# BLIND SPEECH DEREVERBERATION IN THE PRESENCE OF COMMON ACOUSTICAL ZEROS

*Xiang (Shawn) Lin, Nikolay D. Gaubitch and Patrick A. Naylor*

Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, UK
E-mail: {shawn.lin04, ndg, p.naylor}@imperial.ac.uk

## ABSTRACT

*Speech acquisition in rooms with microphones positioned at a distance from the talker suffers degradation in quality due to reverberation caused by multiple reflections of the sound from surrounding walls and objects. Speech reverberation is therefore an important problem in, for example, hands-free telecommunication applications. One approach to this problem is to perform dereverberation using blind multichannel system identification and inversion. However, most such methods rely on the assumption that the room transfer functions (RTFs) do not share common zeros. In this paper, a two-stage approach is proposed which separately identifies and equalizes the common and the non-common zeros components of the RTFs. Experimental results indicate a considerable improvement using the new method compared to other existing methods.*

## 1. INTRODUCTION

Reverberation arises when acoustic signals are emitted in non-anechoic environments. Reverberation may reduce considerably the intelligibility and speech recognizer performance in hands-free telecommunication systems. The perceptual effects of room acoustics are often considered to comprise two distinct properties: the coloration caused by the early reflections and the reverberation caused by the reverberant tail of the room impulse response.

The aim of blind speech dereverberation is to recover the clean speech using only the observed microphone signals. It is a blind problem since neither the acoustic system nor the source signal are available. Also, typical room impulse responses may contain several thousands of taps, making the recovery more challenging. Existing speech dereverberation algorithms can be generally divided into three categories: (i) Blind System Identification and Inversion (BSII) algorithms [1, 2] which blindly estimate and equalize the room impulse responses to recover the source signal, (ii) algorithms based on the speech enhancement [3] using, for example, LPC method to modify the characteristics of reverberant speech signal without estimating the room impulse responses, and (iii) beamforming [4]. The method proposed in this paper belongs to category (i).

Most multichannel BSII-based dereverberation algorithms rely on the assumption that there are no zeros common to all RTFs [1, 2, 3]. However, it was demonstrated in [5] that it is very likely for multichannel acoustic systems with impulse responses of thousands of taps to have common zeros. Such systems are thus vulnerable to zeros that are close enough to degrade the performance of Blind System Identification (BSI) algorithms.

In [5], a two-stage method for blind identification of SIMO systems with common zeros was proposed based on the concept of channel decomposition, where the common zeros and non-common zeros were identified separately. In this paper, we apply this concept to blind speech dereverberation. We demonstrate the importance of correct estimation of the order of characteristic zeros and show that this can be found through the eigenvalues of the data correlation matrix. Furthermore, we employ an efficient way to implement the common zeros identification for real speech signals and measured room impulse responses.

The remainder of this paper is organized as follows. In Section 2, the problem of blind acoustic system identification is formulated. In Section 3, it is demonstrated that correct estimation of the number of characteristic zeros is essential to the overall system identification performance. The proposed two-stage algorithm is described in Section 4. Section 5 demonstrates the performance of the new algorithm with several illustrative simulation results and conclusions are drawn in Section 6.

## 2. PROBLEM FORMULATION

A typical acoustic environment containing one talker and multiple microphones can be considered a linear SIMO system where the relationship between the speech signal $s(n)$ and $m^{\text{th}}$ output $x_m(n)$ is given by

$$x_m(n) = \sum_{i=0}^{L-1} h_{m,i}(n)s(n-i) + b_m(n), \qquad m = 1,2,\ldots,M \tag{1}$$

where $h_{m,i}(n),\ i = 0,\ldots,L-1$ are the coefficients for the $m^{\text{th}}$ channel with $L$ taps and $b_m(n)$ is additive noise. As we are concentrating on the dereverberation problem, henceforth we consider the noise-free case in this paper, i.e., $b_m(n) = 0,\ m = 1,2,\ldots,M$. Equation (1) can now be written in vector form,

