Single-Microphone LP Residual Skewness-Based Inverse Filtering of the Room Impulse Response

Saeed Mosayyebpour, *Student Member, IEEE*, Hamid Sheikhzadeh, *Senior Member, IEEE*, T. Aaron Gulliver, *Senior Member, IEEE*, and Morteza Esmaeili, *Member, IEEE*

***Abstract*—This paper presents a method based on higher order statistics (HOS), namely the normalized third-order moment (skewness), for blind estimation of the inverse filter of the room impulse response (RIR). Skewness is used as a measure of asymmetry, and a comprehensive comparison with the commonly used metric (kurtosis) is presented. It is shown that a sufficiently long linear predictive (LP) residual of the speech signal has an asymmetric pdf with sufficient skewness to be used as a score function for the HOS-based approach. The proposed algorithm is optimized for the inverse filter estimation problem. This optimization includes an efficient initialization for high reverberation intensities, enabling the method to be employed in highly reverberant rooms. The direct-to-reverberation ratio (DRR) as well as the equalized impulse response clearly show that our method can estimate the inverse filter even in highly reverberant environments. In addition, performance results using recorded background noise and in time-varying environments illustrate that our approach is applicable in real world situations. The proposed method is shown to be superior to the method by Wu and Wang, particularly in terms of reducing the coloration effect. Experiments under different acoustic conditions confirm the effectiveness of the proposed method for time delay estimation (TDE). Finally, the proposed algorithm is used as the first-stage of monaural segregation, and it is shown to improve the performance under different conditions.**

***Index Terms*—Higher order statistics (HOS), inverse filtering, linear prediction (LP) residual, room impulse response (RIR), single-microphone, skewness.**

I. Introduction

SPEECH signals recorded with a distant microphone in an enclosed room usually contain reverberation artifacts caused by reflections from walls, floors, and ceilings. This leads to spectral coloration causing a deterioration of the signal quality and intelligibility in many communication environments such as hands-free telephony and audio-conferencing [24]. Such deterioration often seriously degrades applications such as automatic speech recognition, speech separation, and source localization. These detrimental effects are magnified when the speaker to microphone distance increases. Inverse filtering of the room impulse response (RIR) is considered an attractive approach to overcome this problem [24]. It has been used in many applications such as sound reproduction, sound-field equalization, cross-talk cancellation, speech dereverberation, speech separation, and source localization.

Inverse filtering of the RIR is a difficult problem since the speech signal has colored and nonstationary characteristics, and the length of the RIR filter, typically approximated by a finite impulse response (FIR) filter, can be long. This necessitates a long inverse filter for the single channel case, which imposes a high computational burden especially when adaptive solutions are considered. Another inverse filtering challenge is the non-minimum phase nature of acoustic systems arising from the late energy in the RIR [2]. As a result, direct inversion of the RIR filter may result in poles near or outside the unit circle, leading to slow decay or instability, respectively. There are two distinct techniques used to invert the RIR filter: 1) minimum-phase, all-pass decomposition and 2) linear least squares. In the first technique, the speech signal is passed through the inverted stable minimum-phase component. The remaining all-pass component of the mixed phase RIR causes phase distortion (attributed to spikes in the group delay function) [2]. Thus, this magnitude-only equalization is inadequate. The least squares method has a delay included to compensate for zeros outside the unit circle. It provides better results, but suffers from residual artifacts with a tradeoff between filter length and quality [3].

The inverse filter of the RIR can be estimated using multi-channel or single-channel based approaches. Compared to single-microphone techniques which are limited to temporal processing, multiple microphones allow for spatial processing. Therefore, single-microphone inverse filtering is a significantly more challenging problem. Nonetheless, a number of single-microphone inverse filtering methods have been proposed in the literature since it may be that only one microphone is available.

Inverse filtering can be performed with or without channel estimation. The latter case estimates the RIR inversion directly while the former estimates the RIR followed by an inversion. Channel estimation errors will have an adverse effect on the RIR inversion, so the second approach has limited value in practical applications [10]. Bees *et al.* [4] employed a cepstrum-based method to estimate the RIR and used a least squares technique

Manuscript received April 24, 2011; revised September 03, 2011 and November 23, 2011; accepted January 16, 2012. Date of publication February 03, 2012; date of current version March 21, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Lewis Seltzer.

S. Mosayyebpour and T. A. Gulliver are with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8W 3P6, Canada (e-mail: saeedm@ece.uvic.ca; agullive@ece.uvic.ca).

H. Sheikhzadeh is with the Department of Electrical Engineering, Amirkabir University of Technology, Tehran 15875-4413, Iran (e-mail: hsheikh@aut.ac.ir).

M. Esmaeili is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8W 3P6, Canada, and also with the Department Mathematical Sciences, Isfahan University of Technology, Isfahan 84156, Iran (e-mail: mesmaeil@ece.uvic.ca; emorteza@cc.iut.ac.ir).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2186804

to invert it. They obtained satisfactory channel estimation results only for minimum-phase or mixed-phase responses with a few zeros outside the unit circle, which is unrealistic. Alternatively, there exist other approaches which do not require channel estimation. They can be categorized into two classes: higher order statistics (HOS) and information theory-based methods. The main idea behind both categories is that the reverberant speech samples, which are assumed to be delayed, weighted versions of independent and identically distributed (i.i.d.) speech samples, approximate a Gaussian distribution due to the central limit theorem.

Gillespie *et al.* [5], proposed a HOS-based approach which determines an adaptive inverse filter by maximizing the fourth-order statistic (kurtosis), of the linear prediction (LP) residuals. This algorithm is computationally efficient for moderate reverberation intensities. Tonelli and Mitianoudis [6] proposed a maximum-likelihood approach to maximize the LP residual kurtosis so that the sensitivity of the method in [5] to outlier values is decreased. In [7], Wu and Wang proposed a single-channel frequency-domain approach which provides reasonable results when the reverberation time is bounded in the range 0.2 to 0.4 s. While satisfactory inverse channel estimation accuracy can be obtained with reasonable computational complexity, this method cannot be used when the reverberation intensity is high.

From an information theory point of view, the reverberant speech can be viewed as redundant samples and the inverse filter can be obtained by minimizing the statistical dependency between samples. This concept was established by Bell and Sejnowski [8], where the mutual information between samples is minimized. They considered all HOS of the reverberant speech signal rather than only the fourth-order. The HOS of the reverberant speech were obtained by a nonlinear transformation of the signal [8]. The result was then weighted via maximizing the entropy of the transformation output. This is because maximizing the entropy corresponds to minimizing the mutual information. Erdogmus *et al.* [9] proposed a method based on this approach which minimizes the entropy of the reverberant speech and does not depend on *a priori* knowledge of the input signal. The main drawback of the methods in this category is their high computational complexity, which means they cannot be used in moderate to highly reverberant rooms. Satisfactory results are obtained only for low reverberation intensities.

From the above discussion, it is clear that blind single-microphone inverse filtering of the RIR for high reverberation intensity conditions with low computational complexity and direct estimation of the inverse channel is a very challenging and practical research problem. In this paper, an HOS-based algorithm for blind inverse filtering of the RIR is proposed. We show that the proposed method is more effective than methods based on fourth-order moments, and that it can be applied in a highly reverberant room with moderate background noise. Our contributions are listed below.

