BLIND SYSTEM IDENTIFICATION FOR SPEECH DEREVERBERATION WITH FORCED SPECTRAL DIVERSITY

Xiang (Shawn) Lin,[1] *Andy W. H. Khong,*[2] *and Patrick A. Naylor*[1]

[1]Department of Electrical and Electronic Engineering, Imperial College London, UK
[2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ABSTRACT

The common zeros problem for blind system identification (BSI) is well known. It degrades the performance of classic BSI algorithms and therefore imposes the limit on the performance of subsequent speech dereverberation. The effect of near-common zeros has recently been studied in terms of channel diversity and the degradation in performance of BSI and multichannel equalization algorithms has been shown. We now introduce a novel approach to improve channel diversity which we refer to as Forced Spectral Diversity (FSD). The FSD concept uses a combination of spectral shaping filters and effective channel undermodelling. Simulation results show that the proposed approach achieves improved performance with reduced complexity for multichannel BSI in a room acoustics example.

***Index Terms*—** blind system identification, speech dereverberation, near-common zeros, channel diversity

1. INTRODUCTION

Speech acquisition in rooms with microphones positioned at a distance from the talker suffers degradation in quality due to reverberation. This is caused by multiple reflections of the sound from surrounding walls and objects [1]. Speech reverberation is therefore an important problem in, for example, hands-free telecommunication applications. One approach to this problem is to perform multichannel blind system identification (BSI) to estimate the room impulse responses and to recover an estimate of the original speech through multichannel equalization [2].

One of the identifiability conditions for most second order statistics (SOS)-based BSI algorithms is that the channels must be coprime, i.e., they do not share common zeros [3]. Otherwise, BSI algorithms fail to identify the channels correctly as they cannot distinguish whether the common terms are due to the input signal or the acoustic channels. Although increasing spatial diversity with more microphones is effective in reducing the number and negative impact of common (or near-common) zeros, it is computationally expensive and practically limited. In addition, the coprime property is also required for equalization of a SIMO system using the Bezout theorem [2][4]. However, it has been shown that the presence of near-common zeros (NCZs) degrades the performance of both BSI and equalization algorithms and that the effect can be quantified in terms of channel diversity [5]. For multichannel systems with high order, the problem of NCZs can be very significant since zeros of channel responses tend to cluster around the unit circle [6]. The NCZs problem in speech dereverberation has not been specifically addressed so far in the literature.

In this paper, we propose a novel method to overcome the NCZs problem for speech dereverberation based on the concept of forced spectral diversity (FSD), which combines the use of spectral shaping filters and effective channel undermodelling [7]. We first show how NCZs affect the dereverberation performance in Section 3. The FSD concept is proposed in Section 4 with illustrative examples. In Section 5, we apply this concept to blind identification of SIMO acoustic systems for speech dereverberation. Simulation results of BSI and dereverberation incorporating FSD is presented in Section 6.

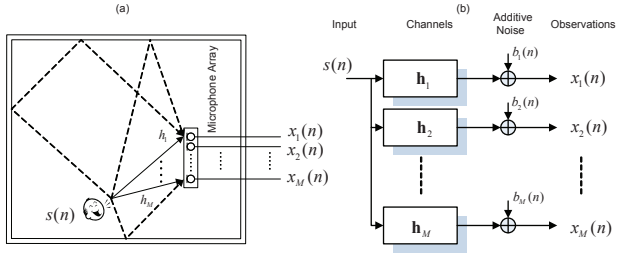

Fig. 1. Diagram of (a) M-channel SIMO acoustic system and (b) the problem of BSI.