$$\mathbf{x}_m(n) = \mathbf{H}_m(n)\mathbf{s}(n), \qquad m = 1,2,\ldots,M \tag{2}$$

where

$$\mathbf{H}_m(n) = \begin{bmatrix} \mathbf{h}_m^T(n) & 0 & \cdots & 0 \\ 0 & \mathbf{h}_m^T(n) & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \mathbf{h}_m^T(n) \end{bmatrix}_{L\times(2L-1)},$$

$$\mathbf{h}_m(n) = [h_{m,0}(n),\ h_{m,1}(n),\ \ldots,\ h_{m,L-1}(n)]^T,$$
$$\mathbf{x}_m(n) = [x_m(n),\ x_m(n-1),\ \ldots,\ x_m(n-L+1)]^T,$$
$$\mathbf{s}(n) = [s(n),\ s(n-1),\ \ldots,\ s(n-2L+2)]^T,$$

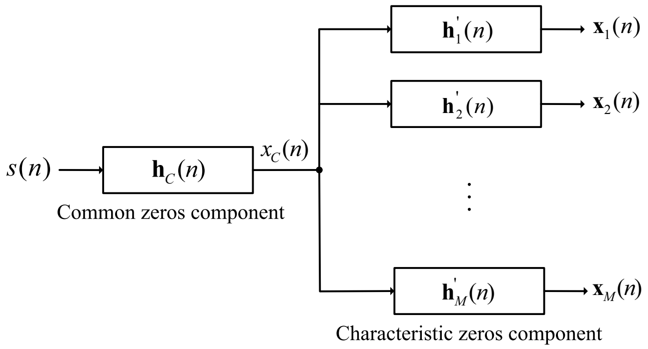

Figure 1: Channel decomposition.

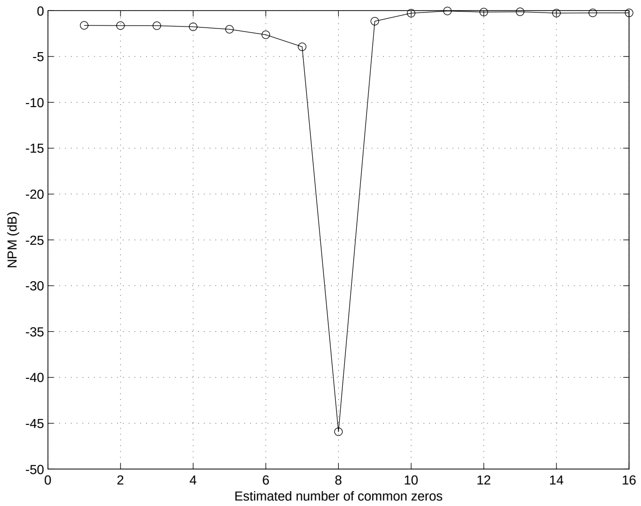

Figure 2: NPM vs. Estimated number of common zeros.

and $[\cdot]^T$ is the matrix transpose operator. Thus, the problem is to find an estimate $\hat{\mathbf{h}}_m(n) = [\hat{h}_{m,0}(n),\ \hat{h}_{m,1}(n),\ \ldots,\ \hat{h}_{m,L-1}(n)]^T$ given only $\mathbf{x}_m(n)$ such that an estimate $\hat{\mathbf{s}}(n)$ of $\mathbf{s}(n)$ can be formed.

## 3. EFFECTS OF COMMON ZEROS

It was shown in [6] that a multichannel system is identifiable if the following conditions are satisfied: (i) the autocorrelation matrix of input signal is full-rank and (ii) the multiple channels do not share common zeros. It was also shown in [5, 7] that zeros which are very close but not exactly common (i.e., near common zeros) degrade the performance of BSI algorithms. A SIMO system with common zeros can be considered to contain two parts: one with common zeros, $\mathbf{h}_C(n) = [h_{C,0}(n),\ h_{C,1}(n),\ \ldots,\ h_{C,L_C-1}(n)]^T$, and one with characteristic (non-common) zeros, $\mathbf{h}_m'(n) = [h_{m,0}'(n),\ h_{m,1}'(n),\ \ldots,\ h_{m,L'-1}'(n)]^T$, i.e.,