1) We show that a sufficiently long speech LP residual signal (e.g., more than 32 ms), has an asymmetric probability distribution function (pdf) with high skewness as a measure of asymmetry. This allows the use of the third-order moment of the input speech signal in blind deconvolution.
2) We propose an adaptive gradient-ascent algorithm for the input LP residual of a reverberant speech signal based on skewness instead of the commonly used metric (kurtosis).
3) We optimize the algorithm for implementation. This includes an effective algorithm for estimating the expected value of the feedback function, and an efficient procedure for filter initialization, which can be used with very high reverberation times (above 2 s).
4) A comprehensive comparison between skewness and kurtosis-based adaptive methods is given which shows the performance improvement using skewness.
5) We investigate three different applications of our inverse filtering method, namely single channel dereverberation, time delay estimation (TDE), and monaural speech segregation. The results show that our inverse filtering approach provides good performance for all three applications.

The remainder of this paper is organized as follows. In Section II, we investigate the reverberant speech characteristics in the LP residual domain. Section III describes blind inverse filtering of the RIR. The proposed algorithm including model description and implementation issues are also given. Section IV presents the performance results which demonstrate the effectiveness of the proposed method in highly reverberant rooms with moderate noise levels. Three applications are investigated to evaluate the performance of our approach. Finally, some concluding remarks are provided in Section V.

II. Reverberant Speech in the LP Residual Domain

In this section, we examine the characteristics of reverberant speech in the LP residual domain. This motivates our approach to speech processing for RIR inversion. The following model is considered for reverberant speech

$$x[n] = s[n] + \sum_{k=1}^{K} b_k s[n - n_k] = s[n] * h_r[n] \qquad (1)$$

where $s[n]$ is the clean speech signal, b_k is the relative amplitude of the reflection arriving after a delay of n_k samples, K is the number of reflections, and $h_r[n]$ is the time-invariant RIR model. For the purposes of analysis, the noise is assumed to be below perceivable levels and therefore is ignored.[1]

In order to reduce the effects of the vocal tract filter on the inverse filtering, pre-whitening based on LP analysis is performed. This topic has been discussed extensively in the literature (e.g., [16]). The LP residual of the reverberant speech is then obtained using an LPC filter of length 10 with a 32-ms frame size.

A. Characteristics of the LP Residual Signal for Reverberant Speech

The LP residual signal samples exhibit a Gaussian-like pdf especially in the reverberant tail regions, and hence the entropy is high [11]. Therefore, comparing the pdf of the LP residual of clean speech with that of the reverberant speech, it is expected that the pdf for the LP residual of the clean speech will be more outlier-prone and asymmetric.

[1] However, the performance of the proposed method will be evaluated in noisy conditions.

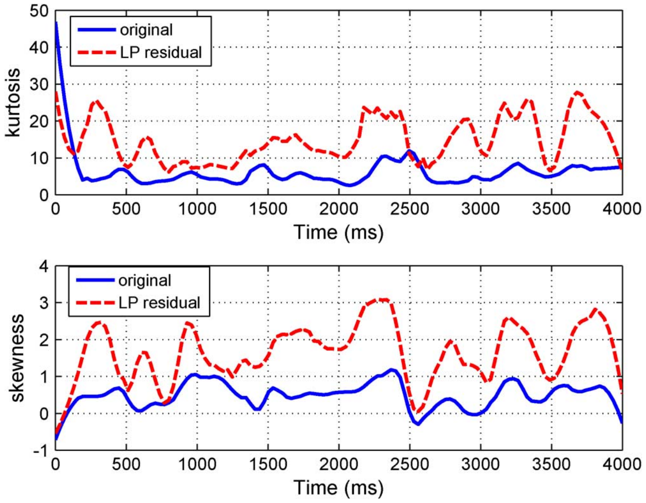


Fig. 1. Comparison of the kurtosis (upper) and skewness (lower) of the LP residual signal and the original speech signal (frame size $= 32$ ms).

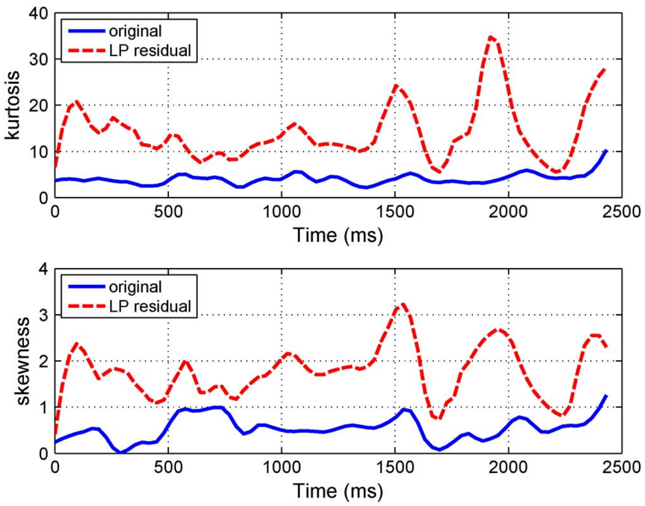


Fig. 2. Comparison of the kurtosis (upper) and skewness (lower) of the LP residual and original signals for only the voiced parts (frame size $= 32$ ms).

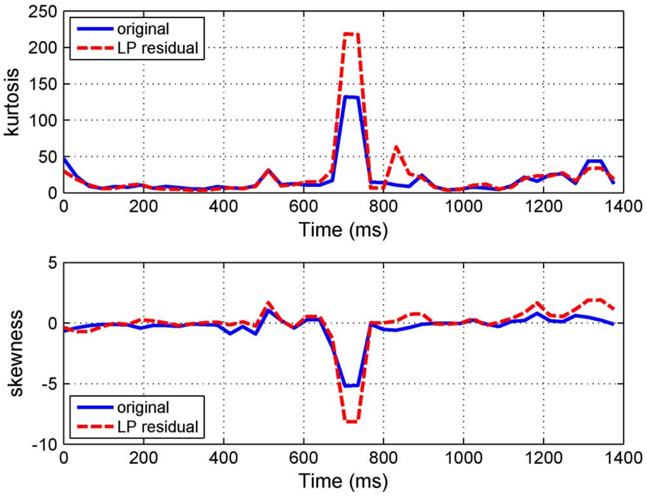


Fig. 3. Comparison of the kurtosis (upper) and skewness (lower) of the LP residual and original signals considering only the unvoiced parts (frame size $=$ 32 ms).

The skewness (third standardized moment), is a measure of the asymmetry of the data around the sample mean and is zero for a normal (Gaussian) distribution. The kurtosis (fourth standardized moment), is based on high-order statistical moments and so is a good indicator of how outlier-prone a distribution is. The skewness (γ_1) and kurtosis (γ_2) are defined as

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \tag{2}$$

and

$$\gamma_2 = \frac{\mu_4}{\sigma^4} \tag{3}$$

where μ_3 and μ_4 are the third and fourth moments about the mean, and σ is the standard deviation, respectively.

B. Evaluation of Kurtosis and Skewness as Measures of Gaussianity

To measure the Gaussianity of a signal, score functions based on higher order central moments are commonly used. Second-order statistics such as the autocovariance of the signal and the corresponding power spectrum do not provide an indication of the Gaussianity, and thus higher-order statistics are needed. Moments of order greater than four are rarely considered because they provide no real benefits over third- and fourth-order moments [14]. The kurtosis (normalized fourth-order moment) was used in [5]–[7] to construct a score function to estimate the RIR inverse filter. While this criterion indicates how outlier-prone a distribution is, the skewness (normalized third-order moment), is a measure of the asymmetry of the data around the sample mean. Third-order moments are rarely considered for blind deconvolution because the signals usually have a symmetric distribution with zero skewness, prohibiting the use of odd-order moments. However, it was recently shown (e.g., in [13]–[15] and [21]), that skewness can be used in blind deconvolution when the input signal has an asymmetric pdf.

