Statistical analysis of the autoregressive modeling of reverberant speech

Nikolay D. Gaubitch,[a] Darren B. Ward, and Patrick A. Naylor
Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, London SW7 2AZ, United Kingdom

(Received 15 February 2006; revised 27 July 2006; accepted 29 August 2006)

Hands-free speech input is required in many modern telecommunication applications that employ autoregressive (AR) techniques such as linear predictive coding. When the hands-free input is obtained in enclosed reverberant spaces such as typical office rooms, the speech signal is distorted by the room transfer function. This paper utilizes theoretical results from statistical room acoustics to analyze the AR modeling of speech under these reverberant conditions. Three cases are considered: (i) AR coefficients calculated from a single observation; (ii) AR coefficients calculated jointly from an M-channel observation ($M > 1$); and (iii) AR coefficients calculated from the output of a delay-and sum beamformer. The statistical analysis, with supporting simulations, shows that the spatial expectation of the AR coefficients for cases (i) and (ii) are approximately equal to those from the original speech, while for case (iii) there is a discrepancy due to spatial correlation between the microphones which can be significant. It is subsequently demonstrated that at each individual source-microphone position (without spatial expectation), the M-channel AR coefficients from case (ii) provide the best approximation to the clean speech coefficients when microphones are closely spaced (<0.3m). © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2356840]

PACS number(s): 43.72.Ar, 43.55.Hy [DOS] Pages: 4031–4039

I. INTRODUCTION

Many hands-free telecommunication applications involving, for example, speech coding and speech enhancement, make use of autoregressive (AR) analysis techniques such as linear predictive coding (LPC). These applications are often employed in systems used inside rooms where the observed speech signal becomes reverberant due to the enclosed space. There is an interest in AR modeling of degraded speech, and the properties of the AR coefficients have been studied in the context of parameter quantization noise and ambient acoustical noise.[1–3] Several dereverberation algorithms have been proposed which operate on the linear prediction (LP) residual under the explicit or implicit assumptions that the AR coefficients are not affected by reverberation.[4–8] These methods utilize known features of the LP residual of speech signals to attenuate components due to reverberation. Yegnanarayana and Satyanarayana[6] provided a comprehensive study on the effects of reverberation on the LP residual. We now present an investigation of the effects of reverberation on the AR coefficients.

We utilize tools from statistical room acoustics (SRA) theory[9–11] for the analysis of the relation between the sets of AR coefficients obtained from clean speech and those obtained from reverberant speech. SRA provides a means for describing the sound field in a room that is mathematically tractable compared to, for example, wave theory.[9] SRA has been shown useful for the analysis of signal-processing techniques in reverberant environments and has recently been applied by several researchers. Radlović *et al.*,[10] Talantzis *et al.*,[12] and Bharitkar *et al.*[13] utilized SRA to investigate the robustness of channel equalization. Further, Talantzis *et al.*[14] investigated the performance of blind source separation, Gustafsson *et al.*[15] analyzed the performance of sound source localization, and Ward[16] used SRA to measure the performance of acoustic crosstalk cancellation in reverberant environments.

In our study, we will consider three cases: (i) AR coefficients calculated from a single observation; (ii) AR coefficients jointly calculated from an M-channel observation ($M > 1$); and (iii) AR coefficients obtained from the output of a delay-and-sum beamformer (DSB). Extending the work in Ref. 17, we will show in terms of spatial expectation that the AR coefficients obtained from reverberant speech are approximately equal to those from clean speech for cases (i) and (ii), while the AR coefficients obtained from the output of the delay-and-sum beamformer differ due to spatial correlation between the microphones. Furthermore, it will be demonstrated that the M-channel AR coefficients from (ii) provide the best estimate of the clean speech coefficients compared to the other two cases under consideration. We believe that our results here also relate to and explain the following statement in Ref. 4: "*…it has been recognized that any practical or typical room transfer function has certain properties that make it possible to accurately determine the speaker's vocal tract transfer function from the reverberative speech signal.*" and "*…arrays of plural microphones can also be used to advantage…*" which continues "*For this case, each new microphone requires its own correlation computer. The new outputs from this computer $R'(\tau_1)$, $R'(\tau_2), \ldots, R'(\tau_{14})$ are added to the other $R(\tau)$'s of other microphones thus giving more accurate data for the coefficient computer.*"

[a] Electronic mail: ndg@imperial.ac.uk

The remainder of this paper is organized as follows. In Sec. II we review the statistical room acoustic model including the conditions under which the theory is valid. The simulation environment is defined in Sec. III. In Sec. IV an analysis of the effects of reverberation on the AR coefficients and on the residual signal is presented for the single channel case. Section V presents the analysis of the two multichannel AR modeling cases. Simulation results are presented in Sec. VI and finally conclusions regarding AR modeling of reverberant speech are drawn in Sec. VII.

II. STATISTICAL ROOM ACOUSTICS

In this section, the statistical model of room reverberation and the conditions under which this is assumed valid are summarized. Within the framework of SRA, the sound field at a point in a room consists of the superposition of many acoustic plane waves arriving from all possible directions and with randomly distributed amplitudes and phases such that they form a uniform, diffuse sound field.[9,11] Subsequently, the room transfer function (RTF) of the acoustic channel from the source to the mth microphone can be expressed as the sum of a direct component, $H_{d,m}(e^{j\omega})$ and a reverberant component, $H_{r,m}(e^{j\omega})$, such that

$$H_m(e^{j\omega}) = H_{d,m}(e^{j\omega}) + H_{r,m}(e^{j\omega}), \quad m = 1,2,\ldots,M. \tag{1}$$

Under the conditions stated at the end of this section, and due to the different propagation directions and the random relation of the phases of the direct component and all the reflected waves, it can be assumed that the direct and the reverberant components are uncorrelated.[9,11] Hence, the spatial expectation of the cross terms of the squared magnitude of (1) is zero[10] and the spatially expected energy density spectrum of the RTF can be written

$$\mathcal{E}\{|H_m(e^{j\omega})|^2\} = |H_{d,m}(e^{j\omega})|^2 + \mathcal{E}\{|H_r(e^{j\omega})|^2\}, \tag{2}$$

where $\mathcal{E}\{\cdot\}$ is the spatial expectation operator, with the spatial expectation defined over all allowed microphone-source positions in a room.[15,11] Only the reverberant component varies with position, and its spatial expectation is independent of the microphone index m. The computation of $\mathcal{E}\{\cdot\}$ is described in Sec. III. The direct component of the RTF is the free-space Green's function, defined as[11]

$$H_{d,m}(e^{j\omega}) = \frac{e^{jkD_m}}{4\pi D_m}, \tag{3}$$

where D_m is the distance from the source to the mth microphone and $k=2\pi f/c$ is the wave number, with f denoting frequency and c the speed of sound in air, which we take at room temperature as $c=344$ m/s. From SRA, the expected density spectrum of the reverberant component is given by[9,10]

$$\mathcal{E}\{|H_r(e^{j\omega})|^2\} = \left(\frac{1-\alpha}{\pi A\alpha}\right), \tag{4}$$

with A being the total surface area of the room and α the average absorption coefficient of the room walls.