$$h_{m,i}(n) = \sum_{k=0}^{L'-1} h_{m,k}'(n) h_{C,i-k}(n), \qquad m = 1,\ 2,\ \ldots,\ M \quad (3)$$

where $h_{C,i}(n),\ i = 0,\ldots,L_C-1$ are the coefficients of $\mathbf{h}_C(n)$ and $h_{m,k}'(n),\ k = 0,\ldots,L'-1$ are the coefficients of $\mathbf{h}_m'(n)$, respectively. Therefore, the number of common zeros is equal to $L_C - 1$. This is illustrated in Fig. 1, where $x_C(n)$ is the output of the common zeros component, i.e., $x_C(n) = \sum_{i=0}^{L_C-1} h_{C,i}(n)s(n-i)$. An experiment was performed to demonstrate that only when the number of common zeros is correctly estimated can we blindly identify the system part with non-common zeros. As an example, the subspace algorithm [2] was employed with a randomly generated SIMO system containing two FIR channels of length $L = 32$ that share $L_C = 8$ common zeros. The identification performance was measured using the Normalized Projection Misalignment (NPM) defined as [8]

$$\mathrm{NPM}(n) = 20\log_{10}\left(\frac{1}{\|\mathbf{h}(n)\|}\left\|\mathbf{h}(n) - \kappa(n)\hat{\mathbf{h}}(n)\right\|\right)\mathrm{dB}, \quad (4)$$

where

$$\kappa(n) = \frac{\mathbf{h}^T(n)\hat{\mathbf{h}}(n)}{\hat{\mathbf{h}}^T(n)\hat{\mathbf{h}}(n)},$$

and $\mathbf{h}(n) = [\mathbf{h}_1^T(n),\ \mathbf{h}_2^T(n),\ \ldots,\ \mathbf{h}_M^T(n)]^T$ is the true channel vector and $\hat{\mathbf{h}}(n) = [\hat{\mathbf{h}}_1^T(n),\ \hat{\mathbf{h}}_2^T(n),\ \ldots,\ \hat{\mathbf{h}}_M^T(n)]^T$ is the vector of channel estimates. In Fig. 2, the NPM is plotted against the estimated number of common zeros. As we can see, when the number of common zeros is not correctly estimated, the identification fails. This is because the subspace algorithm, in its simplest form, fails when the channel order is not correctly estimated, although it can be extended to work in the case of overestimation. The common zeros have an effect of overestimation as will be discussed further in Section 4.

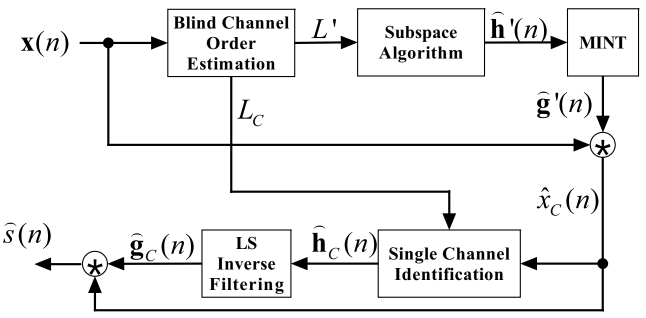

Figure 3: Two-stage speech dereverberation.

## 4. TWO-STAGE SPEECH DEREVERBERATION

A new speech dereverberation approach robust to common zeros is proposed based on the BSII scheme. First, the order of characteristic channel components is blindly estimated. Then, using only the multichannel observation of the reverberant speech signals, $\mathbf{h}_m'(n)$ is identified and inverted to form $x_C(n)$. Secondly, the component associated with the common zeros is estimated using a single channel approach. Finally, the dereverberated speech signal is obtained by using single channel inverse filtering. A system diagram is shown in Fig. 3.