It has been demonstrated that the distribution of speech samples is well described by a Laplacian distribution (LD) (or more generally a super Gaussian pdf) [22]. These symmetric distributions prohibit the use of skewness as a criterion to construct the score function for deconvolution. However, the excitation signal of speech sources has a distribution with higher peakedness and asymmetry, and lower entropy compared to the original speech signal. This results in higher kurtosis and skewness because the excitation signal can be considered as a combination of a periodic train of impulsive signals (for voiced segments) and noise-like signals (for unvoiced segments). The periodic impulse train of the LP residuals has a distribution which is far from Gaussian compared to the original speech signal, while the noise-like signal in unvoiced segments has a Gaussian-like distribution. It is well-known that a speech signal contains mostly voiced components. As a result, *a sufficiently long LP residual of the speech signal has lower entropy and higher kurtosis and skewness compared to the original speech signal*. This was confirmed experimentally using large amounts of speech data from the TIMIT database (sampled at 16 kHz), one example of which is shown in Figs. 1–3 for a frame size of 32 ms. The kurtosis

TABLE I
AVERAGE KURTOSIS AND SKEWNESS OF THE LP RESIDUALS FOR DIFFERENT FRAME SIZES

Frame Size	Kurtosis	Skewness
2 ms	4.92	0.55
4 ms	6.17	0.69
8 ms	7.30	0.76
16 ms	9.54	0.88
32 ms	13.41	**1.11**
64 ms	19.29	1.59
128 ms	23.7	1.96
256 ms	28.98	2.42
512 ms	32.05	2.55
1024 ms	41.51	2.88
2048 ms	58.59	3.42

and skewness of the LP residual and the original speech signal are shown in Fig. 1, while the corresponding voiced and unvoiced parts are shown in Figs. 2 and 3, respectively. These figures show that the LP residual signal for the voiced parts has greater skewness and a more peaked pdf than the original voice signal. This is not true for the unvoiced parts since they have a Gaussian-like pdf. Since a typical speech signal is dominated by the voiced parts, the skewness and kurtosis are higher for the LP residuals of the entire segment containing speech, as illustrated in Fig. 1.

In addition, speech segments with a longer duration tend to have a pdf which resembles a Gamma distribution or generalized Gaussian distribution [22], leading to higher third-order and fourth-order central moments. This also holds for the LP residual signal. Table I presents the kurtosis and skewness of the LP residual of the speech signal corresponding to Figs. 1–3 for different frame sizes. These results clearly indicate that increasing the frame size results in higher kurtosis and skewness. As can be seen in Table I, for a frame of 32 ms or greater, the mean skewness is larger than one. Thus it is preferable to use a frame length of 32 ms or more to construct a score function based on skewness, as then the pdf of the LP residual for the original speech signal will be sufficiently asymmetric with sufficient skewness to allow its use in a score function for adaptive inverse filter construction.

As discussed previously, reverberation affects the pdf of the LP residual signal so that the kurtosis and skewness are both decreased. Moreover, the kurtosis and skewness possess the following property: the sum of $K + 1$ independent random variables results in a random variable with a kurtosis and skewness the same as the original kurtosis and skewness divided by $K+1$ and the square root of $K + 1$, respectively [12] [see (1)]. Our experimental results confirmed this property. For example, the skewness and kurtosis were computed for 32-ms frames every 16 ms of the LP residual signals (for the reverberant and clean speech of Figs. 1–3), and the results are shown in Fig. 4. In addition, the average kurtosis and skewness for different RIRs (with different reverberation times (RT60) and speaker-microphone distances (d)), were computed for a speech segment of length 1.76 s, and this is shown in Fig. 5. It is clear from the figure that the reverberation metrics (skewness and kurtosis) generally decrease as the signal degradation increases (increased RT60 or increased d).

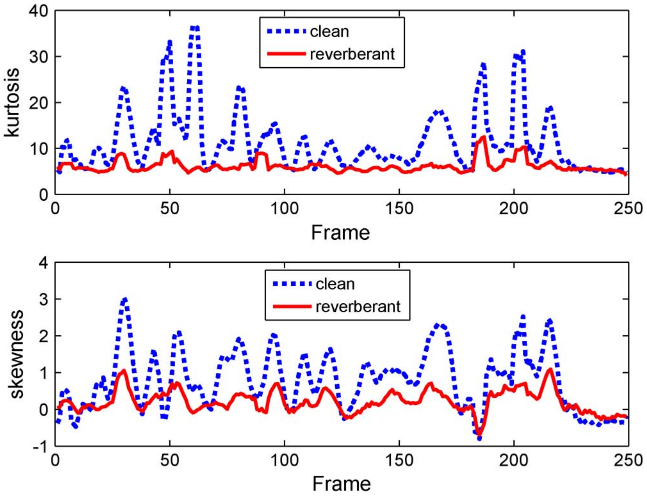


Fig. 4. LP residual kurtosis (upper) and skewness (lower) of the original and reverberant speech signals (frame size $= 32$ ms).

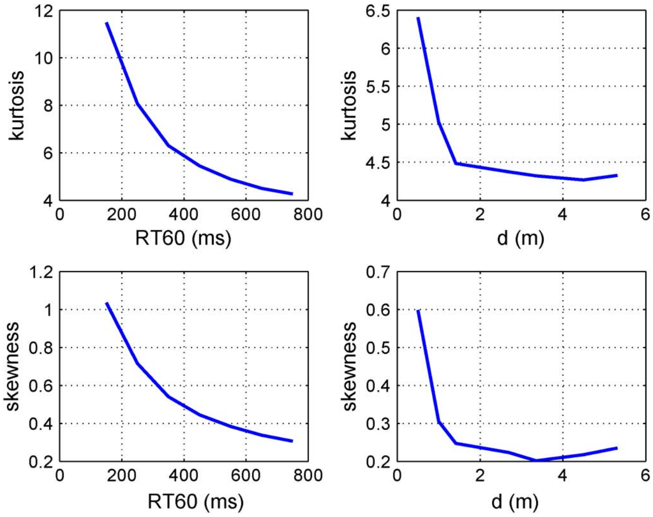


Fig. 5. LP residual skewness and kurtosis for different RIR values: for the left plots $\mathrm{d} = 0.5$ m, and for the right plots $\mathrm{RT60} = 300$ ms.

Finally, we determined that kurtosis and skewness are appropriate metrics, i.e., the greater the kurtosis/skewness, the more accurate the estimated inverse filter is. This was established via simulation results which show that this property holds, and we present one example here to illustrate this. Inverse filter estimation for a RIR (with $\mathrm{RT60} = 400$ ms and $\mathrm{d} = 2$ m), was performed using both skewness-based and kurtosis-based methods. This was done for different inverse filter lengths and initialization filters. The estimated inverse filter was used to obtain the inverse-filtered signal, and the average LP residual skewness and kurtosis over 32-ms frame lengths was computed. The direct-to-reverberation ratio (DRR)[2] for the equalized impulse response versus the LP residual skewness and kurtosis is presented in Fig. 6. This figure shows that the performance of the inverse filter estimation (which is directly related to the DRR values), is proportional to both the skewness and kurtosis. Therefore, skewness and kurtosis are suitable measures of reverberation and can be used to estimate the inverse filter of the RIR. Our goal is to develop an adaptive algorithm that uses these in the estimation. In the next section, two online adaptive inverse filter

[2] The definition of DRR is presented in Section IV-A.