The spatial cross correlation of the reverberant paths between the mth and the nth channels has been shown to be[16]

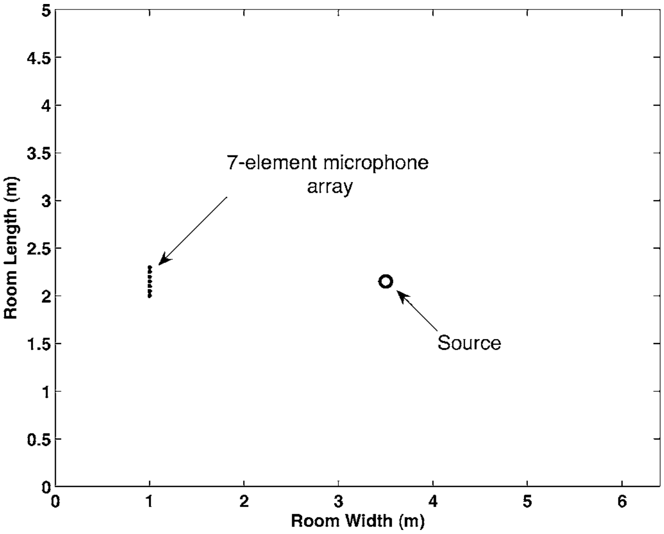

FIG. 1. Plan view of the simulated room environment with the initial position of the microphones ($\cdot$) and the source ($\circ$).

$$\mathcal{E}\{H_{r,m}(e^{j\omega})H_{r,n}^*(e^{j\omega})\} = \left(\frac{1-\alpha}{\pi A\alpha}\right)\frac{\sin k\|\ell_m - \ell_n\|}{k\|\ell_m - \ell_n\|}, \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean norm and ℓ_m is the three-dimensional position vector of the mth microphone, with the origin at $(x,y,z)=(0,0,0)$.

These approximations are known to represent closely the acoustic properties of a room provided that the following conditions are satisfied:[9,10]

(1) The dimensions of the room are large relative to the wavelength at all frequencies of interest.
(2) The average spacing between the resonant frequencies of the room is smaller than one-third of their bandwidth. This can be satisfied at all frequencies above the Schroeder frequency defined as

$$f_{\text{Sch}} = 2000\sqrt{\frac{T_{60}}{V}}\ \text{Hz}, \tag{6}$$

where T_{60} is the reverberation time and V is the volume of the room in cubic meters.
(3) Speaker and microphones are situated in the room interior, at least a half-wavelength from the surrounding walls.

These conditions usually hold for most practical situations over the significant speech bandwidth. Also, image method[18] simulations and measured impulse responses of a real room have been demonstrated to coincide closely with SRA theory.[15]

III. EXPERIMENTAL ENVIRONMENT

We consider a room with a single source and an array of microphones as depicted in Fig. 1. All results presented in this paper are based on computer simulations with the simulated environment defined as follows. The dimensions of the room were set to 4×5×6.4 m. These dimensions were specifically chosen to conform with the ratio (1:1.25:1.6), as in Ref. 10, in order to obtain the best approximation of a diffuse sound field so as to satisfy the conditions above. Unless stated otherwise, the microphones were positioned in a linear array configuration with the distance between adjacent microphones set to $\|\boldsymbol{\ell}_m-\boldsymbol{\ell}_{m+1}\|=0.05$ m. The source was at a distance $D=2.5$ m from the center of the array. The source and the microphones were assumed omnidirectional and were always at least a half-wavelength from the surrounding walls, where the wavelength is taken with respect to the lowest frequency component in the signal. The source-image method for modeling small room acoustics,[18] modified to accommodate fractional sample delays according to Ref. 19, was used to generate finite room impulse responses, $h(n)$. The room transfer function, $H(e^{j\omega})$, was then found by taking the Fourier transform of $h(n)$. Anechoic speech samples were taken from the APLAWD database[20] and all the signals under consideration were bandlimited to 300–7000 Hz with a sampling frequency $f_s=16$ kHz.

To compute the spatial expectation, $\mathcal{E}\{\cdot\}$, we utilized the method used by Radlovic *et al.*[10] and Gustafsson *et al.*[15] An initial position for the source, $\boldsymbol{y}_0$, and for each of the microphones, $\boldsymbol{\ell}_{m,0}$, was selected. A random translation vector, $\boldsymbol{\theta}$, and a random rotation matrix, $\boldsymbol{\mathcal{R}}$, were generated and applied to the initial coordinates of the source-receiver configuration to obtain the ith realization coordinates $\boldsymbol{y}_i=\boldsymbol{\mathcal{R}}\boldsymbol{y}_0+\boldsymbol{\theta}$ and $\boldsymbol{\ell}_{m,i}=\boldsymbol{\mathcal{R}}\boldsymbol{\ell}_{m,0}+\boldsymbol{\theta}$. In this way, the distance between the source and the microphones and between successive microphones is kept constant for all $i=1,2,\ldots,N$. An estimate of $\mathcal{E}\{\cdot\}$ is obtained by taking the average of the N outcomes.

IV. SINGLE-CHANNEL AR MODELING OF REVERBERANT SPEECH

In this section, we consider AR modeling of speech using linear prediction[21,22] and we present the analysis of the effects of reverberation on the AR coefficients obtained from a single channel. We also discuss the consequences this has on the linear prediction residual.

A. AR modeling of speech

In order to introduce the notation used in the rest of the paper, we provide a brief summary of the AR modeling of speech. A speech signal, $s(n)$, can be expressed as a linear combination of its p past samples using a linear predictor[21]

$$s(n) = -\boldsymbol{a}^T\boldsymbol{s}(n-1) + e(n), \tag{7}$$

where $\boldsymbol{a}=[a_1 \ \ a_2 \ \ \cdots \ \ a_p]^T$ is a $p\times 1$ vector of AR coefficients with $[\cdot]^T$ denoting matrix transpose, $\boldsymbol{s}(n-1)=[s(n-1)s(n-2)\cdots s(n-p)]^T$ is a vector of input samples at time n, $e(n)$ is the LP residual, and p is the prediction order. The prediction error filter and the all-pole predictor are, respectively,

$$A(z) = 1 + \boldsymbol{a}^T\boldsymbol{z} \tag{8}$$

and

$$V(z) = 1/A(z), \tag{9}$$

where $\boldsymbol{z}=[z^{-1} \ \ z^{-2} \ \ \cdots \ \ z^{-p}]$.

