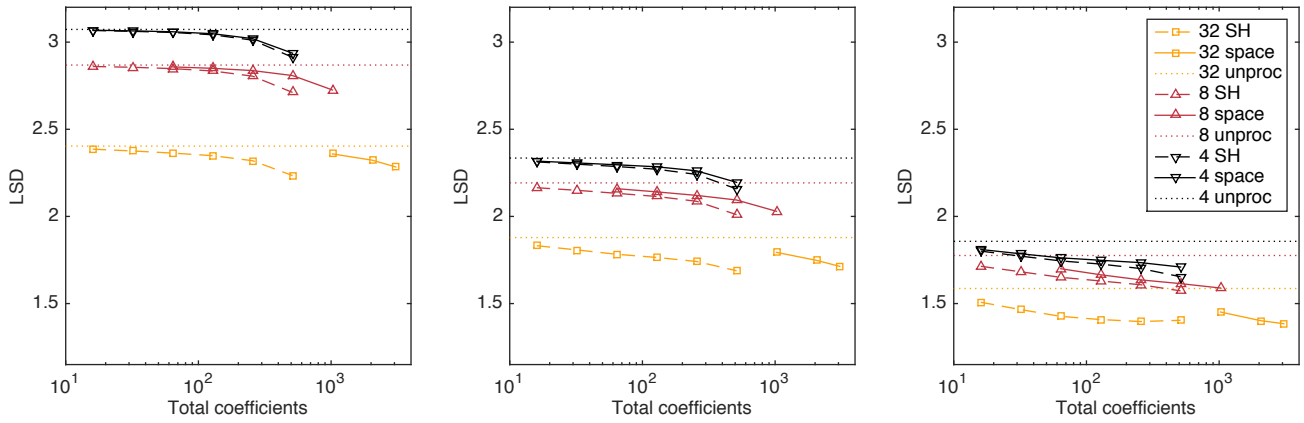
The results of dereverberation processing are shown as a function of the total number of coefficients, $N_g \times N_q^2$ for space domain dereverberation and $N_g \times N_p^2$ for SH domain dereverberation, which is related to the computational cost. As justification, Fig. 1 shows the real-time factor (computation time/signal duration) as a function of the total number of coefficients for each of the conditions at DINR of 0 dB. As can be seen, the real-time factor increases monotonically with number of coefficients regardless of the number of microphone channels or processing domain.

Figure 2 shows the SRMR of the beamformed signal for all the test conditions. The effect of varying the number of microphones is most pronounced when the DINR is low (0 dB). Even without dereverberation processing, using many microphones greatly improves the SRMR of the beamformed signal. With only 4 microphones, SH domain dereverberation gives an improvement of between 3% and 11% over space domain processing with longer filters giving more improvement. With $N_q = 8$, again SH domain dereverberation performs better. Of the conditions for which a direct comparison is possible, the improvement ranges from 6% to 10%. At $N_q = 32$ the smallest number of coefficients for space domain processing (1024) is already larger than the maximum number of coefficients



(a) (b) (c)

Fig. 2. Effect of the number of filter coefficients on SRMR for DINR of (a) 0 dB, (b) 10 dB and (c) 20 dB. Legend indicates the number of microphones and the processing domain.



(a) (b) (c)

Fig. 3. Effect of the number of filter coefficients on LSD for DINR of (a) 0 dB, (b) 10 dB and (c) 20 dB. Legend indicates the number of microphones and the processing domain.

considered for SH domain processing, so a like-for-like comparison is not possible. Instead one can interpret the space domain performance as shifting the SH performance curve to the right. That is, in the space domain it requires 1024 coefficients to achieve what can be done in the SH domain with only 32. Comparing the '32 SH' condition to the '4 space' condition, an overall improvement of between 63% and 71% is obtained with the same number of filter coefficients.

At higher DINRs the same trends are evident, although the benefit of more microphone channels is less pronounced. At 20 dB DINRs, the noise is sufficiently low that beamforming only performance is almost the same, independent of the number of microphones. However, it is still clear that dereverberation in the SH domain offers improved performance compared to the space domain and using more microphones further increases this advantage.

Figure 3 shows that LSD performance follows the same trends as for SRMR with SH processing achieving improved (lower) scores over space domain processing in all cases.

5. CONCLUSIONS

A novel formulation of the WPE algorithm which operates in the SH domain has been presented. In numerical simulations it has been demonstrated that for a 4-channel array by performing dereverberation in the SH domain between 3 and 10% improvement in SRMR can be obtained. Furthermore, increasing the number of channels to 32 led to an SRMR improvement of between 63 and 71%, without increasing the computational complexity of the dereverberation.

6. REFERENCES

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

[2] Y. Peled and B. Rafaely, "Method for dereverberation and noise reduction using spherical microphone arrays," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010.

[3] A. H. Moore, C. Evers, and P. A. Naylor, "Multichannel equalisation for high-order spherical microphone arrays using beamformed channels," in *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, July 2015.

[4] H. A. Javed, A. H. Moore, and P. A. Naylor, "Spherical microphone array acoustic rake receivers," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016.

[5] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The Reverb Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2013.

[6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[7] T. Yoshioka and T. Nakatani, "Generalization of multichannel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[8] B. Rafaely, *Fundamentals of Spherical Array Processing*, Springer Topics in Signal Processing. Springer, Berlin Heidelberg, 2015.

[9] 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.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," NIST Interagency/Internal Report (NISTIR) 4930, National Institute of Standards and Technology (NIST), Feb. 1993.

[11] "Objective measurement of active speech level," Mar. 1993.

[12] M. H. Acoustics, "EM32 Eigenmike microphone array release notes (v17.0)," Oct. 2013.

[13] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "3D source localization in the spherical harmonic domain using a pseudointensity vector," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aug. 2010.

[14] E. Mabande, H. Sun, K. Kowalczyk, and W. Kellermann, "Comparison of subspace-based and steered beamformer-based reflection localization methods," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aug. 2011.

[15] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1494–1505, Oct. 2014.

[16] C. Evers, A. H. Moore, and P. A. Naylor, "Multiple source localisation in the spherical harmonic domain," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, July 2014.

[17] A. H. Moore, C. Evers, P. A. Naylor, D. L. Alon, and B. Rafaely, "Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015.

[18] B. Rafaely, "Phase-mode versus delay-and-sum spherical microphone array processing," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 713–716, Oct. 2005.

[19] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[20] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2014.