2. PROBLEM FORMULATION

A speech signal $s(n)$ in a reverberant room propagates from its source to an array of M microphones as shown in Fig. 1(a). This can be modelled by a SIMO system where the observed signal at the mth microphone, as shown in Fig. 1(b), is given by

$$x_m(n) = \mathbf{h}_m^T \mathbf{s}(n) + b_m(n), \quad m = 1, \ldots, M \tag{1}$$

where $\mathbf{h}_m = [h_{m,0} \ \ldots \ h_{m,L-1}]^T$ is the L-tap impulse response between the source and the mth microphone, $\mathbf{s}(n) = [s(n) \ \ldots \ s(n-L+1)]^T$ is the input signal vector, $b_m(n)$ is the additive noise and $[\cdot]^T$ denotes the vector transpose operator. The aim of BSI is to estimate blindly the impulse responses $\mathbf{h}_m$ from the observations $x_m(n)$. Among various BSI algorithms, SOS-based algorithms have become popular [8]. A typical approach is to utilize the cross-relation (CR) between two channels [3], i.e., $\mathbf{x}_i^T(n)\mathbf{h}_j = \mathbf{x}_j^T(n)\mathbf{h}_i$ for $i, j = 1, 2, \ldots, M$, $i \neq j$, where $\mathbf{x}_m(n) = [x_m(n) \ \ldots \ x_m(n-L+1)]^T$. In the presence of noise, a cost function can be obtained by considering all combinations of M channels, i.e.,

$$\mathbf{R}\mathbf{h} = \mathbf{e}, \tag{2}$$

where $\mathbf{h} = [\mathbf{h}_1^T \ \ldots \ \mathbf{h}_M^T]^T$ is a vector of concatenated channel responses and $\mathbf{R}$ is a correlation-like matrix [3]. The estimated channel responses $\widehat{\mathbf{h}}$ can then be found, up to a scaling factor, by minimizing (2) using adaptive or closed-form algorithms such as [9][10].

For the subsequent dereverberation of speech, a second stage is required where a system of inverse filters can be obtained using multichannel equalization algorithms, such as multichannel inverse theorem (MINT) [4].

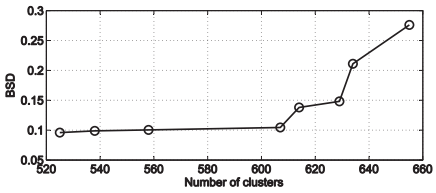

Fig. 2. BSD vs. number of clusters for simulated impulse responses.

3. EFFECT OF NEAR-COMMON ZEROS ON DEREVERBERATION

Defining $H_m(z) = K\prod_{p=1}^{L-1}(z - z_m(p))$ as the z-transform of the mth channel impulse response where $z_m(p)$ denotes the pth zero and K is the gain constant, NCZs are clusters of zeros that satisfy the following conditions [11]: (i) each cluster contains only M zeros with each channel contributing one zero, and (ii) the Euclidean distance between any pair of zeros in a cluster lies within a pairwise tolerance ξ where $\xi \geq 0$. We note that given these two conditions, any zero can be included in more than one cluster, and NCZs become exactly-common zeros when $\xi = 0$.

Since the presence of NCZs degrades the performance of BSI and equalization algorithms, we show how they can affect the overall performance of dereverberation. We simulated a set of impulse responses using the method of images [12] with room dimensions $10 \times 10 \times 3$ m. The sampling frequency and reverberation time are 8 kHz and $T_{60} = 0.2$ s, and the generated impulse responses were truncated to 1024 coefficients. A speech sample containing both male and female utterance was used and captured by a linear microphone array with uniform spacing of 5 cm. Employing normalized multichannel frequency domain least-mean-squares (NMCFLMS) [9] and MINT [4], the performance of dereverberation is measured using Bark spectral distortion (BSD) [13] and shown in Fig. 2 for a set of two-channel systems extracted from the generated impulse responses, where the step-size was set 0.2, the signal-to-noise ratio (SNR) was 60 dB, and the number of NCZ clusters were found using the GMC-ST algorithm [11] with $\xi = 2 \times 10^{-3}$. It can be clearly seen that BSD value increases with number of clusters indicating the degradation of the dereverberation performance.