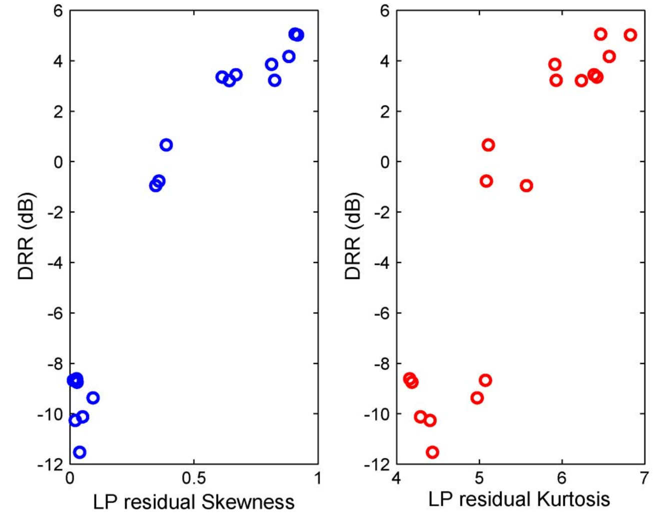


Fig. 6. Relationship between the DRR of the equalized impulse response, the LP residual skewness (left plot) and LP residual kurtosis (right plot), of the inverse-filtered signal. The inverse filters are estimated using different initialization values and inverse filter lengths for both approaches.

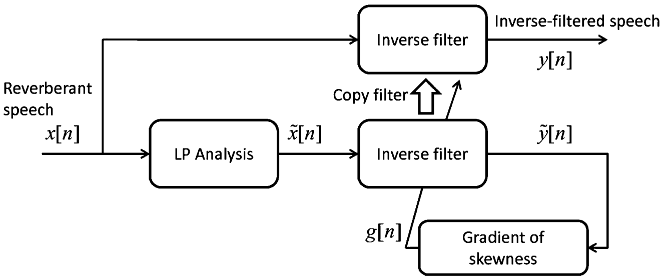


Fig. 7. Diagram of our proposed inverse filtering algorithm.

design algorithms based on skewness and kurtosis are developed and compared.

III. Blind Inverse RIR Filtering

As discussed in the previous section, the LP residual skewness and kurtosis for the original speech signal are higher than those for the reverberant signal. Maximizing these factors can be employed to obtain the inverse filter of the RIR. Here they are used to construct a score function for an adaptive gradient-ascent algorithm. The following sections present the adaptive algorithm for estimating a single-microphone RIR inverse filter.

A. System Model

Assume that an unknown speech signal $s[n]$ is received by a microphone in a room resulting in signal $x[n]$. The LP residual of the received reverberant speech is denoted by $\widetilde{x}[n]$. The adaptive inverse filter $\mathbf{h}$ is assumed to be an FIR filter of order L. After r iterations the filter has coefficients

$$\mathbf{h}^{(r)} = \left[h_0^{(r)}, h_1^{(r)}, \ldots, h_{L-1}^{(r)}\right]^T. \tag{4}$$

The LP residual of the reverberant speech after passing through the inverse filter is

$$\widetilde{y}_n = \left(\mathbf{h}^{(r)}\right)^T \widetilde{\mathbf{x}}_n \tag{5}$$

where $\widetilde{\mathbf{x}}_n$ is a vector of length L containing elements n to $n - L + 1$ of $\widetilde{x}[n]$.

The objective function of $\widetilde{y}_n$ is denoted as $\Psi(\widetilde{y}_n)$, and we wish to find the filter $\mathbf{h}$ that maximizes $\Psi(\widetilde{y}_n)$. Using a gradient-ascent algorithm, $\mathbf{h}$ is iteratively updated to maximize the objective function. The filter update rule is then

$$\mathbf{h}^{(r+1)} = \mathbf{h}^{(r)} + \mu \nabla_{\Psi(\mathbf{h}^r)} \tag{6}$$

where μ is the step-size and $\nabla_{\Psi(\mathbf{h}^r)}$ is the gradient of Ψ with respect to $\mathbf{h}^{(r)}$

$$\nabla_{\Psi(\mathbf{h}^r)} = \left[\frac{\partial \Psi}{\partial h_0^{(r)}}, \frac{\partial \Psi}{\partial h_1^{(r)}}, \ldots, \frac{\partial \Psi}{\partial h_{L-1}^{(r)}}\right]^T. \tag{7}$$

A block diagram of our algorithm is shown in Fig. 7. We consider two objective functions based on the kurtosis (normalized forth-order moment) and skewness (normalized third-order moment) of the filter output y_n

$$\text{kurtosis} \rightarrow \Psi^{(k)}(\widetilde{y}_n) = \frac{E\left\{\widetilde{y}_n^4\right\}}{E^2\left\{\widetilde{y}_n^2\right\}} \tag{8}$$

$$\text{skewness} \rightarrow \Psi^{(s)}(\widetilde{y}_n) = \frac{E\left\{\widetilde{y}_n^3\right\}}{E^{\frac{3}{2}}\left\{\widetilde{y}_n^2\right\}}. \tag{9}$$

The corresponding gradients with respect to $\mathbf{h}$ are

$$\frac{\partial \Psi^{(k)}(\tilde{y})}{\partial \mathbf{h}} = 4\left(\frac{E\{\tilde{y}^2\}E\{\tilde{y}^3\tilde{\mathbf{x}}\} - E\{\tilde{y}^4\}E\{\tilde{y}\tilde{\mathbf{x}}\}}{E^3\{\tilde{y}^2\}}\right) \tag{10}$$

$$\frac{\partial \Psi^{(s)}(\tilde{y})}{\partial \mathbf{h}} = 3\left(\frac{E\{\tilde{y}^2\}E\{\tilde{y}^2\tilde{\mathbf{x}}\} - E\{\tilde{y}^3\}E\{\tilde{y}\tilde{\mathbf{x}}\}}{E^{\frac{5}{2}}\{\tilde{y}^2\}}\right). \tag{11}$$

In the above equations, the time-dependency is omitted for simplicity. The gradients can be approximated by

$$\frac{\partial \Psi^{(k)}(\tilde{y})}{\partial \mathbf{h}} \approx 4\left(\frac{\tilde{y}^3E\{\tilde{y}^2\} - \tilde{y}E\{\tilde{y}^4\}}{E^3\{\tilde{y}^2\}}\right)\tilde{\mathbf{x}} = f\tilde{\mathbf{x}} \tag{12}$$

$$\frac{\partial \Psi^{(s)}(\tilde{y})}{\partial \mathbf{h}} \approx 3\left(\frac{\tilde{y}^2E\{\tilde{y}^2\} - \tilde{y}E\{\tilde{y}^3\}}{E^{\frac{5}{2}}\{\tilde{y}^2\}}\right)\tilde{\mathbf{x}} = g\tilde{\mathbf{x}} \tag{13}$$

where f and g are the feedback functions, and are both subsequently represented by q when there is no ambiguity. The kurtosis was previously used as a feedback function in [5]–[7]. Here we propose using g as the feedback function based on skewness and compare it with the results using f.

In order to calculate the feedback function, the corresponding expected values should be accurately estimated. To achieve this, the following procedure is employed.

1) The signal y is segmented into blocks of length N, denoted y_j.
2) y_j^2, y_j^3 and y_j^4 are computed for each block, denoting the second, third, and fourth powers of the elements of y_j, respectively.
3) The mean values of each block (y_j^2, y_j^3, and y_j^4) are computed, denoted by $M_j^{(2)}$, $M_j^{(3)}$ and $M_j^{(4)}$, respectively.
4) The expected values in (12) and (13) are replaced with the corresponding means $M_j^{(i)}$ and the estimated feedback

function values for each block are concatenated to obtain the feedback vector.

