UNDER-MODELLED BLIND SYSTEM IDENTIFICATION FOR TIME DELAY ESTIMATION IN REVERBERANT ENVIRONMENTS

Wei Xue, Mike Brookes, Patrick A. Naylor

Dept. of Electrical and Electronic Engineering, Imperial College London, UK
Email: {w.xue, mike.brookes, p.naylor}@imperial.ac.uk

ABSTRACT

In multichannel systems, acoustic time delay estimation (TDE) is a challenging problem in reverberant environments. Although blind system identification (BSI) based methods have been proposed which utilize a realistic signal model for the room impulse response (RIR), their TDE performance depends strongly on that of the BSI, which is often inaccurate in practice when the identified responses are under-modelled. In this paper, we propose a new under-modelled BSI based method for TDE in reverberant environments. An under-modelled BSI algorithm is derived, which is based on maximizing the cross-correlation of the cross-filtered signals rather than minimizing the cross-relation error, and also exploits the sparsity of the early part of the RIR. For TDE, this new criterion can be viewed as a generalization of conventional cross-correlation-based TDE methods by considering a more realistic model for the early RIR. Depending on the microphone spacing, only a short early part of each RIR is identified, and the time delays are estimated based on the peak locations in the identified early RIRs. Experiments in different reverberant environments with speech source signals demonstrate the effectiveness of the proposed method.

Index Terms— Time Delay Estimation, Room Acoustics, Reverberation, Microphone Arrays.

1. INTRODUCTION

Time delay estimation (TDE) is an important problem in microphone array signal processing as the direction of arrival (DOA) of the sound source can be straightforwardly computed from the TDE results. Reliable DOA estimation is important on applications such as speaker tracking [1, 2], beamforming-based multichannel speech enhancement [3, 4, 5] and dereverberation [6, 7].

The classical and most widely used methods for TDE are the generalized cross-correlation (GCC) based methods [8], among which GCC with Phase-Transform (GCC-PHAT) has received the most attention. When using more than two microphones, the Steering Response Power PHAT (SRP-PHAT) method [9] is a generalization of the GCC-PHAT. In addition, methods based on the parametric spatial correlation matrix [10] and on spatial linear prediction [11] have been proposed that formulate the multichannel TDE problem in a more compact mathematical structure.

Because the above methods generally assume the signals at different microphones are simply scaled and delayed versions of the source signal, their performances degrade dramatically in highly reverberant environments. To overcome this problem, methods based on a more realistic signal model have been developed [12, 13, 14] in which the time delays are computed from the room impulse responses (RIRs) estimated by multichannel blind system identification (BSI). However, the practical use of these methods is limited by the accuracy of multichannel BSI, which is based on cross-relation (CR) error minimization. Multichannel BSI generally requires a priori knowledge of the true RIR length, which is always unknown in practice. Moreover, the BSI performance is sensitive to noise [15]. In noisy and system under-modelling conditions, the multichannel BSI is unable to yield accurate RIR estimates, which consequently leads to erroneous TDE values.

In this paper, we propose a new under-modelled BSI based TDE method in reverberant conditions. Based on the previous work [16], an under-modelled BSI algorithm for TDE is derived, in which instead the CR error being minimized, the cross-correlation of the cross-filtered signals is maximized and the sparsity of the early RIR is exploited. For TDE, this method can be viewed as a generalization of the traditional cross-correlation based TDE methods that uses a more realistic model for the early RIR than a unit impulse. Only the early RIRs that correspond to the inter-microphone delays are estimated and the time delays are determined from the peak locations in the early RIR estimates. We demonstrate the effectiveness of the proposed method by conducting experiments in different reverberant environments with speech source signals.

2. SIGNAL MODEL

In a reverberant environment with a single sound source and an M-element microphone array, the time-domain signal received

This research has received funding from the EU 7th Framework Programme (FP7/2007-2013) project DREAMS under grant agreement ITN-GA-2012-316969.

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

by the i-th microphone at time index n can be expressed as:

$$x_i(n) = \mathbf{h}_i^T \mathbf{s}(n) + v_i(n),\ i = 1, 2, ..., M, \tag{1}$$

where $\mathbf{h}_i = [h_{i,0}\ h_{i,1}\ \dots\ h_{i,L-1}]^T$ is the $L \times 1$ true RIR of the i-th channel, $\mathbf{s}(n) = [s(n)\ s(n-1)\ \dots\ s(n-L+1)]^T$ is the source signal vector, and $v_i(n)$ is additive noise which is assumed to be white Gaussian and uncorrelated both with the source signal and with $v_j(n)$ for $j \neq i$.