### 4.1 Stage 1: Characteristic zeros component identification and inversion

Since $\mathbf{h}_m'(n)$ do not share any common zeros, we can identify them blindly using, for example, the subspace method [9], which is based on eigenvalue decomposition. Like most cur-

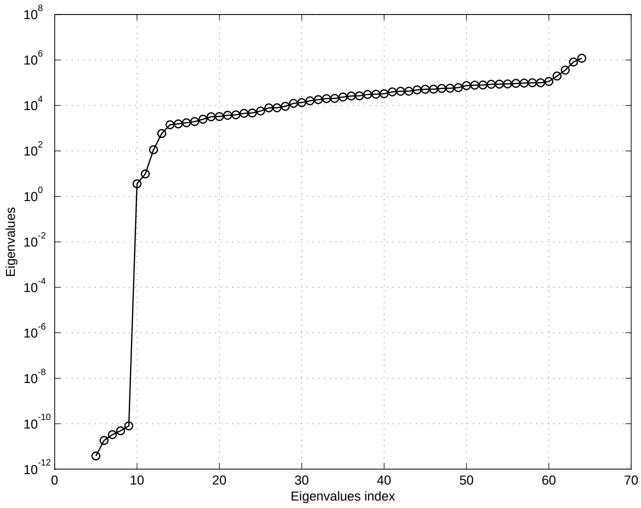

Figure 4: Channel order estimation.

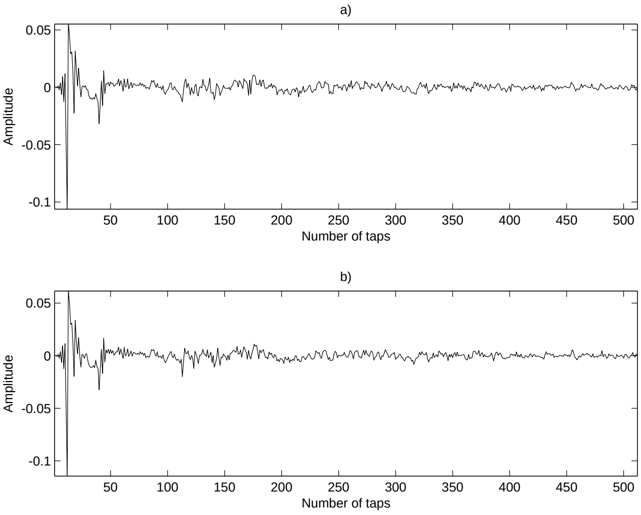

Figure 5: Measured room impulse responses from the MARDY database for a) channel 1, b) channel 2.

rent second-order-statistics-based BSI algorithms, the overall system order is assumed to be known or over-estimated. As shown in Section 2, for the case where there are common zeros across the RTFs, knowledge of $L'$ is required for $\mathbf{h}'_m(n)$ to be identified. We apply the eigenvalue-based method proposed in [10] to estimate the order of the characteristic component without having to factorize the polynomials arising from the RTFs.

Using the expression in (2), a system equation can be obtained by concatenating all $M$ outputs of (2) as follows:

$$\mathbf{x}(n) = \mathbf{H}(n)\mathbf{s}(n), \tag{5}$$

where $\mathbf{x}(n) = [\mathbf{x}_1^T(n),\ \mathbf{x}_2^T(n),\ \ldots,\ \mathbf{x}_M^T(n)]^T$ and $\mathbf{H}(n) = [\mathbf{H}_1^T(n),\ \mathbf{H}_2^T(n),\ \ldots,\ \mathbf{H}_M^T(n)]^T$. From (5), the autocorrelation matrix of the observed data $\mathbf{x}(n)$ can be written

$$\mathbf{R}_x = E\{\mathbf{x}(n)\mathbf{x}^T(n)\} = \mathbf{H}(n)\mathbf{R}_s\mathbf{H}^T(n), \tag{6}$$

where $\mathbf{R}_s = E\{\mathbf{s}(n)\mathbf{s}^T(n)\}$ is the autocorrelation matrix of the input signal.