4. THE CONCEPT OF FSD

Since increasing spatial diversity through the use of larger microphone arrays is limited practically, the motivation underlying FSD is to introduce spectral shaping filters in combination with undermodelling in order to increase channel diversity. We introduce the FSD concept by describing how these two processing steps are combined using illustrative examples and numerical results.

4.1. Illustrative examples

Channel undermodelling was introduced for BSI in [7], where it was shown that the L_mth-order least-squares (LS) method [3] can estimate the L_mth-order "significant" part of the full L-order impulse responses for $L_m < L$ provided that it offers sufficient diversity. For $\mathbf{h}_m$ defined in (1), the L_mth-order "significant" part can be found by $\mathbf{h}_m^{L_m} = \mathbf{h}_m - \mathbf{d}_m^{L_m}$, where $\mathbf{h}_m^{L_m} = [h_{m,0} \ \ldots \ h_{m,L_m-1} \ 0 \ \ldots \ 0]^T$ and $\mathbf{d}_m^{L_m} = [0 \ \ldots \ 0 \ h_{m,L_m} \ \ldots \ h_{m,L-1}]^T$. Similarly, define in z-domain $A(z) = z - z_A$, $B(z) = z - z_B$ and

$$C(z) \ = \ A(z)B(z) \ = \ z^2 - (z_A + z_B)z + z_A z_B, \qquad (3)$$

where $z_A, \ z_B \neq 0$, the 1st-order part of $C(z)$ is thus given by,

$$C'(z) \ = \ z - (z_A + z_B) \ = \ z - z_{C'}. \qquad (4)$$

We refer to such procedure as the channel undermodelling.

Utilizing $A(z)$, $B(z)$ and $C(z)$, we consider a SIMO system with two identical channels, $H_1(z) = H_2(z) = A(z)$, which produces the output signals, $X_m(z) = S(z)H_m(z), \ m = 1, 2$, where $S(z)$ is the z-transform of the source signal. BSI algorithms are not expected to work successfully as $H_m(z)$ is equivalently a single channel system. However, if we introduce $B(z)$ to obtain

$$\overline{X}_1(z) \ = \ X_1(z)B(z) \ = \ S(z)H_1(z)B(z) \ = \ S(z)C(z), \quad (5)$$

and employ BSI with order $L = 2$ on $\overline{X}_1(z)$ and $X_2(z)$, it is equivalent to identifying a modified system $H'_m(z)$ of order $L = 2$ with $H'_2(z) = H_2(z)$ and $H'_1(z) = C'(z)$. Since it is found that the distance between zeros has been increased from $|z_A - z_A| = 0$ to $|z_{C'} - z_A| = |z_B| > 0$, $H'_m(z)$ now contains no exactly-common zeros, and thus the BSI algorithm is able to successfully estimate both $H_2(z)$ and $H'_1(z)$. This indicates that extra diversity, quantified by $|z_B|$, is introduced into $H_m(z)$ using linear convolution and channel undermodelling. Therefore, as long as $|z_B|$ is sufficiently large such that the first condition for NCZs to exist, as described in Section 3, is not satisfied, $H'_m(z)$ can be identified.

In some cases [7], undermodelling is sufficient without extra zeros for increasing channel diversity. Consider another example system with $H_1(z) = A(z)B(z)$ and $H_2(z) = B(z)C'(z)$ where z_B is the common zero, repeating similar procedures in (4) for both channels results in a modified system with zeros of each channel being $z'_1 = z_A + z_B$ and $z'_2 = z_B + z_{C'}$, respectively, from which it is found that the common zero can be eliminated without introducing extra zeros since $|z'_2 - z'_1| = |z_B|$. However, extra zeros obtained from *spectral shaping filters* have been found necessary because the zeros of acoustic systems tend to cluster around the unit circle; the effect of the extra zeros in combination with undermodelling is that the zeros in the modified system are located differently to those in the original system, adding additional channel diversity. For a M-channel system, up to M spectral shaping filters can be employed.