Assuming the direct path component in $\mathbf{h}_i$ appears at sample τ_i, then the sample delay between the i-th channel and j-th channel can be estimated by comparing the time indexes of the direct paths as:

$$\tau_{ij} = \tau_i - \tau_j,\ i, j = 1, 2, ..., M, i \neq j. \tag{2}$$

According to the array geometry, τ_{ij} satisfies:

$$|\tau_{ij}| \leq \tau_{ij,\max} = \lfloor \frac{d_{ij} f_s}{c} \rfloor, \tag{3}$$

where d_{ij} is the microphone spacing, f_s is the sampling rate, and c is the sound speed. The floor function $\lfloor a \rfloor$ returns the largest integer not exceeding a.

3. TDE FROM A CROSS-FILTERING PERSPECTIVE

In GCC based methods [8], the signals are first pre-filtered to reduce the auto-correlation of the signals, and then the cross-correlations corresponding to different time delays are computed and used as the GCC function for the each time delay. The TDE is performed by finding the sample index of the largest peak in the GCC function.

In Fig. 1, we present a new generalized TDE framework from the cross-filtering perspective, where $G_i(z)$ is the Z-transform of the pre-filter, $y_i(n)$ is the pre-filtered signal, and $\hat{H}_i(z)$ is the Z-transform of estimated RIR of the i-th channel $\hat{\mathbf{h}}_i$. For a microphone pair, i and j, by convolving the pre-filtered signal of each channel with the estimated RIR of the other channel, two cross-filtered signals $\tilde{y}_{ji}(n)$ and $\tilde{y}_{ij}(n)$ can be obtained. The GCC function measures the cross-correlation of these cross-filtered signals. The conventional GCC based methods can be seen as a special case of the TDE framework, by simplifying the RIR $\hat{\mathbf{h}}_i$ as a unit-impulse function which is written in vector form as

$$\hat{\mathbf{h}}_i = [\underbrace{0\ \dots\ 0}_{\hat{\tau}_i}\ 1\ \underbrace{0\ \dots\ 0}_{L-\hat{\tau}_i-1}]^T,\ i = 1, 2, ..., M, \tag{4}$$

where $\hat{\tau}_i$ is the estimation of τ_i. Then filtering $y_j(n)$ with $\hat{h}_i$ yields the time-delayed version of $y_j(n)$, and the GCC function reflects the cross-correlation between the time-delayed signals of $y_i(n)$ and $y_j(n)$. When the assumed $\hat{\tau}_i$ for $i = 1, 2, ..., M$ matches the true delay, the delayed signals are time-aligned, and the cross-correlation attains its maximum.

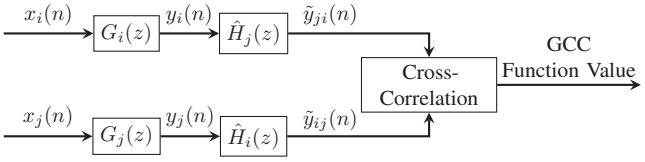

Fig. 1. Illustration of the GCC based methods for TDE from the cross-filtering perspective.

The problem in reverberant environments is that the RIRs cannot be simply modelled as a unit-impulse, and because of the late reverberation components, the cross-correlation does not necessarily have a maximum at the true direct path delay. A straightforward idea is to generalize the $\hat{\mathbf{h}}_i$ from unit-impulse function to a more realistic model of the RIRs. The unit-impulse function can be seen as an under-modelled representation of the RIR in which only a single non-zero tap is included. It has been shown that even with this simplified RIR, we can still often achieve an acceptable TDE performance by using the GCC methods in environments with low and moderate reverberation. This indicates that the cross-correlation based criterion has an inherent robustness against system under-modelling when assessing the correctness of the assumed RIRs. Therefore, it can be expected that the robustness to reverberation can be improved if we generalize the unit-impulse functions to include the early reflections. This motivates the proposed method which is presented in Sec. 4.

4. PROPOSED METHOD

Following Fig. 1, by considering early reflections in the RIR, a new method for TDE is proposed. Unlike the TDE methods which only concern the direct path location, it is much harder to determine the peak locations and amplitudes of the direct path and reflections by exhaustively checking all combinations of possible values. Here, an adaptive method is derived to find the optimal estimate of the early RIR, by maximizing the cross-correlation of the cross-filtered signals. In addition, as RIRs are always sparse over the period of early reflections, sparsity regularization is exploited to further improve the estimation performance.