The AR coefficients can be obtained by minimizing the sum of the squared prediction error,

$$J = \sum_{n=-\infty}^{\infty} e^2(n) = \sum_{n=-\infty}^{\infty} (s(n) + \boldsymbol{a}^T\boldsymbol{s}(n-1))^2, \tag{10}$$

with respect to each of the coefficients in $\boldsymbol{a}$. Equivalently, by Parseval's theorem, a frequency domain formulation of the error in (10) can be expressed as[21]

$$J = \frac{1}{2\pi}\int_{-\pi}^{\pi} |E(e^{j\omega})|^2 d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} |1 + \boldsymbol{a}^T\boldsymbol{d}|^2|S(e^{j\omega})|^2 d\omega, \tag{11}$$

where $S(e^{j\omega})$ and $E(e^{j\omega})$ are the Fourier transforms of $s(n)$ and $e(n)$ respectively, and $\boldsymbol{d}=[e^{-j\omega} \ \ e^{-j2\omega} \ \ \cdots \ \ e^{-jp\omega}]^T$ is a $p\times 1$ DFT vector. The optimum set of p AR coefficients that minimize the error J is

$$\boldsymbol{a}_{\text{opt}} = \arg\min_{\boldsymbol{a}} J = -\boldsymbol{R}^{-1}\boldsymbol{r}, \tag{12}$$

where

$$\boldsymbol{R} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S(e^{j\omega})|^2 \boldsymbol{d}\boldsymbol{d}^H d\omega \tag{13}$$

is a $p\times p$ autocorrelation matrix, with $[\cdot]^H$ denoting Hermitian (complex conjugate) matrix transpose and

$$\boldsymbol{r} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |S(e^{j\omega})|^2 \boldsymbol{d} d\omega \tag{14}$$

is a $p\times 1$ vector of autocorrelation coefficients. In practice, the error signal is evaluated over finite windowed frames;[21] however, in this paper the effects of the window will not be considered.

B. Effect of reverberation on the AR coefficients

Consider a speech signal, $s(n)$, produced at a point in a noiseless, reverberant room. The observation by a single microphone positioned at some distance from the speaker is denoted

$$x(n) = \boldsymbol{h}^T\boldsymbol{s}(n), \tag{15}$$

where $\boldsymbol{h}=[h_0 \ \ h_1 \ \ \cdots \ \ h_{L-1}]^T$ is the L-tap impulse response of the acoustic channel from the source to the microphone and $\boldsymbol{s}(n)=[s(n) \ \ s(n-1) \ \ \cdots \ \ s(n-L+1)]^T$ is the input vector at time n. The relation between the AR coefficients obtained by linear prediction from $s(n)$ and those from $x(n)$ is summarized in Theorem 1.

Theorem 1 *Let* $\boldsymbol{a}_{opt}=[a_{opt,1} \ \ a_{opt,2} \ \ \cdots \ \ a_{opt,p}]^T$ *be the optimum set of AR coefficients obtained from the clean speech signal*, $s(n)$, *and* $\boldsymbol{b}_{opt}=[b_{opt,1} \ \ b_{opt,2} \ \ \cdots \ \ b_{opt,p}]^T$ *the optimum set of AR coefficients obtained from the reverberant speech signal*, $x(n)$. *The spatially expected values of the reverberant speech AR coefficients are approximately equal to those of the AR coefficients calculated from clean speech, i.e.,*

$$\mathcal{E}\{\boldsymbol{b}_{\text{opt}}\} \cong \boldsymbol{a}_{\text{opt}}. \tag{16}$$

Proof: We apply LP analysis on the reverberant speech signal, $x(n)$, to obtain the optimum set of AR coefficients

$$\boldsymbol{b}_{\mathrm{opt}} = -\boldsymbol{Q}^{-1}\boldsymbol{q}, \tag{17}$$

with

$$\boldsymbol{Q} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 |S(e^{j\omega})|^2 \boldsymbol{d}\boldsymbol{d}^H d\omega \tag{18}$$

and

$$\boldsymbol{q} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{j\omega})|^2 |S(e^{j\omega})|^2 \boldsymbol{d} d\omega, \tag{19}$$

where $\boldsymbol{Q}$ is a $p \times p$ autocorrelation matrix and $\boldsymbol{q}$ is a $p \times 1$ vector of autocorrelation coefficients.

In order to study the AR coefficients of reverberant speech, we take the expectation on both sides of (17)

$$\mathcal{E}\{\boldsymbol{b}_{\mathrm{opt}}\} = -\mathcal{E}\{\boldsymbol{Q}^{-1}\boldsymbol{q}\}. \tag{20}$$

However, we would like to consider the expectation of each term of (20). Adopting the approach used in Ref. 10 and Ref. 12, we use the zeroth-order Taylor series expansion to write $\mathcal{E}\{g(x)\} \cong g(\mathcal{E}\{x\})$, as detailed in the Appendix, and therefore (20) can be written as

$$\mathcal{E}\{\boldsymbol{b}_{\mathrm{opt}}\} \cong -\mathcal{E}\{\boldsymbol{Q}\}^{-1}\mathcal{E}\{\boldsymbol{q}\}. \tag{21}$$

This reduces the problem to studying the properties of the AR coefficients in terms of the autocorrelation function.

Now, consider the spatial expectation of the uth element of $\boldsymbol{q}$ in (19)

$$\mathcal{E}\{q_u\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{E}\{|H(e^{j\omega})|^2\} |S(e^{j\omega})|^2 e^{-j\omega u} d\omega, \tag{22}$$

for $u = 1, 2, \ldots, p$. The term $S(e^{j\omega})$ is taken outside the spatial expectation since it is independent of the source-microphone position.

From (2)–(4) the SRA expression for the expected energy density spectrum of the RTF is

$$\mathcal{E}\{|H(e^{j\omega})|^2\} = \frac{1}{(4\pi D)^2} + \left(\frac{1-\alpha}{\pi A \alpha}\right) = \gamma. \tag{23}$$

Since γ is independent of frequency, by substitution of (23) into (22) we arrive at

$$\mathcal{E}\{q_u\} = \frac{\gamma}{2\pi}\int_{-\pi}^{\pi} |S(e^{j\omega})|^2 e^{-j\omega u} d\omega = \gamma r_u, \tag{24}$$

for $u = 1, 2, \ldots, p$, where r_u is the uth element of the clean speech autocorrelation vector $\boldsymbol{r}$, and by similar reasoning the (u,v)th element of $\boldsymbol{Q}$ in (18) becomes

$$\mathcal{E}\{Q_{u,v}\} = \gamma R_{u,v}, \quad u,v = 1,2, \ldots, p, \tag{25}$$

where $R_{u,v}$ is the (u,v)th element of the clean speech autocorrelation matrix $\boldsymbol{R}$. Substituting the results from (24) and (25) into (21) gives (16). □

This result states that if LP analysis is applied to reverberant speech, the coefficients $\boldsymbol{a}_{\mathrm{opt}}$ and $\boldsymbol{b}_{\mathrm{opt}}$ are not necessarily equal at a single observation point in space. However, in terms of spatial expectation, the AR coefficients from reverberant speech are approximately equal to those from clean speech. The accuracy of the approximation depends on the accuracy of estimation of the spatial expectation of the autocorrelation function.

C. Effect of reverberation on the prediction residual

Consider a frequency domain formulation of the source-filter model described in Sec. IV A. The speech signal is expressed as

$$S(e^{j\omega}) = E(e^{j\omega})V(e^{j\omega}), \tag{26}$$

where $E(e^{j\omega})$ is the Fourier transform of the LP residual and $V(e^{j\omega})$ is the transfer function all-pole filter from (9) evaluated for $z = e^{j\omega}$.