4.2. Numerical results

We now present numerical results to further illustrate how FSD processing affects simple SIMO systems with common zeros, from which important characteristics of the FSD processing can be summarized. Let $\mathbf{h}$ denote a two-channel SIMO system with real coefficients of length $L = 7$ and $\mathbf{h}'$ as the modified system after FSD processing, respectively. The zeros of each channel are defined as $\mathbf{z}_1$ and $\mathbf{z}_2$, and $\mathbf{z}_1 = \mathbf{z}_2$. Since these zeros exist as complex conjugate pairs, Fig. 3(a) shows $\mathbf{z}_1$ and $\mathbf{z}_2$ on the upper half of the z-plane, respectively denoted by circles and squares. A set of spectral shaping filters with two zeros defined as $\mathbf{z}_{f,m} = [z_{f,m}(1) \ z_{f,m}(2)], \ m = 1, 2$ are then randomly generated such that $z_{f,m}(1) = z^*_{f,m}(2)$. In order to show how the effect of FSD processing varies with $\mathbf{z}_{f,1}$ and $\mathbf{z}_{f,2}$, the modulus $\rho_1 = |z_{f,1}(1)| = |z_{f,1}(2)|$ was set to range from 0.05 to 1, and Fig. 3(b) shows $\mathbf{z}_{f,1}$ and $\mathbf{z}_{f,2}$ where $\rho_1 = 1$. Denote the zeros of $\mathbf{h}'$ as $\mathbf{z}'_m = [z'_m(1) \ z'_m(2)], \ m = 1, 2$, Fig. 3(c) and Fig. 3(d) show $\mathbf{z}'_m$ for the example cases $\rho_1 = 0.45$ and 0.7. Comparing $\mathbf{z}'_1$ with $\mathbf{z}_1$, it is seen that by introducing $\mathbf{z}_{f,1}$ with increasing ρ_1, $\mathbf{z}'_1$ is located further away from $\mathbf{z}_2$ indicating the increment in channel diversity of $\mathbf{h}'$ compared to $\mathbf{h}$.

We compare various measurements over $\mathbf{h}'$ against different ρ_1. In Fig. 4(a), the mean distance between $\mathbf{z}'_1$ and $\mathbf{z}'_2$ is plotted against ρ_1. As can be seen, the channel diversity for $\mathbf{h}'$, quantified by $|\mathbf{z}'_1 -$

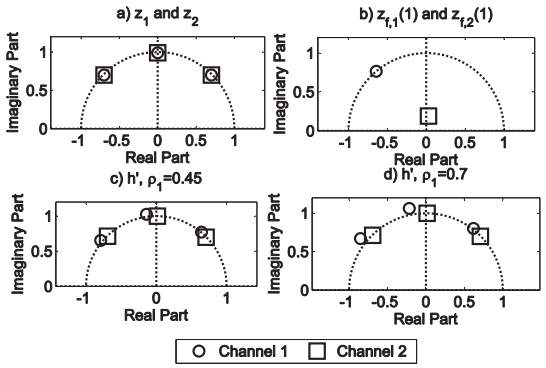

Fig. 3. Effect of FSD: a) $\mathbf{z}_1$ and $\mathbf{z}_2$; b) $\mathbf{z}_{f,1}$ and $\mathbf{z}_{f,2}$ with $\rho_1 = 1$ for $\mathbf{z}_{f,1}$; c) $\mathbf{z}'_1$ and $\mathbf{z}'_2$ for $\rho_1 = 0.45$; d) $\mathbf{z}'_1$ and $\mathbf{z}'_2$ for $\rho_1 = 0.7$.