4.1. Pre-Filtering

To reduce the negative effect of autocorrelation of speech signals on the cross-correlation based cost function, pre-filtering is usually needed. The PHAT weighting function has been shown to be effective compared with other weighting functions [17]. In our method, we therefore adopt the PHAT to pre-filter the signals $x_i(n)$ for $i = 1, 2, ..., M$. For each channel, the captured signal is first converted into the short-time Fourier transform (STFT) domain, and then in each time-frequency bin, the amplitude is normalized to be 1. Finally, the pre-

filtered signal $y_i(n)$ is obtained by transforming the normalized STFT spectrum back into the time domain by using the overlap-save method [18].

4.2. Optimization Problem for Under-modelled BSI

For each channel, we estimate an RIR with a length $K = 2\tau_{\max} + 3$ where $\tau_{\max} = \max\{\tau_{i1,\max}\}$. The reason for choosing such value of K is that the inter-microphone time delays are limited by (3), and the flooring effect in (3) should be considered such that the maximum true time delay can always be identified. Without loss of generality, we assume the direct path of the first channel lies at the central sample of its estimated RIR.

Denoting the $K\times 1$ vector $\hat{\mathbf{h}}_i = [\hat{h}_{i,0}\ \hat{h}_{i,1}\ \dots\ \hat{h}_{i,K-1}]^T$ as the under-modelled channel estimation, then following Fig. 1, cross-filtering $y_j(n)$ with $\hat{\mathbf{h}}_i$ results in $\tilde{y}_{ij}(n) = \hat{\mathbf{h}}_i^T\mathbf{y}_j(n)$ where $\mathbf{y}_j(n) = [y_j(n)\ y_j(n-1)\ \dots y_j(n-K+1)]^T$, and the cross-correlation between $\tilde{y}_{ij}(n)$ and $\tilde{y}_{ji}(n)$ is

$$\gamma_{ij} = \hat{\mathbf{h}}_i^T\mathbb{E}\{\mathbf{y}_j(n)\mathbf{y}_i^T(n)\}\hat{\mathbf{h}}_j, \tag{5}$$

where $\mathbb{E}\{\cdot\}$ denotes expectation.

By using all microphone pairs, to facilitate the derivation of least mean squares (LMS) updating in the next subsection, we define a cost function which combines the instantaneous values of γ_{ij} for all microphone pairs in (5):

$$\Upsilon(\hat{\mathbf{h}}, n) = \sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}\hat{\mathbf{h}}_i^T\mathbf{y}_j(n)\mathbf{y}_i^T(n)\hat{\mathbf{h}}_j = \hat{\mathbf{h}}^T\mathbf{R}(n)\hat{\mathbf{h}}, \tag{6}$$

where $\hat{\mathbf{h}} = [\hat{\mathbf{h}}_1^T\ \hat{\mathbf{h}}_2^T\ \dots\ \hat{\mathbf{h}}_M^T]^T$, and $\mathbf{R}(n)$ is an $(MK)\times(MK)$ matrix having the form

$$\mathbf{R}(n) = \begin{bmatrix} \mathbf{0}_{K\times K} & \mathbf{R}_{21}(n) & \cdots & \mathbf{R}_{M1}(n) \\ \mathbf{R}_{12}(n) & \mathbf{0}_{K\times K} & \cdots & \mathbf{R}_{M2}(n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{1M}(n) & \mathbf{R}_{2M}(n) & \cdots & \mathbf{0}_{K\times K} \end{bmatrix} \tag{7}$$

with $\mathbf{R}_{ji}(n) = \mathbf{y}_j(n)[\mathbf{y}_i(n)]^T$.

We aim to find the early RIRs which maximize the cross-correlation of the cross-filtered signals while also being sparse. Note that simply maximizing the cross-correlation will lead to the infinite solution of $\hat{\mathbf{h}}$. To avoid this, the unit-norm constraint is imposed on $\hat{\mathbf{h}}$. We finally define the cross-correlation based cost function as

$$J(\hat{\mathbf{h}}, n) = -\frac{\Upsilon(\hat{\mathbf{h}}, n)}{\|\hat{\mathbf{h}}\|_2^2}. \tag{8}$$

By adding the l_1-norm sparsity regularization, the early RIRs of different channels are estimated by solving the following optimization problem:

$$\hat{\mathbf{h}} = \arg\min_{\hat{\mathbf{h}}}\{\bar{J}(\hat{\mathbf{h}}) + \rho|\hat{\mathbf{h}}|_1\},\ \text{s.t. } \|\hat{\mathbf{h}}\|_2^2 = 1. \tag{9}$$

where $\bar{J}(\hat{\mathbf{h}}) = \mathbb{E}\{J(\hat{\mathbf{h}}, n)\}$ and ρ is a regularization parameter.