Now, consider the speech signal produced in a reverberant room as defined in (15), which in the frequency domain leads to

$$X(e^{j\omega}) = S(e^{j\omega})H(e^{j\omega}) = E(e^{j\omega})V(e^{j\omega})H(e^{j\omega}). \tag{27}$$

Referring to (16), an inverse filter, $B(e^{j\omega}) = 1 + \Sigma_{k=1}^{p} b_k e^{j\omega k}$, can be obtained such that $\mathcal{E}\{B(e^{j\omega})\} \cong A(e^{j\omega})$, where $A(e^{j\omega})$ is given by (8) for $z = e^{j\omega}$. Filtering the reverberant speech signal with this inverse filter, whose coefficients are obtained from the reverberant speech signal, results in

$$\hat{E}(e^{j\omega}) \cong E(e^{j\omega})H(e^{j\omega}), \tag{28}$$

where $\hat{E}(e^{j\omega})$ is the Fourier transform of the LP residual, $\hat{e}(n)$, obtained from the reverberant speech signal. Thus, in the time domain, the LP residual obtained from reverberant speech is approximately equal to the clean speech residual convolved with the room impulse response. The approximation in (28) arises from the AR modeling. Therefore, if the AR coefficients used were identical to those from clean speech, the approximation would be an equivalence.

In summary, we have shown that the AR coefficients obtained from reverberant speech are approximately equal to those from clean speech in terms of spatial expectation. Furthermore, the LP residual obtained from a reverberated speech signal is approximately equal to the clean speech residual convolved with the room impulse response. This approximation depends on the accuracy of the estimation of the AR coefficients. Intuitively, the result in (16) suggests that using a microphone array in a manner so as to approximate the taking of the spatial expectation will give a more accurate estimation of the AR coefficients than use of a single observation alone. This motivates our study of multichannel AR modeling in the following section.

V. MULTICHANNEL AR MODELING OF REVERBERANT SPEECH

Several microphone array techniques have been applied as preprocessing in speech applications, proving advantageous to single-channel algorithms.[23] In this section, we investigate the use of a microphone array to obtain the AR coefficients and how these compare to the AR coefficients from clean speech. Two alternative approaches are considered. In the first alternative, the AR coefficients are obtained by formulating an estimation procedure that jointly minimizes the squared errors over all M channels. In the second alternative, the AR coefficients are obtained from the output of an M-channel array using delay-and-sum beamforming.

A. *M*-channel AR coefficients

The speech signal observed at the mth microphone in an array of M microphones can be expressed as

$$x_m(n) = \boldsymbol{h}_m^T \boldsymbol{s}(n), \quad m = 1,2,\ldots,M, \tag{29}$$

where $\boldsymbol{h} = [h_{m,0} \;\; h_{m,1} \;\; \cdots \;\; h_{m,L-1}]^T$ is the L-tap room impulse response from the source to the mth microphone.

In linear prediction terms, the observation at the mth sensor from (29) can be written as

$$x_m(n) = -\boldsymbol{b}_m^T \boldsymbol{x}_m(n-1) + e_m(n), \quad m = 1,2,\ldots,M, \tag{30}$$

where $\boldsymbol{b}_m = [b_{m,1} \;\; b_{m,2} \;\; \cdots \;\; b_{m,p}]^T$ are the prediction coefficients, $\boldsymbol{x}_m(n-1) = [x_m(n-1)x_m(n-2)\cdots x_m(n-p)]^T$ is the mth microphone observation vector at time n, and $e_m(n)$ is the prediction residual obtained from the mth microphone signal. From (30), a joint M-channel error function can be formulated as[23]

$$\begin{aligned} J_M &= \frac{1}{M}\sum_{m=1}^{M}\sum_{n=-\infty}^{\infty} e_m^2(n) \\ &= \frac{1}{M}\sum_{m=1}^{M}\sum_{n=-\infty}^{\infty} (x_m(n) + \boldsymbol{b}_m^T \boldsymbol{x}_m(n-1))^2. \end{aligned} \tag{31}$$

The optimum set of coefficients that minimize this error, similarly to (12), is given by

$$\hat{\boldsymbol{b}}_{\text{opt}} = -\hat{\boldsymbol{Q}}^{-1}\hat{\boldsymbol{q}}, \tag{32}$$

with

$$\hat{\boldsymbol{Q}} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{Q}_m \tag{33}$$

and

$$\hat{\boldsymbol{q}} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{q}_m, \tag{34}$$

where $\hat{\boldsymbol{Q}}$ and $\hat{\boldsymbol{q}}$ are, respectively, the $p \times p$ mean autocorrelation matrix and the $p \times 1$ mean autocorrelation vector across the M microphones. The relation between the clean speech coefficients and the coefficients obtained using (32) is summarized in Corollary 1.

Corollary 1 *Replacing* (18) and (19) *with their averages considered over M microphones* (33) and (34) *and then following the steps of the proof of Theorem 1, it can be shown that the spatial expectation of the AR coefficients obtained from minimization of* (31) *is approximately equal to those from clean speech. That is,*

$$\mathcal{E}\{\hat{\boldsymbol{b}}_{\text{opt}}\} \cong \boldsymbol{a}_{\text{opt}}. \tag{35}$$

This result implies that the optimal AR coefficients obtained using a spatial expectation over M channels are equivalent to the spatial expectation of the AR coefficients in the single-microphone case in (16). However, at each individual position, the M-channel case provides a more accurate estimation of the clean speech AR coefficients than that obtained with a single reverberant channel, as will be shown by simulations in Sec. VI. This is because the averaging of the autocorrelation functions in (33) and (34) is equivalent in effect to the calculation of the spatial expectation operation in the single-channel case (21).

B. AR coefficients from a DSB output

The output of a delay-and-sum beamformer can be written[24] as

$$\bar{x}(n) = \frac{1}{M}\sum_{m=1}^{M} x_m(n-\tau_m), \tag{36}$$

where τ_m is the propagation delay in samples from the source to the mth sensor. Assuming that the time delays of arrival are known for all microphones, linear prediction can be performed on the beamformer output, $\bar{x}(n)$, following the approach in Sec. IV B. This is summarized in Theorem 2.

Theorem 2 *Let* $\boldsymbol{a}_{opt} = [a_{opt,1} \;\; a_{opt,2} \;\; \cdots \;\; a_{opt,p}]^T$ *be the optimum set of AR coefficients obtained from the clean speech signal,* $s(n)$, *and* $\bar{\boldsymbol{b}} = [\bar{b}_{opt,1} \;\; \bar{b}_{opt,2} \;\; \cdots \;\; \bar{b}_{opt,p}]^T$ *be the optimum set of AR coefficients obtained from the DSB output,* $\bar{x}(n)$. *The spatial expectation of the AR coefficients calculated by linear prediction from the output of the DSB is*

$$\mathcal{E}\{\bar{\boldsymbol{b}}_{\text{opt}}\} \cong \boldsymbol{T}\boldsymbol{a}_{\text{opt}} - \boldsymbol{t}, \tag{37}$$

with $\boldsymbol{T} = \boldsymbol{I} - (1/\bar{\gamma})\boldsymbol{R}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Gamma}^H(1/\bar{\gamma})\boldsymbol{R}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^H$ *and* $\boldsymbol{t} = (\bar{\gamma}\boldsymbol{R} + \boldsymbol{\Xi})^{-1}\boldsymbol{\xi}$, *where these terms are defined in the following proof.*