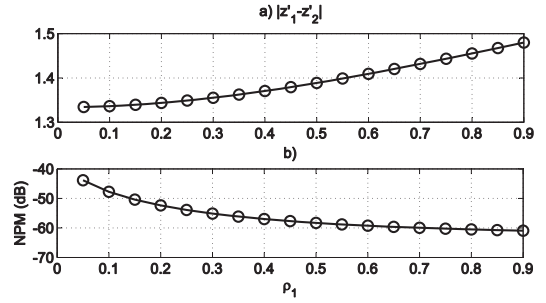

Fig. 4. Various results on $\mathbf{h}'$ for the effect of FSD filtering against various ρ_1: a) $|\mathbf{z}'_1 - \mathbf{z}'_2|$; b) NPM performance for WGN input.

$\mathbf{z}'_2|$, increases due to the increasing ρ_1 for the extra zeros of spectral shaping filters. We then employ white Gaussian noise (WGN) as the source signal with 50 dB SNR to simulate the BSI performance over $\mathbf{h}'$ using the subspace algorithm [10]. As shown in Fig. 4(b), the improvement of BSI performance measured using NPM [14] due to increased diversity is clearly seen, which is expected to result in a better performance for channel equalization.

In summary, the FSD processing involves two important components: (i) spectral shaping filters provide extra zeros, and (ii) effective undermodelling of the system with these extra zeros gives rise to additional diversity compared to the original system. It is now important to note that an estimate of $\widehat{\mathbf{s}}(n)$ can be obtained without the need to "undo" the effect of FSD processing by equalizing $\widehat{\mathbf{h}}'_m$ using $\widehat{\mathbf{g}}'_m = \min_{\mathbf{g}'_m} \|\widehat{\mathbf{H}}'^T_m \mathbf{g}'_m - \delta(n-\tau)\|^2$ where τ is the modelling delay, $\widehat{\mathbf{H}}'_m$ is the convolutive matrix of $\widehat{\mathbf{h}}'_m$ and $\|\cdot\|$ denotes l_2-norm, so that the FSD concept can be directly applied to dereverberation, although the accuracy of the inversion process can be limited due to noise amplification if spectral shaping filters have zeros close to the unit circle, which then indicates a potential tradeoff between increasing diversity using the FSD and the equalization performance.

5. FSD PROCESSING FOR SIMO ACOUSTIC SYSTEMS

We now apply the FSD concept to SIMO system for dereverberation. For simplicity, we assume the system to be noise-free. A two-channel system diagram is shown in Fig. 5, where the microphone signal $\mathbf{x}_m(n)$ is filtered by the spectral shaping filters and the resulting outputs $\overline{\mathbf{x}}_m(n)$ can be written

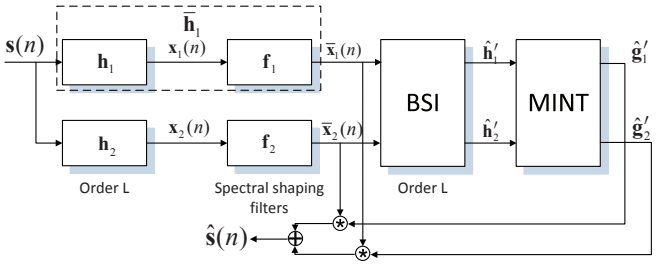

Fig. 5. Schematic for a two-channel SIMO system with FSD processing for speech dereverberation.

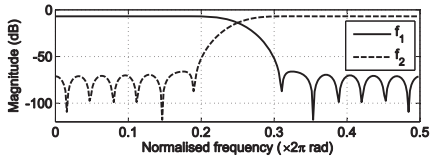

Fig. 6. Frequency responses of the spectral shaping filters, where the solid line denotes $\mathbf{f}_1$ and the dotted line denotes $\mathbf{f}_2$.