4.3. Adaptive LMS Updating

The cost function in (9) is a combination of a convex differential term and an l_1-norm term, and we minimize this using a split Bregman approach [19] similar to [14]. Temporarily omitting the unit-norm constraint, (9) can be reformulated as:

$$(\hat{\mathbf{h}}, \hat{\mathbf{d}}) = \arg\min_{\mathbf{h},\mathbf{d}}\{\bar{J}(\mathbf{h}) + \rho|\mathbf{d}|_1 + \lambda\|\mathbf{d}-\mathbf{h}\|_2^2\}, \tag{10}$$

where $\mathbf{d}$ is a $(MK)\times 1$ auxiliary variable vector, with $\hat{\mathbf{d}}$ as the estimate, and λ is a Lagrange multiplier.

The problem in (10) can be further solved by the split Bregman iteration method [19] as

$$(\hat{\mathbf{h}}, \hat{\mathbf{d}})^{k+1} = \arg\min_{\mathbf{h},\mathbf{d}}\{\bar{J}(\mathbf{h}) + \rho|\mathbf{d}|_1 + \lambda\|\mathbf{b}^k + \mathbf{h} - \mathbf{d}\|_2^2\}, \tag{11a}$$

$$\mathbf{b}^{k+1} = \mathbf{b}^k + \hat{\mathbf{h}}^{k+1} - \hat{\mathbf{d}}^{k+1}, \tag{11b}$$

where $\mathbf{b}$ is a $(KM)\times 1$ Bregman variable vector, and k denotes the iteration index. Reinserting the unit norm constraint, (11a) can finally be transformed into two sub-problems [14]:

$$\hat{\mathbf{h}}^{k+1} = \arg\min_{\mathbf{h}}\{\bar{J}(\hat{\mathbf{h}}) + \lambda\|\mathbf{b}^k + \mathbf{h} - \hat{\mathbf{d}}^k\|_2^2\},\ \text{s.t. } \|\mathbf{h}\|_2^2 = 1 \tag{12a}$$

$$\hat{\mathbf{d}}^{k+1} = \arg\min_{\mathbf{d}}\{\rho|\mathbf{d}|_1 + \lambda\|\mathbf{b}^k + \hat{\mathbf{h}}^{k+1} - \mathbf{d}\|_2^2\}. \tag{12b}$$

Solving (12a): Analogous to an LMS update, $\hat{\mathbf{h}}$ is updated for each n and $\bar{J}(\hat{\mathbf{h}})$ is replaced by the instantaneous value $J(\hat{\mathbf{h}}^n, n)$. By using gradient descent, we obtain:

$$\begin{aligned}\hat{\mathbf{h}}^{n+1} &= \hat{\mathbf{h}}^n - \mu\frac{\partial J^n(\mathbf{h}, n)}{\partial\mathbf{h}} - \mu\lambda\frac{\partial\|\hat{\mathbf{d}}^n - \mathbf{h} - \mathbf{b}^n\|_2^2}{\partial\mathbf{h}} \\ &= \hat{\mathbf{h}}^n + 2\mu\frac{\mathbf{R}^n\hat{\mathbf{h}}^n + J^n(\hat{\mathbf{h}}^n, n)\hat{\mathbf{h}}^n}{\|\hat{\mathbf{h}}^n\|_2^2} \\ &\quad - 2\mu\lambda(\hat{\mathbf{h}}^n + \mathbf{b}^n - \hat{\mathbf{d}}^n),\end{aligned} \tag{13}$$

where μ is the step size. Then enforcing the unit-norm constraint, we finally have:

$$\hat{\mathbf{h}}^{n+1} = \frac{\hat{\mathbf{h}}^n + 2\mu[\mathbf{R}(n)\hat{\mathbf{h}}^n - \Upsilon(\hat{\mathbf{h}}^n, n)\hat{\mathbf{h}}^n - \lambda(\hat{\mathbf{h}}^n + \mathbf{b}^n - \hat{\mathbf{d}}^n)]}{\|\hat{\mathbf{h}}^n + 2\mu[\mathbf{R}(n)\hat{\mathbf{h}}^n - \Upsilon(\hat{\mathbf{h}}^n, n)\hat{\mathbf{h}}^n - \lambda(\hat{\mathbf{h}}^n + \mathbf{b}^n - \hat{\mathbf{d}}^n)\|_2^2}. \tag{14}$$