Proof: Consider a speech signal source, $s(n)$, observed using M microphones and combined using a DSB to give a signal $\bar{x}(n)$. In the frequency domain this can be expressed as

$$\bar{X}(e^{j\omega}) = \left(\frac{1}{M}\sum_{m=1}^{M} H_m(e^{j\omega})e^{-j2\pi f \tau_m}\right) S(e^{j\omega}) = \bar{H}(e^{j\omega})S(e^{j\omega}), \tag{38}$$

where $\bar{X}(e^{j\omega})$ is the Fourier transform of $\bar{x}(n)$, $S(e^{j\omega})$ is the Fourier transform of $s(n)$, $H_m(e^{j\omega})$ is the RTF with respect to the mth microphone, and $\bar{H}(e^{j\omega})$ is the averaged RTF at the DSB output. The AR coefficients, $\bar{\boldsymbol{b}}_{\text{opt}}$, at the beamformer output are calculated as in Sec. IV A,

$$\bar{\boldsymbol{b}}_{\text{opt}} = -\bar{\boldsymbol{Q}}^{-1}\bar{\boldsymbol{q}}, \tag{39}$$

with

$$\bar{\boldsymbol{Q}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\bar{H}(e^{j\omega})|^2 |S(e^{j\omega})|^2 \boldsymbol{d}\boldsymbol{d}^H d\omega \tag{40}$$

and

$$\bar{\boldsymbol{q}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\bar{H}(e^{j\omega})|^2 |S(e^{j\omega})|^2 \boldsymbol{d} d\omega, \tag{41}$$

where $\bar{\boldsymbol{Q}}$ is a $p \times p$ autocorrelation matrix and $\bar{\boldsymbol{q}}$ is a $p \times 1$ vector.

From this point on, we omit the frequency index for reasons of clarity. The expected energy density spectrum of the averaged RTFs can be written as

$$\mathcal{E}\{|\bar{H}|^2\} = \frac{1}{M^2}\left[\sum_{m=1}^{M}\mathcal{E}\{|H_m|^2\} + \sum_{m=1}^{M}\sum_{\substack{n=1\\ n\neq m}}^{M}\mathcal{E}\{H_mH_n^*\}e^{-j2\pi f(\tau_m-\tau_n)}\right]. \qquad (42)$$

From Sec. II, the expected energy density for the mth channel is

$$\mathcal{E}\{|H_m|^2\} = \frac{1}{(4\pi D_m)^2} + \left(\frac{1-\alpha}{\pi A\alpha}\right), \qquad (43)$$

and the expected cross correlation between the mth and the nth microphones is

$$\mathcal{E}\{H_mH_n^*\} = \frac{e^{jk(D_m-D_n)}}{16\pi^2 D_mD_n} + \left(\frac{1-\alpha}{\pi A\alpha}\right)\frac{\sin k\|\boldsymbol{\ell}_m-\boldsymbol{\ell}_n\|}{k\|\boldsymbol{\ell}_m-\boldsymbol{\ell}_n\|}. \qquad (44)$$

By substituting (43) and (44) into (42) and with $\tau_m = D_m/c$, we obtain the following expression for the mean energy density at the DSB output:

$$\mathcal{E}\{|\bar{H}|^2\} = \bar{\gamma} + \psi(\omega), \qquad (45)$$

with

$$\bar{\gamma} = \frac{1}{(4\pi M)^2}\sum_{m=1}^{M}\sum_{n=1}^{M}\frac{1}{D_mD_n} + \left(\frac{1-\alpha}{M\pi A\alpha}\right)$$

and

$$\psi(\omega) = \left(\frac{1-\alpha}{M^2\pi A\alpha}\right)\sum_{m=1}^{M}\sum_{\substack{n=1\\ n\neq m}}^{M}\frac{\sin k\|\boldsymbol{\ell}_m-\boldsymbol{\ell}_n\|}{k\|\boldsymbol{\ell}_m-\boldsymbol{\ell}_n\|} \times\cos(k[D_m-D_n]),$$

where $\bar{\gamma}$ is a frequency-independent component and $\psi(\omega)$ is a component due to spatial correlation.

Now, let

$$\xi_u = \frac{1}{2\pi}\int_{-\pi}^{\pi}\psi(\omega)|S(e^{j\omega})|^2 e^{-j\omega u}d\omega \qquad (46)$$

and

$$\Xi_{u,v} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\psi(\omega)|S(e^{j\omega})|^2 e^{-j\omega(u-v)}d\omega \qquad (47)$$

be the uth element of a vector $\boldsymbol{\xi}$ and the (u,v)th element of a matrix $\boldsymbol{\Xi}$, respectively. The expected value of the uth element of $\bar{\boldsymbol{q}}$ from (41) then becomes

$$\mathcal{E}\{\bar{q}_u\} = \bar{\gamma}r_u + \xi_u, \quad u = 1,2,\ldots,p, \qquad (48)$$

where r_u is the uth element of the vector $\boldsymbol{r}$ in (14). Similarly, the expected value of the (u,v)th element of $\bar{\boldsymbol{Q}}$ from (40) is

$$\mathcal{E}\{\bar{Q}_{u,v}\} = \bar{\gamma}R_{u,v} + \Xi_{u,v}, \quad u,v = 1,2,\ldots,p, \qquad (49)$$

where $R_{u,v}$ is the (u,v)th element of the matrix $\boldsymbol{R}$ in (13). The expected set of coefficients for the DSB output is therefore

$$\mathcal{E}\{\bar{\boldsymbol{b}}_{\text{opt}}\} \cong -(\bar{\gamma}\boldsymbol{R} + \boldsymbol{\Xi})^{-1}(\bar{\gamma}\boldsymbol{r} + \boldsymbol{\xi}). \qquad (50)$$

Since $\boldsymbol{\Xi}$ is a Hermitian symmetric matrix, it can be factored as

$$\boldsymbol{\Xi} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^H, \qquad (51)$$

where $\boldsymbol{\Gamma}$ is a matrix of eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues. Using the matrix inversion lemma,[25] we can write

$$(\bar{\gamma}\boldsymbol{R} + \boldsymbol{\Xi})^{-1} = \frac{1}{\bar{\gamma}}\boldsymbol{R}^{-1} - \frac{1}{\bar{\gamma}^2}\boldsymbol{R}^{-1}\boldsymbol{\Gamma} \times\left(\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Gamma}^H\frac{1}{\bar{\gamma}}\boldsymbol{R}^{-1}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}^H\boldsymbol{R}^{-1}. \qquad (52)$$

Finally, substituting the result from (52) into (50) we obtain the result in (37). □