Consider the two-channel SIMO system from Section 3 as an example. The overall system size is known to be $L = 32$, based on which $\mathbf{R}_x$ is constructed according to (6) and the 64 eigenvalues are computed for this two-channel case. The resulting eigenvalues, sorted in an ascending order are plotted in Fig. 4. From the figure, it is observed that the first 9 eigenvalues are distinctly smaller than the remaining ones. Thus the order of the characteristic component is $L' = L - L_C + 1 = 24$. This principle has been used in [2] for order estimation of over-estimated channels. In the case where there are common zeros present, the common zeros component is essentially the channel with over-estimated order when it comes to eigenvalue decomposition.

Using the subspace method, the channel estimates $\hat{\mathbf{h}}'_m(n)$ are the eigenvectors corresponding to the smallest eigenvalue of $\mathbf{R}_x$ and are determined up to an arbitrary scale factor. The signal $x_C(n)$ can then be obtained by

$$x_C(n) = \sum_{m=1}^{M} \sum_{i=0}^{L'_{inv}-1} \hat{g}'_{m,i}(n) x_m(n-i), \tag{7}$$

where $\hat{g}'_{m,i}(n)$, $i = 0, \ldots, L'_{inv} - 1$ are coefficients of the estimated inverse filters of $\hat{\mathbf{h}}'_m(n)$ obtained from MINT [11] by minimizing the error $\hat{\mathbf{g}}'_{m,i}(n) = \arg\min_{\mathbf{g}'_m} \|\sum_{m=1}^{M}\sum_{k=0}^{L'_{inv}-1} g'_{m,k}(n)\hat{h}'_{m,i-k}(n) - 1\|^2$ with $\hat{g}'_m(n) = [\hat{g}'_{m,0}(n),\ \hat{g}'_{m,1}(n),\ \ldots,\ \hat{g}'_{m,L'_{inv}-1}(n)]^T$.

### 4.2 Stage 2: Common zeros component identification and inversion

Using the results obtained from Section 4.1, the single channel associated with common zeros is to be identified and equalized. If the enclosure of the room and the position of the loudspeaker is fixed, we can assume that the coefficients of acoustic impulse responses do not change with time or at least they change very slowly. This enables us to employ a single channel identification approach which is based on the stationarity of channel zeros.

As can be seen from Fig. 1, the zeros associated with $\mathbf{h}_C(n)$ are contained within $x_C(n)$. Since $\mathbf{s}(n)$ is not known, the zeros of the common zeros component $\mathbf{h}_C(n)$ can be identified by exploiting the fixed pattern of their position over duration for which the channel is considered static.

This technique was employed in [12] for single channel dereverberation, where $x_C(n)$ is partitioned into several time-segments that are then factorized. The channel zeros are those which fall into the same position of z-plane for each segment. However, due to the linear convolution between the original speech signal and multiple room impulse responses, the channel zeros spread among the output signal, which can be very long. Thus the length of each segment has to be correspondingly long in order to capture every channel zero. Long factorization not only results in significant delay but also potentially introduces numerical approximation errors which can degrade the performance of zero identification and that of time domain coefficients reconstruction. Instead, multiple short frames of $x_C(n)$ can be used. The channel zeros are then identified sequentially as frames of $x_C(n)$ are acquired. Although it may not be possible to identify all the channel zeros within one frame, several frames are sufficient for the

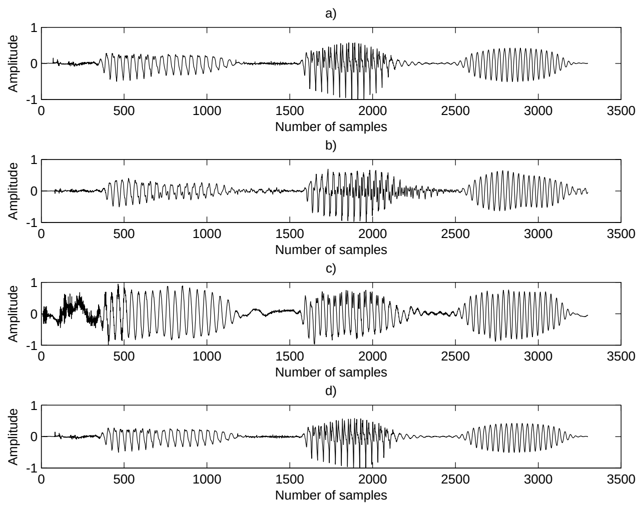

Figure 6: Performance of speech dereverberation: a) original speech, b) reverberant speech, c) dereverberation using standard subspace algorithm, d) dereverberation using proposed approach.