Solving (12b)*:* In order to control the computational complexity, $\hat{\mathbf{d}}$ is updated only once every P samples. Following [19], whenever $(n+1)\mod P = 0$, $\hat{\mathbf{d}}$ is updated as

$$\hat{\mathbf{d}}_i^{n+1} = \text{sign}(\hat{\mathbf{h}}_i^{n+1} + \hat{\mathbf{b}}_i^n)\cdot\max\{|\hat{\mathbf{h}}_i^{n+1} + \hat{\mathbf{b}}_i^n| - \frac{\rho}{2\lambda}, 0\}, \tag{15}$$

where $\mathbf{u}_i$ denotes the i-th element of the vector $\mathbf{u}$. Whenever $\hat{\mathbf{d}}$ is updated, $\mathbf{b}$ is also updated according to (11b).

In the LMS updating, the early RIR of the first channel is initialized to a $K \times 1$ vector $[\mathbf{0}^T_{\tau_{\max}+1}\ 1\ \mathbf{0}^T_{\tau_{\max}+1}]^T$, and the other early RIRs are initialized to the all-zeros $K \times 1$ vector. It is worth noting that for each channel, the number of estimated coefficients is only K for the proposed method where K is determined by the microphone spacing and is always much smaller than the length of identified RIR in the traditional BSI based methods. Thus both the computational complexity and the estimation latency are reduced.

4.4. Time Delay Calculation

Once the early RIRs are identified, the time delays can be computed according to the locations of the direct paths. We should note that practically the direct path does not necessarily appear at an integer sampling point due to the finite sampling rate. Accordingly, the direct path is identified as the maximum peak of the quadratic interpolation of each identified RIR, using the v_findpeaks function in the VOICEBOX [20]. Given the direct path locations, the time delays are computed according to (2).

5. EVALUATION

We evaluate the performance of our proposed method against two alternative TDE methods on simulated data. The SRP-PHAT (or GCC-PHAT in the two-channel case) and a recently proposed frequency-domain BSI based method [14], are used for comparison.

A rectangular room is modelled with size 5 m $\times$ 4 m $\times$ 3 m. A two-element microphone array is deployed with the elements located at (2.4, 2, 1.6) m and (2.6, 2, 1.6) m respectively. The source is positioned at (1.05, 3.95, 1.67) m. A speech signal with sampling rate of 8 kHz and approximately 18 s duration is used as the source signal, which is generated by concatenating six sentences randomly selected from the IEEE sentences database [21]. The microphone signals are obtained by first convolving the source signal with the RIRs simulated by the image-source method [22], and then adding uncorrelated white Gaussian noise to achieve a 40 dB signal-to-noise ratio. The length of the simulated RIRs is set as $L = RT_{60} f_s$, where RT_{60} is the reverberation time [23].

Throughout the experiments, the parameters of the proposed method are empirically chosen as: $\mu = 0.8$, $\rho = 4 \times 10^{-4}$, $\lambda = 5 \times 10^{-3}$, $P = 30$. The analysis window for STFT in the pre-filtering step is a 512 sample Hamming window with 50% overlap. The method in [14] uses the traditional CR error based criterion for BSI, and requires oracle knowledge of the assumed filter length. For this method, we test two variations, which assume the true filter length to be $\hat{L} = 2048$ and $\hat{L} = 1024$, respectively. The same pre-filtering procedure with the proposed method is also performed. For the GCC-PHAT method, the analysis window is 512 samples hamming window with 50% overlap, and the smoothing factor for cross power spectrum estimation is 0.95.