Theorem 2 states that, in terms of spatial expectation, the AR coefficients obtained by LP analysis of the DSB output, $\bar{x}(n)$, differ from those obtained from clean speech. This difference depends on the spatial cross correlation between the acoustic channels. It can be seen from (5) that the interchannel correlation and its significance are governed by the reverberation time, the distance between adjacent microphones, the source-microphone separation, and on the array size if the speaker is in the near field of the microphone array. Of particular interest is the separation of adjacent microphones. From (45) it is evident that the term $\psi(\omega)$ and, consequently, the matrix $\boldsymbol{\Xi}$ and the vector $\boldsymbol{\xi}$, will tend to zero as the source-microphone separation is increased. Therefore, for large intermicrophone separation the matrix $\boldsymbol{T}$ tends to the identity matrix $\boldsymbol{I}$ and the vector $\boldsymbol{t}$ tends to zero and the result in (37) tends to the result in (16). Furthermore, if estimates of $\boldsymbol{T}$ and $\boldsymbol{t}$ were available, since $\boldsymbol{T}$ is a square matrix the effects of the spatial cross correlation could be compensated as $\boldsymbol{a}_{\text{opt}} \cong \boldsymbol{T}^{-1}(\mathcal{E}\{\bar{\boldsymbol{b}}_{\text{opt}}\} + \boldsymbol{t})$. However, estimating these parameters is difficult in practice. Finally, for the special case where the distance between the microphones is exactly a multiple of a half-wavelength at each frequency and the speaker is far from the microphones, then $\psi(\omega) = 0, \forall\,\omega$ and thus $\boldsymbol{\Xi}$ and $\boldsymbol{\xi}$ from (46) and (47) are equal to zero. Therefore, the matrix $\boldsymbol{T}$ becomes exactly the identity matrix $\boldsymbol{I}$ and the vector $\boldsymbol{t}$ is exactly zero, which results in the expression in (37) becoming equivalent to that in (16).

VI. SIMULATIONS AND RESULTS

Having established the theoretical relationship between the AR coefficients obtained from clean speech and those obtained from reverberant speech observations, we now present simulation results to demonstrate and to validate the theoretical analysis. In summary, we demonstrate two specific points: (1) On average over all positions in the room, the AR coefficients obtained from a single microphone as in

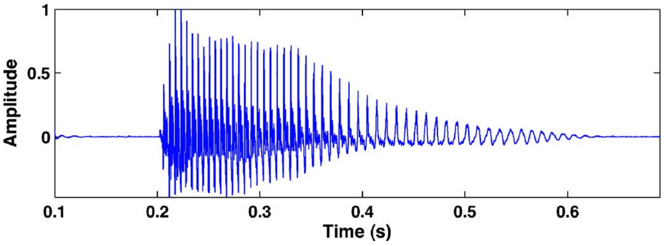

FIG. 2. Speech sample used in the experiments comprising the time-domain waveform of the diphthong /eɪ/ as in the alphabet letter "a" uttered by a male speaker.

case (i) and those calculated from M microphones as in case (ii) are not affected by reverberation while the AR coefficients from the DSB become more dissimilar from the clean speech coefficients with increased reverberation time. (2) The M-channel AR coefficients are the most accurate estimates of the clean speech AR coefficients from the three cases studied.

As an evaluation metric for the similarity between two sets of AR coefficients we use the Itakura distance measure,[22] defined as

$$d_I = \log\left(\frac{\hat{\boldsymbol{a}}^T \boldsymbol{R} \hat{\boldsymbol{a}}}{\boldsymbol{a}^T \boldsymbol{R} \boldsymbol{a}}\right), \tag{53}$$

where $\boldsymbol{a}$ is the set of clean speech coefficients and $\hat{\boldsymbol{a}}$ are the coefficients under test. The Itakura distance can be interpreted as the log ratio of the minimum mean squared errors (MMSE) obtained with the true and the estimated coefficients. The denominator represents the optimal solution for the clean speech and thus $d_I \geqslant 0$. For the experiments, the diphthong /eɪ/ as in the alphabet letter "a" uttered by a male speaker was used as an example and is depicted in Fig. 2. We performed the LP analysis on that sample employing selective linear prediction[26] with a frame length equal to the length of the vowel and a prediction order $p=21$ with sampling frequency $f_s=16$ kHz. The prediction order was chosen using the relation $p=f_s/1000+5$ as recommended in Ref. 22. Thus, this gives a pole pair per kHz of Nyquist sampling frequency and some additional poles to model the glottal pulse. For the selective linear prediction we consider the spectrum in the range 0.3–7 kHz in order to avoid errors due to bandlimiting filters.

A. Experiment 1

The spatial expectation was calculated using $N=200$ realizations of the source-array positions, and thus an average autocorrelation function was calculated for each of the cases under consideration. This was repeated, varying the reverberation time, T_{60}, from 0.1 to 0.9 s in steps of 0.2 s. For each case the Itakura distance measure was applied to the estimation of the spatial expectation of the coefficients. Figure 3 shows the results in which the Itakura distance of the spatially expected coefficients is plotted versus reverberation time for (a) a single channel; (b) $M=7$ channels; and (c) the DSB output simulation and the theoretical expression for the DSB output in (37) (dashed). It can be seen that the experimental outcome closely corresponds to the theoretical results where the coefficients from the M-channel case and from a single channel are close to the clean speech coefficients with an Itakura distance close to zero. In contrast, the difference between the results from the DSB output and the clean speech increases proportionally to the reverberation time, where the Itakura distance varies from $d_I=0.0028$ for $T_{60}=0.1$ s to $d_I=0.07$ at $T_{60}=0.9$ s.

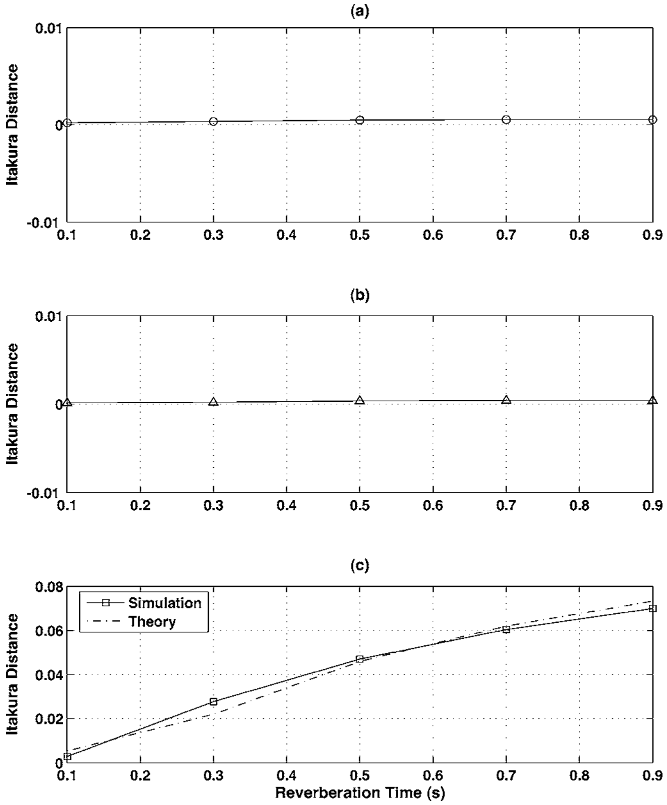

FIG. 3. Itakura distance vs reverberation time for the spatially expected AR coefficients of (a) a single channel; (b) $M=7$ channels; and (c) DSB output simulation and the theoretical expression for the DSB output (37) (dashed).