As direct use of the time-domain implementation may have slow or no convergence [5], a frequency-domain implementation of the adaptive filter is used. In this formulation, the reverberant speech signal $x[n]$ is segmented into blocks of length L. The blocks are increased to $2L$ samples by zero-padding, and a fast Fourier transform (FFT) of length $2L$ is computed for each block. The feedback function q is segmented into blocks of length $2L$ with L samples overlapping, and an FFT of length $2L$ is computed for each block. Denote the number of blocks by T. The adaptation in the frequency-domain for calculating the inverse filter is the same as in [7], and is given by

$$\mathbf{H}'^{(r+1)} = \mathbf{H}^{(r)} + \frac{\mu}{T}\sum_{i=1}^{T}\mathbf{Q}_i\widetilde{\mathbf{X}}_i^* \tag{14}$$

$$\mathbf{H}^{(r+1)} = \frac{\mathbf{H}'^{(r+1)}}{\left|\mathbf{H}'^{(r+1)}\right|} \tag{15}$$

where $\mathbf{H}^{(r)}$ is the FFT of the inverse filter $\mathbf{h}$ in the rth iteration. $\mathbf{Q}_i$ and $\widetilde{\mathbf{X}}_i$ denote, respectively, the FFT of q and $\widetilde{x}$ for the ith block. The superscript $*$ denotes complex conjugate. The inverse filter is initialized with a simple all-pass filter

$$\mathbf{H}^{(0)} = [1\,1\,1\ldots 1]^T. \tag{16}$$

Equation (15) ensures that the inverse filter is normalized. This is necessary to keep the algorithm numerically stable since an increasing $\widetilde{y}$ increases $\Psi(\widetilde{y}_n)$ without improving the inverse filter estimation, in which case the norm of $\mathbf{h}^{(r)}$ grows rapidly [15].

B. Performance Analysis of the Kurtosis and Skewness Objective Functions

The adaptation algorithm searches for local maximum points, i.e., points where $\nabla_{\Psi_{(\mathbf{h})}}$ is zero and the Hessian[3] is negative definite. The set of maximum points is a subset of the points on which the function surface has gradient $\nabla_{\Psi_{(\mathbf{h})}}$ equal to zero. The other set members are minimum or saddle points. The superscript r denoting the rth iteration of the inverse filter will be omitted for simplicity.

As described in Section III-A, the objective functions for kurtosis and skewness can be approximated by

$$\nabla_{\Psi^{(k)}_{(\mathbf{h})}} \approx 4\left(\frac{\tilde{y}^3 M^{(2)} - \tilde{y}M^{(4)}}{M^{(2)^3}}\right)\tilde{\mathbf{x}} \tag{17}$$

$$\nabla_{\Psi^{(s)}_{(\mathbf{h})}} \approx 3\left(\frac{\tilde{y}^2 M^{(2)} - \tilde{y}M^{(3)}}{M^{(2)^{\frac{5}{2}}}}\right)\tilde{\mathbf{x}} \tag{18}$$

respectively. Using the method of Lagrange multipliers [20], the stationary points of $\nabla_{\Psi_{(\mathbf{h})}}$ under unit-norm $\mathbf{h}$ are the stationary points of the Lagrangian function

$$\Phi_{(\mathbf{h},\lambda)} = \Psi_{(\mathbf{h})} - \lambda(\mathbf{h}^T\mathbf{h} - 1). \tag{19}$$

Setting the gradient of $\Phi_{(\mathbf{h},\lambda)}$ equal to zero gives the following expressions:

$$\nabla_{\Psi^{(k)}_{(\mathbf{h})}} - 2\lambda\mathbf{h} = \mathbf{0} \tag{20}$$

$$\nabla_{\Psi^{(s)}_{(\mathbf{h})}} - 2\lambda\mathbf{h} = \mathbf{0} \tag{21}$$

for kurtosis and skewness, respectively. Expanding (20) and (21), and using (17) and (18), leads to the following system of L nonlinear equations in the L parameters $\{h_m\}_{m=0,\ldots,L-1}$:

$$\begin{aligned}\frac{4}{M^{(2)^3}}\Bigg(M^{(2)}\Bigg(&\sum_{i=0}^{L-1} h_i^3 C(0,0,i-m)\\ &+3\sum_{i,j=0_{j\neq i}}^{L-1} h_i^2 h_j C(0,i-j,i-m)\\ &+\sum_{i,j,k=0_{j\neq i\neq k}} h_i h_j h_k C(i-j,i-k,i-m)\Bigg)\\ &-M^{(4)}\left(\sum_{i=0}^{L-1} h_i x_{n-i+N-1}x_{n-N+m+1}\right)\Bigg)\\ +2\lambda h_m = 0&\end{aligned} \tag{22}$$

$$\begin{aligned}\frac{3}{M^{(2)^{\frac{5}{2}}}}\Bigg(M^{(2)}\Bigg(&\sum_{i=0}^{L-1} C(0,i-m)b_i^2\\ &+\sum_{i,j=0_{j\neq i}}^{L-1} C(i-j,i-m)b_i b_j\Bigg)\\ &-M^{(3)}\left(\sum_{i=0}^{L-1} b_i x_{n-i+L-1}x_{n-L+m+1}\right)\Bigg)\\ +2\lambda b_m = 0&\end{aligned} \tag{23}$$

where $C(a,b,c) = (x_n x_{n+a} x_{n+b} x_{n+c})$, and $C(a,b) = (x_n x_{n+a} x_{n+b})$.

The highest polynomial degree in (22) is 3, which gives a Bezout [23] upper bound on the number of solutions, i.e., the number of stationary points on the function surface is 3^L. The highest polynomial degree in (23) is 2, resulting in only 2^L solutions. This was also shown in [14] for third- and fourth-order moment objective functions. Therefore, the objective function based on kurtosis has many more stationary points to which the adaptive algorithm can converge to. These stationary points contain not only the minima but also saddle points which can stall the adaptation. A lower order score function provides a *simpler* objective surface, which, in general, implies fewer saddle points [14]. Thus, when the reverberation time is high, the probability of converging to an undesirable saddle point for the objective function based on kurtosis is much higher than the one based on skewness. The following properties were observed during our experiments with the adaptation algorithms based on kurtosis and skewness.

- There is usually no single maximum point. Typically, many local maxima can exist which provide similar performance. The local maximum obtained depends on the ***length and initialization of the inverse filter h***.

[3] The Hessian is the square matrix of second-order partial derivatives of a function.