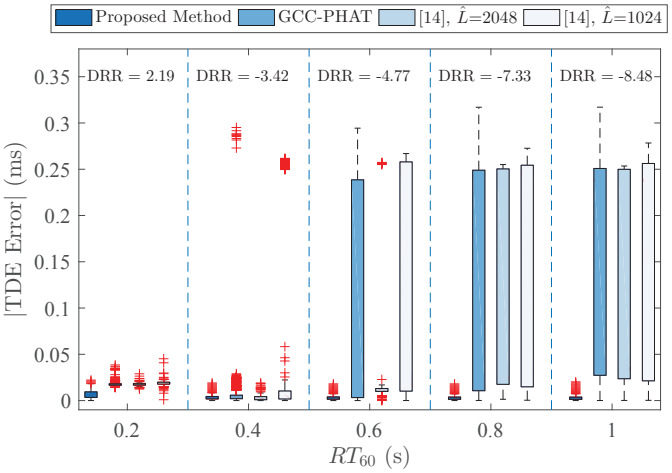

Fig. 2. Box plots of absolute TDE errors as a function of RT_{60}. Whiskers are 1.5 times the interquartile range. 1 sample time delay corresponds to 0.125 ms duration. DRRs are shown in dB.

The algorithms are tested in simulated rooms with five reverberation times in the range 0.2 to 1 s. For each method, 10 Monte Carlo simulations were conducted, and TDE estimates were performed whenever speech was active but excluding the initial 256 ms of speech signal to allow the BSI-based methods to converge. The box plots of the absolute TDE errors of different algorithms are compared in Fig. 2, and the direct-to-reverberant ratio (DRR) in dB for each senario is also shown. We can see that although the performances of comparison methods tend to degrade when RT_{60} increases, the proposed method can still achieve an accurate estimation even when $RT_{60} = 1$ s. On the other hand, when assuming $\hat{L} = 2048$, the method in [14] can generally yield better performance than the GCC-PHAT algorithm, but the TDE performance drops clearly when either increasing RT_{60} or decreasing the assumed true RIR length to $\hat{L} = 1024$. This is mainly because in both cases, with the assumed true filter length, the RIRs cannot fully describe the true system, and the traditional BSI cannot work reliably in under-modelled situations.

6. CONCLUSION

In this paper, we proposed a new method for TDE in reverberant environments. The proposed method is based on under-modelled BSI which maximizes the cross-correlation of the cross-filtered signals and exploits the sparsity of the early RIR. The new method can be viewed as a generalization of the traditional cross-correlation based methods to consider a more realistic model for the early RIR. The time delays are finally estimated based on the early RIR estimates. By conducting experiments in different reverberant conditions, we demonstrate that the proposed method outperforms GCC-PHAT and the cross-relation error BSI based TDE methods.

7. REFERENCES

[1] S. Wangsiripitak and D. W. Murray, "Avoiding moving outliers in visual SLAM by tracking moving objects," in *Proc. Intl. Conf. Robotics and Automation*, May 2009, pp. 375–380.

[2] C. Evers, A. H. Moore, and P. A. Naylor, "Acoustic simultaneous localization and mapping (a-SLAM) of a moving microphone array and its surrounding speakers," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016.

[3] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[4] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[5] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4925–4935, Sep. 2010.

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

[7] 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)*, Shanghai, China, Mar. 2016.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[9] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[10] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, 2003.

[11] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sep. 2004.

[12] Y. Huang and J. Benesty, "Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization," in *Adaptive Signal Processing*. Springer Berlin Heidelberg, 2003, pp. 227–247.

[13] Y. Lin, J. Chen, Y. Kim, and D. Lee, "Blind Sparse-Nonnegative (BSN) channel identification for acoustic Time-Difference-of-Arrival estimation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2007, pp. 106–109.

[14] K. Kowalczyk, E. A. P. Habets, W. Kellermann, and P. A. Naylor, "Blind system identification using sparse learning for TDOA estimation of room reflections," *IEEE Signal Process. Lett.*, vol. 20, pp. 653–656, 2013.

[15] M. A. Haque, M. S. A. Bashar, P. A. Naylor, K. Hirose, and M. K. Hasan, "Energy constrained frequency-domain normalized LMS algorithm for blind channel identification," *Signal, Image and Video Processing*, vol. 1, no. 3, pp. 203–213, Apr. 2007.

[16] W. Xue, M. Brookes, and P. A. Naylor, "Cross-correlation based under-modelled multichannel blind acoustic system identification with sparsity regularization," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, Aug. 2016.

[17] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2008, pp. 2565–2568.

[18] L. R. Rabiner and R. W. Schafer, Eds., *Theory and Applications of Digital Signal Processing*. Pearson, 2010.

[19] T. Goldstein and S. Osher, "The split Bregman method for L1-regularized problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 323–343, Jan. 2009.

[20] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997–2016.

[21] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio and Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.

[22] 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.

[23] H. Kuttruff, *Room Acoustics*, 4th ed. London: Taylor & Francis, 2000.