B. Experiment 2

This experiment illustrates the individual outcomes for the three cases at the $N=200$ different locations. Thus, using the same conditions as in experiment 1, the AR coefficients were computed at each individual source-array position using (17), (32), and (39) and the Itakura distance was calculated. Figure 4 shows the resulting plot in terms of the mean Itakura distance versus increasing reverberation time for (a) a single channel; (b) $M=7$ channels; and (c) a DSB output. The error bars indicate the range between the maximum and the minimum errors while the crosses indicate the mean value for all N locations. It can be seen that the M-channel LPC provides the best approximation of the clean speech AR coefficients. The mean Itakura distance is $d_I=0.01$ on average for all reverberation times, with a maximum distance of $d_I=0.057$ and a minimum distance $d_I=0.0015$ compared to a mean Itakura distance of $d_I=0.027$ for the single-channel case, where the maximum and the minimum distances are, respectively, $d_I=0.079$ and $d_I=0.0067$. It can also be seen that the estimation error for the AR coefficients obtained from the DSB output can become significant with increasing

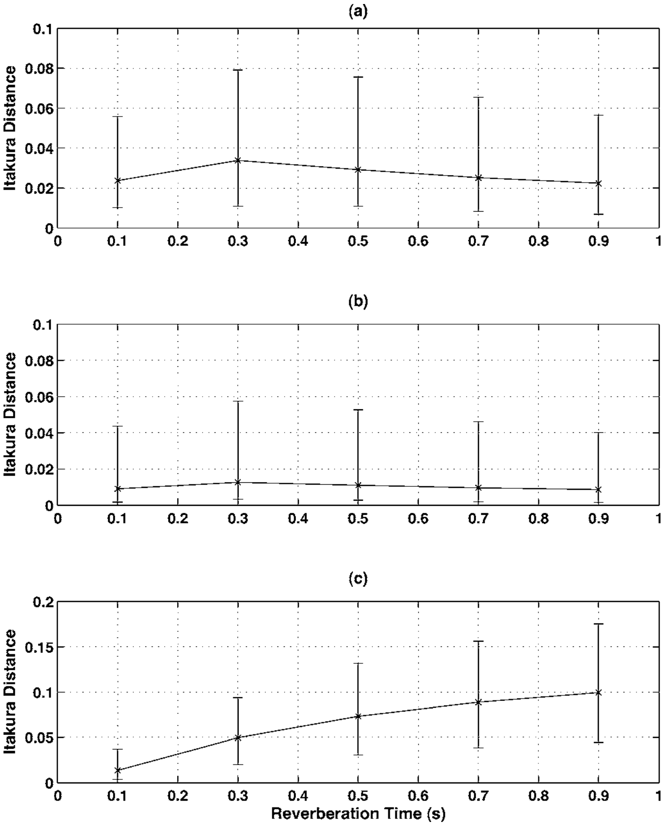

FIG. 4. Itakura distance vs reverberation time in terms of the AR coefficients for each individual outcome for (a) a single channel; (b) $M=7$ channels; and (c) the DSB output. Error bars indicate the maximum and minimum errors while crosses show the mean values.

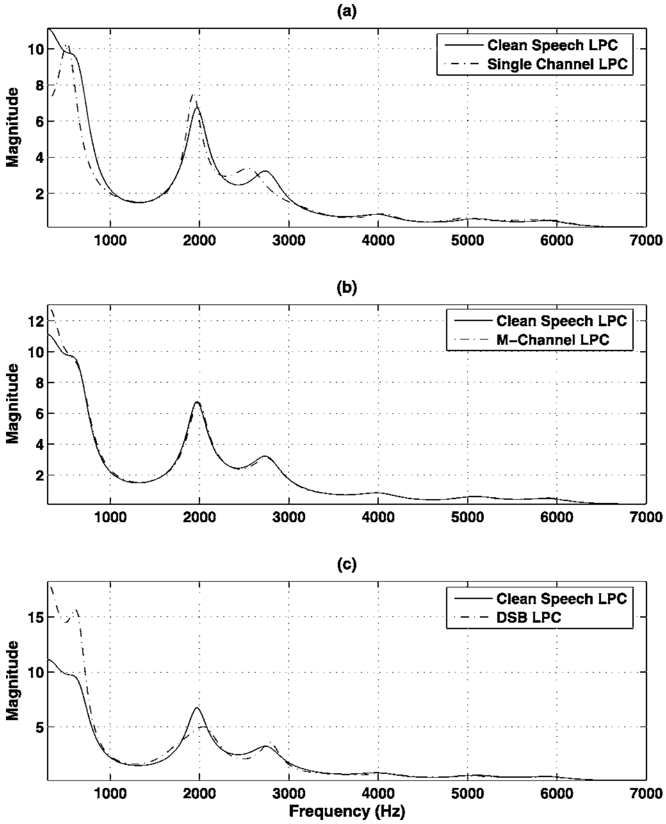

FIG. 5. Spectral envelopes calculated from the AR coefficients of clean speech compared with spectral envelopes obtained from the AR coefficients of (a) a single channel; (b) $M=7$ channels; and (c) the DSB output.

reverberation time, reaching an average of $d_I=0.099$ at $T_{60}=0.9$ with a maximum distance $d_I=0.175$ and minimum $d_I=0.045$ at that reverberation time. This result may appear counterintuitive; however, it conforms with the theoretical expression in (37) and will be clarified further with the following experiment. Figure 5 shows examples of the spectral envelopes from the AR coefficients obtained from reverberant observations using LPC for (a) single channel; (b) $M=7$ channels; and (c) the DSB output. Each case is compared to the resulting spectral envelope from clean speech.

C. Experiment 3

In line with the discussion in Sec. V B, the discrepancy in the estimated AR coefficients at the output of the DSB from those obtained with clean speech is governed mainly by the separation of the microphones. This final experiment demonstrates the effect of the separation between adjacent microphones on the expected AR coefficients obtained at output of a DSB. All the parameters of the room and source-microphone array were kept fixed while the separation, $\|\ell_m - \ell_{m+1}\|$, between adjacent microphones in the linear array was increased from 0.05 to 0.3 m in steps of 0.05 m. The results are shown in Fig. 6, where the Itakura distance is plotted against microphone separation for (a) the theoretical results calculated with (37) (dashed) and the simulated results (crosses) for the spatially expected AR coefficients at the output of the DSB, and (b) the AR coefficients for each individual outcome. Error bars indicate the maximum and the minimum errors while crosses indicate the mean value. It is seen from these results that the estimates at the output of the DSB become more accurate as the distance between the