$$\overline{\mathbf{x}}_m(n) = \boldsymbol{\mathcal{F}}_m^T \mathbf{H}_m \mathbf{s}(n), \qquad m = 1, 2 \tag{6}$$

where $\boldsymbol{\mathcal{F}}_m$ denotes the convolutive matrix associated with the impulse responses of the mth spectral shaping filter of length L_p. $\overline{\mathbf{x}}_m(n)$ can thus be considered as the linear convolution between $\mathbf{s}(n)$ and a SIMO system of length $L + L_p - 1$, given by,

$$\overline{\mathbf{h}}_m \;=\; \boldsymbol{\mathcal{F}}_m^T \mathbf{h}_m, \qquad m = 1, 2. \tag{7}$$

The effective undermodelling is then implemented by employing BSI with order L. This is reasonable since in practice the BSI algorithms are equivalently "blind" to the existence of spectral shaping filter. The modified system to be identified is now given by

$$\mathbf{h}'_m \;=\; \mathbf{U}\overline{\mathbf{h}}_m, \qquad m = 1, 2 \tag{8}$$

where $\mathbf{U} = [\mathbf{I}_{L\times L}\ \mathbf{0}_{L\times(L_p-1)}]$ with $\mathbf{I}_{L\times L}$ and $\mathbf{0}_{L\times(L_p-1)}$ being an identity matrix and a null matrix, respectively.

As indicated in Section 4, the design of $\boldsymbol{\mathcal{F}}_m$ is not trivial as they need to offer sufficient diversity, which is correlated with the distribution of zeros in the original system. Utilizing the characteristic that for long impulse responses, zeros cluster around the unit circle with an approximately uniform distribution, we propose to employ a typical pair of highpass and lowpass FIR filter as spectral shaping filters for the FSD filtering. Such choice increases the FSD effect since the zeros of the filters locate complementarily in the z-plane, so that in the modified systems, the zeros can be "relocated" towards opposite directions in the z-plane. It is also noted that the BSI algorithm is always employed with the order of L regardless of $\boldsymbol{\mathcal{F}}_m$, indicating a $L_p - 1$ undermodelling factor.

6. SIMULATIONS

We demonstrate the performance improvement brought about by FSD using a set of measured acoustic impulse responses obtained from the MARDY database [15] which were resampled at 8 kHz and truncated to 512 taps. A pair of highpass and lowpass FIR shaping filters of length $L_p = 32$ was generated with magnitude responses

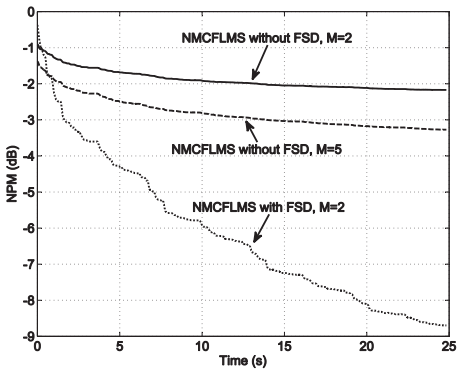

Fig. 7. Comparison of BSI performance for SIMO system with and without FSD processing using speech input.

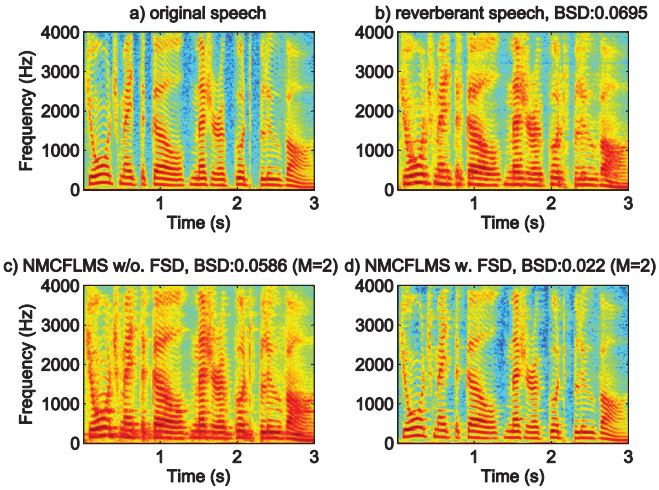

Fig. 8. Spectrogram of a) clean speech; b) reverberant speech; dereverberated speech without FSD; d) dereverberated speech with FSD.

shown in Fig. 6. The speech samples described in Section 3 were employed.