- There are also *undesirable* maxima which provide poor performance. The local maxima which correspond to both maximum kurtosis and maximum skewness are *desirable* maxima. The probability of converging to an *undesirable* maxima for the kurtosis-based function is much larger than that for the skewness-based function. This is because there are many more stationary points for kurtosis than for skewness, as discussed previously. For the kurtosis-based function, the local maxima can be classified into three types, those that correspond to only maximum kurtosis, those that correspond to both maximum kurtosis and maximum skewness, and those that correspond to maximum kurtosis and minimum skewness (maximum skewness in the negative direction). The latter two cases are *desirable* maxima, while the first corresponds to *undesirable* maxima. Note that, in the last case, the equalized impulse response has a negative peak.
- Generally, the convergence for the score function based on kurtosis $(q = f)$ is faster than that based on skewness $(q = g)$. This is because the gradient of the function based on the fourth-order moment is higher than that based on the third-order moment. This is illustrated in Fig. 8 where μ is chosen to be 3×10^{-9}. The solid lines are the kurtosis and skewness when the objective function is based on kurtosis, while the dashed lines are the kurtosis and skewness when the objective function is based on skewness. In this example, the equalized speech signal in each iteration is segmented into blocks of 32 ms, then the criteria is calculated for each block. The average for the blocks (over the entire 24-s speech segment), is computed for each iteration. As can be seen from the figure, the algorithm based on kurtosis has a higher convergence rate (here almost double).
- Despite the higher convergence rate for the objective function based on kurtosis, the resulting estimation is typically poorer than that for the function based on skewness. This is because the local maxima for the function based on kurtosis are more often lower than those for the function based on skewness. This is also clear from Fig. 8 where the algorithm based on skewness $(q = g)$ converges to a higher average value. As a result, the adaptive algorithm based on skewness converges to an inverse filter which moves the signal further away from a Gaussian distribution.
- The main difference in computational complexity between the algorithms is due to the feedback functions f and g. Since f requires one more multiplication than g (six for f and five for g), the computational complexity for the function based on skewness is lower. As discussed in the Introduction, the algorithm presented here is more efficient than information theory-based techniques.
- Additive noise perturbs the objective function surface. We assume the noise to be Gaussian, zero mean, white and independent of the speech signal. The received speech signal is then

$$\mathbf{z}_n = \mathbf{x}_n + \mathbf{v}_n \tag{24}$$

where $\mathbf{x}_n$ is the reverberant speech and $\mathbf{v}_n$ is the noise signal. The inverse-filtered speech signal with the estimated inverse filter $\mathbf{h}$ is

$$\begin{aligned} r_n &= \mathbf{h}^T(\mathbf{z}_n) \\ &= \mathbf{h}^T(\mathbf{x}_n + \mathbf{v}_n) \\ &= d_n + \vartheta_n. \end{aligned} \tag{25}$$

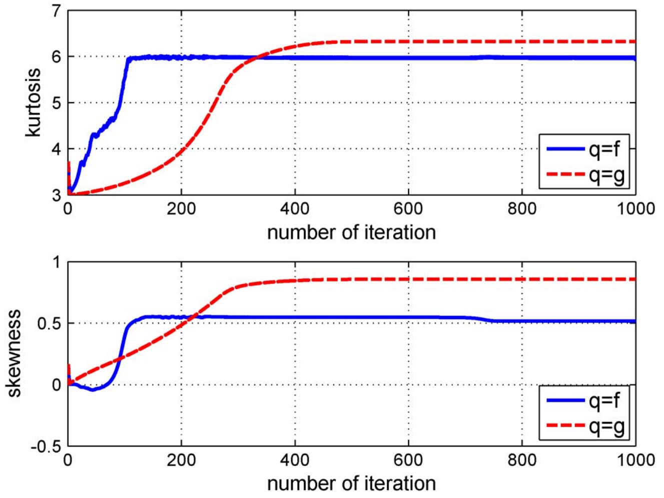


Fig. 8. Average kurtosis (upper) and skewness (lower) of a RIR with $\mathrm{RT60} = 200$ ms for each iteration when the score function is based on kurtosis $(q = f)$ and skewness $(q = g)$, and $\mu = 3 \times 10^{-9}$.

The speech signal is assumed to have an approximately zero mean. Using the above equations and assumptions, the objective functions (8) and (9) can be calculated as

$$\Psi^{(k)}(\widetilde{r}_n) = \frac{E\left\{\tilde{d}_n^4\right\} + \left(\sigma_v^2\right)^2 |\mathbf{h}|^4 + 6\sigma_v^2 E\left\{\tilde{d}_n^2\right\} |\mathbf{h}|^2}{\left(E\left\{\tilde{d}_n^2\right\} + \sigma_v^2\right)^2} \tag{26}$$

$$\Psi^{(s)}(\tilde{r}_n) = \frac{E\left\{\tilde{d}_n^3\right\}}{\left(E\left\{\tilde{d}_n^2\right\} + \sigma_v^2\right)^{\frac{3}{2}}} \tag{27}$$

where σ_v is the standard deviation of the noise. As can be seen from (26) and (27), $\Psi^{(k)}$ and $\Psi^{(s)}$ are both sensitive to noise due to their normalization (denominator of the equations). In addition, the Gaussian noise introduces two additional terms in the numerator of (26). The second term in (26) does not depend on $\mathbf{h}$ and will therefore not change the location of the stationary points. This is because the inverse filter is normalized during adaptation (15). However, the third term depends on $\mathbf{h}$ through d_n, so it can alter the location of the stationary points. As a result, the kurtosis-based inverse filtering is more sensitive to additive noise. To evaluate the effect of additive noise on the skewness-based approach, the feedback function g' is calculated by taking the gradient of $\Psi^{(s)}$ in (27) with respect to $\mathbf{h}$ giving

$$\frac{\partial \Psi^{(s)}(\widetilde{r})}{\partial \mathbf{h}} = 3\frac{E\{\tilde{d}^2\tilde{\mathbf{x}}\}\left(E\{\tilde{d}^2\} + \sigma_v^2\right) - E\{\tilde{d}\tilde{\mathbf{x}}\}E\{\tilde{d}^3\}}{\left(E\{\tilde{d}^2\} + \sigma_v^2\right)^{\frac{5}{2}}} \tag{28}$$

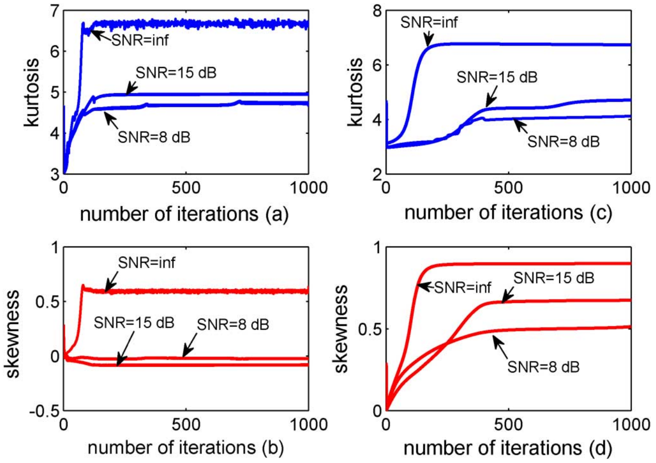


Fig. 9. Average kurtosis (upper) and skewness (lower) for each iteration when the score function is based on kurtosis (a)-(b) and skewness (c)-(d), SNR $= \infty$ (inf), 8 and 15 dB, with $\mu = 3 \times 10^{-9}$.