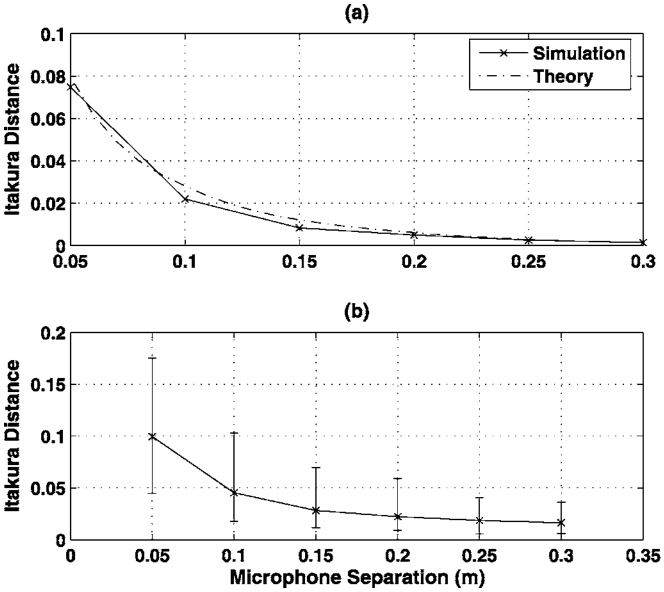

FIG. 6. Itakura distance vs microphone separation for (a) the theoretical results calculated with (37) (dashed) and the simulated results (crosses) for the spatially expected AR coefficients at the output of the DSB and (b) the AR coefficients for each individual outcome. Error bars indicate the maximum and the minimum errors while crosses show the mean value.

microphones is increased. At a microphone separation of $\|\ell_m-\ell_{m+1}\|=0.3$ m the results are comparable to the M-channel case both in terms of spatial expectation where an Itakura distance of $d_I=0.0041$ is observed and for the individual outcomes where the mean Itakura distance is $d_I=0.0164$ with the minimum and maximum distances being $d_I=0.06$ and $d_I=0.036$. This is due to the fact that the spatial correlation between microphones becomes negligible with increased microphone separation.

VII. CONCLUSIONS

We have used statistical room acoustic theory for the analysis of the AR modeling of reverberant speech. Investigating three scenarios, we have shown that, in terms of spatial expectation, the AR coefficients calculated from reverberant speech are approximately equivalent to those from clean speech both in the single-channel case and in the case when the coefficients are calculated jointly from an M-channel observation. Furthermore, it was shown that the AR coefficients calculated at the output of a delay-and-sum beamformer differ from the clean speech coefficients due to spatial correlation, which is governed by the room characteristics and the microphone array arrangement. This difference decreases as the distance between adjacent microphones is increased. It was also demonstrated that AR coefficients calculated jointly from the M-channel observation provide the best approximation of the clean speech AR coefficients at individual source-microphone positions and in particular when the microphone separation is small (<0.3 m). Thus, in general, the M-channel joint calculation of the AR coefficients is the preferred option where such an equivalence is important and specifically in the case of closely spaced microphones. Finally, the findings in this paper are of particular interest in speech dereverberation methods using prediction residual processing, where the main and crucial assumption is that reverberation mostly affects the prediction residual. Since most of these methods utilize microphone arrays for the residual processing, M-channel joint calculation of the AR coefficients should be deployed to ensure the validity of this assumption.

APPENDIX

Consider a function, $g(x_1,x_2,\ldots,x_n)$, of random variables[27] with mean values $E\{x_i\}=\mu_i$, which we write $g(x)$ for brevity. Letting $g'(x)=\partial(g(x))/\partial x_i|_{x=\mu}$, the Taylor series expansion of $g(x)$ about the mean, μ, is $g(x)=g(\mu)+\sum_{i=1}^{n}g'(\mu)(x_i-\mu_i)+\breve{g}(x)$, where $\breve{g}(x)$ are the second-order terms and above. All the partial derivatives up to the first order vanish[10] at $(\mu_1,\mu_2,\ldots,\mu_n)$ and, consequently, we can write $\mathcal{E}\{g(x)\}\cong g(\mathcal{E}\{x\})$ up to the zeroth order of approximation. In practice, the accuracy of this approximation will depend on the estimation of the mean value of the random variables.

[1] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," IEEE Trans. Acoust., Speech, Signal Process. **23**, 309–321 (1975).

[2] M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise," IEEE Trans. Acoust., Speech, Signal Process. **24**, 488–494 (1976).

[3] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," IEEE Trans. Acoust., Speech, Signal Process. **26**, 197–210 (1978).

[4] J. B. Allen, "Synthesis of pure speech from a reverberant signal," U.S. Patent No. 311,731 (1974).

[5] S. M. Griebel and M. S. Brandstein, "Microphone array speech dereverberation using coarse channel estimation," in *Proceedings of the IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol. **1**, pp. 201–204 (2001).

[6] B. Yegnanarayana and P. Satyanarayana, "Enhancement of reverberant speech using LP residual signal," IEEE Trans. Acoust., Speech, Signal Process. **8**, 267–281 (2000).

[7] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proceedings of the IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol. **1**, pp. 541–544 (2002).

[8] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proceedings of the IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol. **6**, pp. 3701–3704 (2001).

[9] H. Kuttruff, *Room Acoustics*, 4th ed. (Taylor & Francis, London, 2000).

[10] B. D. Radlović, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: Robustness results," IEEE Trans. Acoust., Speech, Signal Process. **8**, 311–319 (2000).

[11] P. A. Nelson and S. J. Elliott, *Active Control of Sound* (Academic, London, 1993).

[12] F. Talantzis and D. B. Ward, "Robustness of multichannel equalization in an acoustic reverberant environment," J. Acoust. Soc. Am. **114**, 833–841 (2003).

[13] S. Bharitkar, P. Hilmes, and C. Kyriakakis, "Robustness of spatial average equalization: A statistical reverberation model approach," J. Acoust. Soc. Am. **116**, 3491–3497 (2004).

[14] F. Talantzis, D. B. Ward, and P. A. Naylor, "Expected performance of a family of blind source separation algorithms in a reverberant room," in *Proceedings of the IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol. **4**, pp. 61–64 (Montreal, Canada) (2004).

[15] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," IEEE Trans. Speech Audio Process. **11**, 791–803 (2003).

[16] D. B. Ward, "On the performance of acoustic crosstalk cancellation in a reverberant environment," J. Acoust. Soc. Am. **110**, 1195–1198 (2001).

[17] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of Linear Prediction for dereverberation of speech," in *Proceedings of the Int. Workshop Acoust. Echo Noise Control*, pp. 99–102 (Kyoto, Japan) (2003).

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943–950 (1979).

[19] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," J. Acoust. Soc. Am. **80**, 1527–1529 (1986).

[20] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Technical Report, University College London(1987).

[21] J. Makhoul, "Linear Prediction: A tutorial review," Proc. IEEE **63**, 561–580 (1975).

[22] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time Processing of Speech Signals* (Macmillan, New York, 1993).

[23] *Microphone Arrays: Signal Processing Techniques and Applications*, edited by M. S. Brandstein and D. B. Ward (Springer, Berlin, 2001).

[24] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," IEEE Signal Process. Mag. **5**, 4–24 (1988).

[25] V. Golub and C. H. Gene, *Matrix Computations*, John Hopkins Series in the Mathematical Sciences, 3rd ed. (John Hopkins University Press, London, 1996).

[26] J. Makhoul, "Spectral linear prediction: Properties and applications," IEEE Trans. Acoust., Speech, Signal Process. **23**, 283–296 (1976).

[27] M. Kendall, A. Stuart, and J. K. Ord, *Kendall's Advanced Theory of Statistics*, 6th ed. (Hodder Arnold, 1994), Vol. **1**.