Figure 7 shows the BSI performance in NPM for NMCFLMS algorithm using step-size 0.1 with FSD processing as for $M = 2$. This is compared with the case without FSD processing for $M = 2$. A 6.5 dB improvement in NPM for the FSD processed system is seen over the original system for the case $M = 2$. We further show the performance for the case of $M = 5$ representing greater spatial diversity without FSD processing and it is noted that the two-channel FSD processing still achieves about 6 dB gain of NPM. This indicates that the use of FSD can result in improved performance without the need of larger microphone arrays.

Using the improved accuracy of the FSD channel estimate, the MINT [4] algorithm was then employed to produce corresponding inverse filters. The spectrograms of various speech signals are shown in Fig. 8 for the first 3 seconds for clarity of presentation. Since the FSD processing results in a modified system with fewer NCZs, corresponding improvement for dereverberation is seen from comparing Fig. 8(c) with Fig. 8(d) in both low and high frequencies. The BSD values shown in the figure further supports such improvement of FSD processing for the overall dereverberation performance.

7. CONCLUSIONS

We have introduced the concept of FSD to mitigate the NCZs problem in blind system identification and subsequent dereverberation. We have shown how the performance of BSI is affected by NCZs, and how effective undermodelling can be combined with spectral shaping filters to generate a modified system with sufficient diversity to enable more accurate system identification. It is noted that inversion of the modified system is sufficient for recovery of the original source signal. Simulation results based on real acoustic impulse responses confirmed the effectiveness of the proposed concept for improved BSI and subsequent dereverberation of speech.

8. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," in *Proc. Int. Workshop Acoust. Echo and Noise Control (IWAENC)*, Sept. 2005.

[2] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, pp. 882–895, Sept. 2005.

[3] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[4] 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.

[5] P. A. Naylor, X. Lin, and A. W. H. Khong, "Near-common zeros in blind identification of SIMO acoustic systems," in *Proc. Joint Workshop on Hands-Free Speech Comm. and Microphone arrays (HSCMA)*, Trento, Italy, May 2008, pp. 21–24.

[6] X. Lin, N. D. Gaubitch, and P. A. Naylor, "Two-stage blind identification of SIMO systems with common zeros," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Florence, Italy, Sept. 2006.

[7] A. P. Liavas, P. A. Regalia, and J.-P. Delmas, "Robustness of least-squares and subspace methods for blind channel identification/equalization with respect to effective channel undermodeling/overmodeling," *IEEE Trans. Signal Process.*, vol. 47, no. 6, pp. 1636–1645, June 1999.

[8] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind system identification," *Proc. IEEE*, vol. 85, no. 8, pp. 1310–1322, Aug. 1997.

[9] 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.

[10] S. Gannot and M. Moonen, "Subspace methods for multi-microphone speech dereverberation," *EURASIP J. Applied Signal Process.*, vol. 2003, no. 11, pp. 1074–1090, Oct. 2003.

[11] A. W. H. Khong, X. Lin, and P. A. Naylor, "Algorithms for identifying clusters of near-common zeros in multichannel blind system identification and equalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, USA, Apr. 2008, pp. 389–392.

[12] 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, April 1979.

[13] S. L. Gay and J. Benesty, Eds., *Acoustic Signal Processing For Telecommunicatios*. Kluwer Academic Publishers, 2000.

[14] D. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal Process. Letters*, vol. 5, no. 7, pp. 174–176, July 1998.

[15] J. Y. C. Wen, N. D. Gaubitch, E. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Int. Workshop Acoust. Echo and Noise Control (IWAENC)*, Paris, France, Sept. 2006.