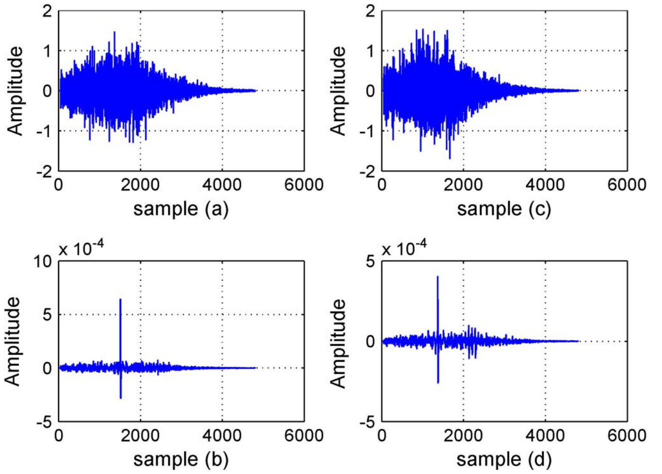


Fig. 10. Equalized impulse response estimated in different noisy conditions (a)-(b) for SNR $= 15$ dB and (c)-(d) for SNR $= 8$ dB, when the score function is based on kurtosis (upper) and skewness (lower).

$$\approx 3\frac{\tilde{d}^2\tilde{\mathbf{x}}\left(E\{\tilde{d}^2\}+\sigma_v^2\right)-\tilde{d}\tilde{\mathbf{x}}E\{\tilde{d}^3\}}{\left(E\{\tilde{d}^2\}+\sigma_v^2\right)^{\frac{5}{2}}}=g'\tilde{\mathbf{x}}. \tag{29}$$

Again, the time-dependency is omitted for simplicity. The feedback function for the skewness-based approach under noisy conditions is then

$$g' = g\left(1-\frac{1}{1+\frac{E\{\tilde{d}^2\}}{\sigma_v^2}}\right)^{\frac{5}{2}} + \frac{3\tilde{d}^2}{(\sigma_v^2)^{\frac{3}{2}}\left(1+\frac{E\{\tilde{d}^2\}}{\sigma_v^2}\right)} \tag{30}$$

where g is the noise free feedback (13). For the LP residual of the speech signal $\tilde{d}_n$, it has been observed that

$$\frac{E\{\tilde{d}^2\}}{\sigma_v^2} \ll 1 \Rightarrow g' < g. \tag{31}$$

Thus, the feedback function based on skewness is smaller under noisy conditions compared to the noise free case, so convergence is slower in the presence of additive noise. As an example, the results of one of our experiments is depicted in Figs. 9 and 10. For this case, the inverse filter for a RIR with $\mathrm{RT60} = 300$ ms and $\mathrm{d} = 1$ m is estimated for SNR $= 8$ and 15 dB, and also no noise, using the skewness and kurtosis based approaches. The average values of the kurtosis and skewness for each iteration with a frame size of 32 ms are given in Fig. 9. This shows that the performance of both methods with no noise $(\mathrm{SNR} = \infty)$ is almost the same due to convergence to *desirable* local maxima (maximizing both the kurtosis and skewness). However, the kurtosis-based method does not converge to *desirable* local maxima under noisy conditions since the skewness does not increase as shown in Fig. 9(a) and (b). In contrast, the skewness-based method converges to *desirable* local maxima as shown in Fig. 9(c) and (d), but the convergence rate is decreased compared with the noise free case. The equalized impulse responses obtained for the noisy conditions are depicted in Fig. 10 and also support this point. Based on the above discussion, the following conclusions can be made.

1) Additive noise alters the positions of the stationary points and thus can degrade the performance of adaptive inverse filtering methods. This is to be expected since additive Gaussian noise shifts the speech signals towards a Gaussian pdf.
2) The skewness-based approach is more robust to noise compared to the kurtosis-based approach. However, it converges to worse local maxima compared to the noise free case.
3) The convergence rate of the skewness-based approach is lower in noise compared to the noise free case.

C. Implementation Issues

Based on the previous arguments, only $q = g$ is considered in the remainder of the paper.

Specification of N: In order to calculate the feedback function, N should be chosen appropriately. N is the number of signal samples (e.g., for $\tilde{y}[n]$) such that the mean can be used as an estimate of the expected value (e.g., $E\{\tilde{y}[n]\}$). This indicates that the proper choice of N depends on the signal stationarity. It is well known that a speech signal has short-term stationary but long-term nonstationary. Considering the typical properties of speech, N is commonly set to a span of 32 ms or less. On the other hand, as discussed in Section II-B, the frame size of the LP residuals should exceed a minimum (e.g., 32 ms), in order to use skewness as a criterion for constructing the objective function. Therefore, N is chosen to span 32 ms (512 samples for $f_s = 16$ kHz).

Specification of L: Generally, the length of the inverse filter L depends on the reverberation time which is directly proportional to the RIR length (approximately $\mathrm{RT60}\ (\mathrm{s}) \times f_s$ (Hz) samples). For high reverberation times, a large value of L is required. On the other hand, to limit the computational complexity, a small value should be used. Thus, the best value is the minimum that provides reasonable performance. By extensive simulations, we have determined good choices for the inverse filter length L for different values of RT60. These results, given in Table II, show that a single value of L can be employed for a

TABLE II
INVERSE FILTER LENGTHS FOR DIFFERENT RT60 VALUES

RT60 (ms)	150-500	600-1100	1200-4000
L (samples)	2000	4000	6000

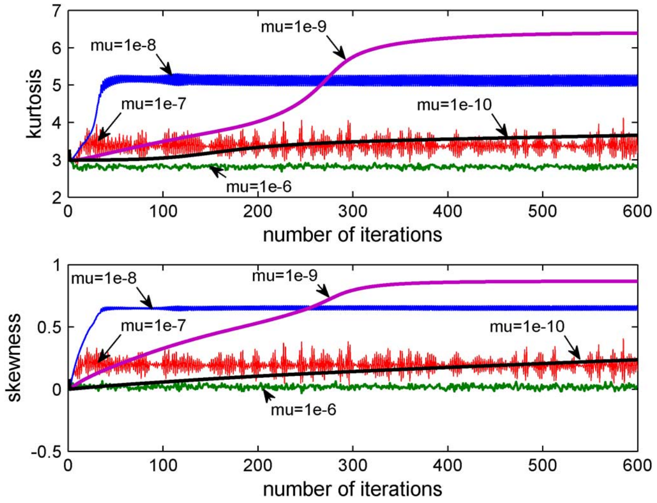


Fig. 11. Average kurtosis (upper) and skewness (lower) for each iteration when the score function is based on skewness using different step sizes μ.

wide-range of RT60 values, and the performance is not very sensitive to variations in L. Note that convolving the inverse filter with the RIR results in an impulse-like response with a maximum at a delay proportional to L.

Specification of μ: μ is the step size in the adaptive inverse filter algorithm that controls the convergence rate. Similar to LMS-like adaptive algorithms, there is a tradeoff between convergence rate and steady-state error. The smaller the value of μ, the lower the convergence rate but the better the performance. The LP residual kurtosis and skewness were calculated for each iteration in the skewness-based approach (proposed method) averaged over 32-ms frames. The results of one experiment with $\mathrm{RT60} = 500$ ms and $\mathrm{d} = 2$ m are shown in Fig. 11 (the results for other RIRs are similar). In this figure, the convergence is shown for five values of μ. When μ is lower than 10^{-9} a *desirable* local maximum (maximum values of kurtosis and skewness) is not obtained and the steady-state error is unacceptable, while higher values do not provide a reasonable convergence rate. As a result, the best value of μ is on the order of 10^{-9}.

The number of iterations: The adaptive inverse filter algorithm is executed recursively until there is no significant change in the criterion being maximized (skewness). Further, convergence depends on the step size μ, and selecting a smaller value results in a higher number of iterations. For example, Fig. 8 shows that about 300 iterations are required for convergence when skewness is used as the score function $(q = g)$ with $\mu = 3 \times 10^{-9}$, which is about half that required when kurtosis is used. This value was confirmed over a large number of simulations.

Inverse filter initialization: As mentioned in Section III-A, the simplest method for initializing the inverse filter is to use (16). This will provide good results for sufficiently large data sizes and number of iterations. We propose an initialization procedure to reduce the necessary data size and improve the algorithm performance. The inverse filter is estimated for RIRs with different values of RT60 using one long (e.g., 24 s), arbitrary speech signal. These estimates are stored to initialize the inverse filter. They are obtained for RIRs with a given RT60 and approximately the same speaker-microphone distance, but arbitrary speaker-microphone positions.