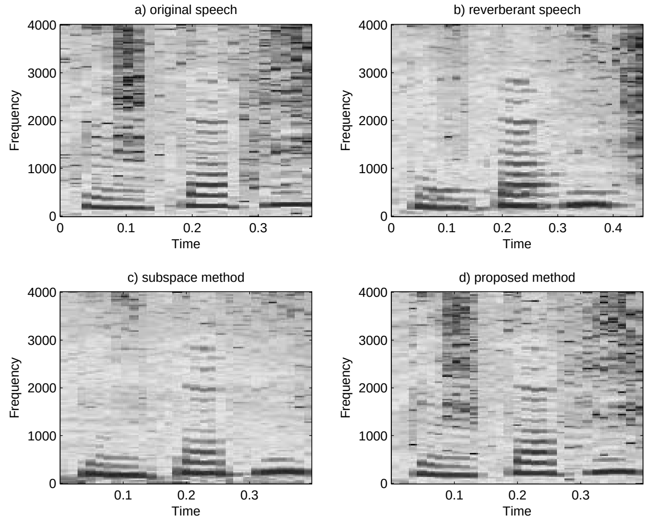

Figure 7: Spectrograms of a) original speech, b) reverberant speech, c) dereverberated speech using subspace method, d) dereverberated speech using proposed method.

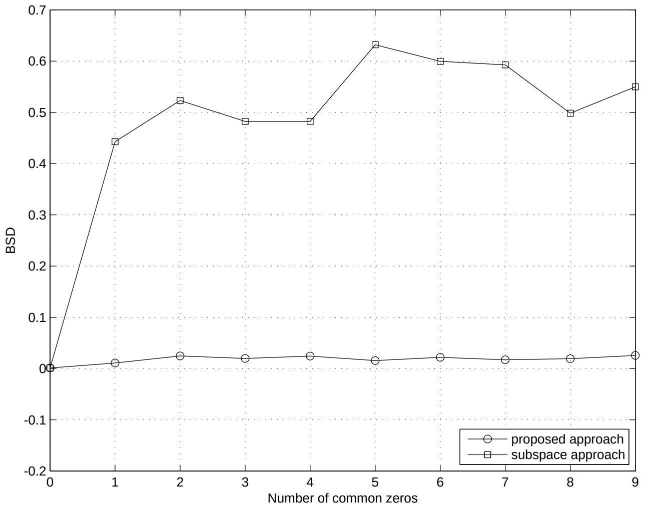

Figure 8: BSD vs. number of common zeros for the subspace method (square) and the proposed method (circle).

complete identification. In practice, the frame length and the number of frames can be determined based on the knowledge of $L_C$, as shown in Section 4.1, $L_C = L - L' + 1$. This implementation makes sure the performance of zero identification is maintained well while introducing only a small processing delay.

After identifying the common channel zeros, $\hat{\mathbf{h}}_C(n)$ is reconstructed and the original speech signal is estimated as

$$\hat{s}(n) = \sum_{i=0}^{L_{Cinv}-1} \hat{g}_{C,i}(n) x_C(n-i), \tag{8}$$

where $\hat{g}_{C,i}(n)$, $i = 0, \ldots, L_{Cinv} - 1$ are the coefficients of the estimated inverse filter of $\hat{\mathbf{h}}_C(n)$ obtained using the least squares method [13], i.e., $\hat{g}_{C,i}(n) = \arg\min_{g_C} \| \sum_{k=0}^{L_{Cinv}-1} g_{C,k}(n)\hat{h}_{C,i-k}(n) - \delta(n-\tau)\|^2$ where $\delta(n)$ denotes an impulse and $\tau$ is chosen to be $L_C/2$.

## 5. SIMULATIONS

Simulation results are next presented to demonstrate the performance of the proposed method. A speech sample comprising an utterance by a female talker sampled at 8 KHz is used as an example. Two measured acoustic impulse responses are obtained from the MARDY database [14]. These are then resampled at 8 KHz and truncated to $L = 512$ taps. The distance between each microphone is 0.05 m and the talker is positioned 3 m away from the microphones. Fig. 5 shows the resulting room impulse responses. To illustrate the effectiveness of our proposed scheme under condition containing common zeros, we super-imposed $L_C = 9$ randomly generated common zeros onto the measured responses from MARDY database. Note that $x_C(n)$ was segmented into frames of length 256 for stage 2, i.e., the single channel identification.

The dereverberation performance was measured using the Bark spectral distortion (BSD) [15], defined as,

$$\mathrm{BSD} = \frac{\sum_{k=0}^{K-1}\sum_{n=kN}^{kN+N-1} |B_s(k,n) - B_{\hat{s}}(k,n)|^2}{\sum_{k=0}^{K-1}\sum_{n=kN}^{kN+N-1} |B_s(k,n)|^2}, \tag{9}$$

where $B_s(k,n)$ is the Bark spectrum of $\mathbf{s}(n)$ and $N$ is the frame length in samples. The simulation result is plotted in Fig. 6, which shows a) the original speech signal, b) the reverberant speech signal, the recovered speech signal using c) the standard subspace algorithm and d) the proposed approach. It is seen that the proposed approach produces a better estimate of the original speech signal compared to the subspace method, with the presence of common zeros. The corresponding spectrogram is shown in Fig. 7, which confirms the improvement of the proposed method over subspace method. In terms of BSD performance for this simulation, the standard subspace method gave a score of 0.4309, while the proposed method gave 0.0016.

A further simulation was performed in which different numbers of randomly generated common zeros were superimposed onto the room impulse responses. The BSD performance for the subspace method and the proposed method was compared to demonstrate robustness to the number of common zeros. The simulation was run for 100 Monte Carlo runs and the result is shown in Fig. 8. It can be seen that the proposed algorithm provides a consistent improvement of BSD over subspace algorithm. Note that the proposed algorithm reduces to the subspace algorithm when there are no exactly common zeros, as indicated in Fig. 8.

## 6. CONCLUSIONS

We have investigated the problem of multichannel speech dereverberation in the presence of common zeros. It has been shown that algorithms based on BSII techniques suffer performance degradation due to common zeros. Consequently, we proposed a two-stage approach, where the characteristic zeros and the common zeros of the room transfer functions are identified and equalized separately. Simulation results based on real room acoustic impulse responses confirmed the improvement obtained with our approach.

## REFERENCES

[1] 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.

[2] 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.

[3] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. Audio, Speech and Language Process.*, vol. 15, no. 2, pp. 430–440, Feb. 2007.

[4] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, 1st ed. Springer-Verlag, Berlin, 2001.

[5] X. S. 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)*, Sept. 2006.

[6] 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.

[7] N. D. Gaubitch, J. Benesty, and P. A. Naylor, "Adaptive common root estimation and the common zeros problem in blind channel identification," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Sept. 2005.

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

[9] E. Moulines, P. Duhamel, J. F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Trans. Signal Process.*, vol. 43, no. 3, pp. 516–525, Feb. 1995.

[10] M. I. Gürelli and C. L. Nikias, "EVAM: An eigenvector-based algorithm for multichannel blind deconvolution of input colored signals," *IEEE Trans. Signal Process.*, vol. 43, no. 1, pp. 134–149, Jan. 1995.

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

[12] F. S. Pacheco and R. Seara, "A single-microphone approach for speech signal dereverberation," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Sept. 2005.

[13] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 7, May 1982, pp. 1858–1861.

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

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