TABLE III
AMOUNT OF SPEECH DATA REQUIRED TO DERIVE THE INVERSE RIR FILTER

RT60 (ms)	150-600	700	800	900-1000	1100-4000
data size (s)	4-5	16	17	18	20

Initialization begins by first estimating the reverberation time (RT60) of the RIR blindly for a given signal. This reverberation time is used to select a stored filter to initialize the inverse filter. The RT60 estimation can be performed using a variety of methods (e.g., the approach in [17]). Here we assume the RT60 values are known for the purpose of choosing the initial filters.

Note that this initialization procedure can be performed once for all RIRs that require the same inverse filter length L. For example, from Table II all RIRs with reverberation times from 600 to 1100 ms need an inverse filter of length $L = 4000$.

The simulation results obtained show that the simple blind initialization of (16) provides satisfactory results for RT60 only in the range 150–1200 ms. Thus, the proposed initialization procedure is crucial for RT60 in the range 1200–4000 ms (i.e., very high reverberation conditions). The proposed initialization procedure has the following advantages:

- reduced estimation data size;
- reduced number of adaptation iterations;
- can be employed for RT60 beyond 1200 ms up to 2000 to 4000 ms.

Data size specification: As the length of the inverse filter L is proportional to RT60, and hence to the FFT size in the frequency domain, it is reasonable that for higher RT60 values, higher data sizes should be used. Further, as the frequency domain inverse filter is calculated by averaging over blocks, and the number of blocks T is related to the data size, for satisfactory results, a minimum number of blocks is needed. Thus, the data size is proportional to the inverse filter length or reverberation time. The required data sizes for different values of RT60 determined by simulation, are shown in Table III.

IV. PERFORMANCE RESULTS

This section presents the results of an extensive experimental investigation of the proposed inverse filtering method. The performance is obtained by convolving segments of 20-s clean speech signals (from four male and four female speakers), from the TIMIT database which are sampled at 16 kHz. The RIRs were constructed using the image method [1]. The microphone is omnidirectional, and is located in a rectangular room with dimensions $5 \times 4 \times 6$ m^3.

A. Inverse Filter Estimation

Generally, the reverberation time and microphone-speaker distance are nearly independent factors which both increase the reverberation [19]. While the first factor smears the speech spectra, the second causes distortion called coloration. Thus, we consider two sets of RIRs: one with different RT60 values and a fixed speaker-microphone distance of $\mathrm{d} = 2$ m, and

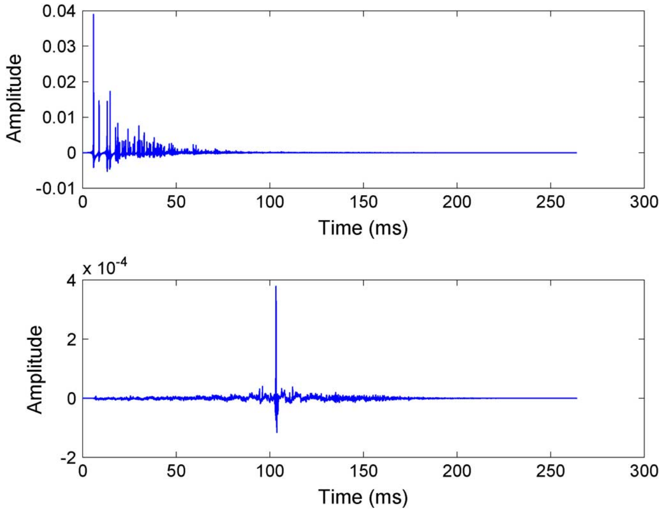


Fig. 12. Actual RIR with $\mathrm{RT60} = 200$ ms (upper plot), and the equalized impulse response using the proposed method (lower plot).

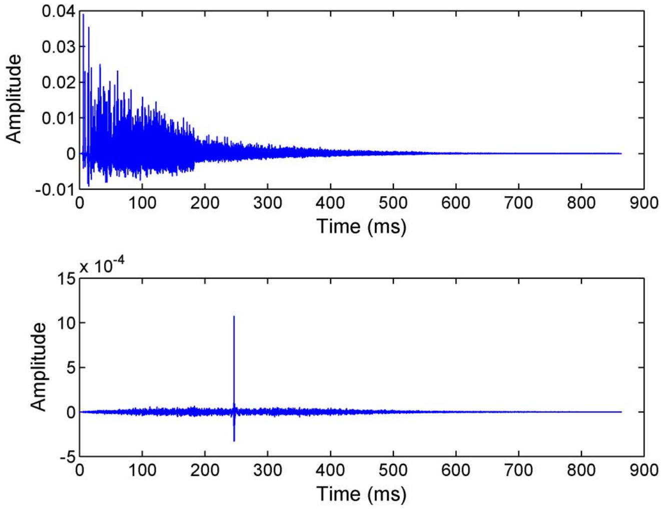
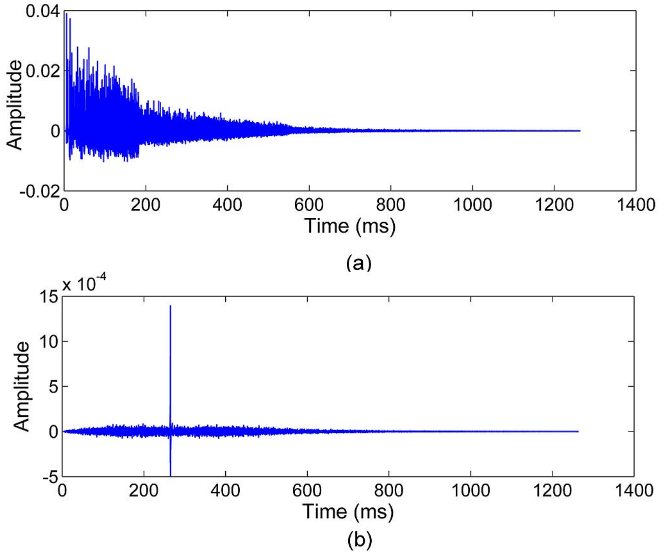


Fig. 13. Actual RIR with $\mathrm{RT60} = 800$ ms (upper plot), and the equalized impulse response using the proposed method (lower plot).

another with different speaker-microphone distances and an RT60 of 800 ms. The initialization procedure introduced in Section III-C was employed only for RT60 values above 2000 ms. For smaller values, the allpass initialization given by (16) is sufficient. The performance is examined via the equalized impulse response and an objective evaluation. The equalized impulse response was obtained by convolving the estimated inverse filter with the corresponding RIR. Four RIRs and the corresponding equalized impulse responses are shown in Figs. 12–15. It is evident from these figures that the proposed algorithm can effectively estimate the inverse filter, as the equalized impulse responses are all impulse-like. However, there are a number of replicas of the direct signal before (pre-echoes) and after (late impulses) its arrival. The late impulses smear the speech spectra and the pre-echoes introduce annoying temporal characteristics and deteriorate the speech quality. In this case, the inverse filtering should be combined with an appropriate secondary method to mitigate both the remaining late-impulses and the pre-echoes. On the other hand, as will be shown in the following sections, our inverse filtering can be used alone in some applications where reverberation is the primary concern.



(a)

(b)

Fig. 14. Actual RIR with $\mathrm{RT60} = 1200$ ms (upper plot), and the equalized impulse response using the proposed method (lower plot).

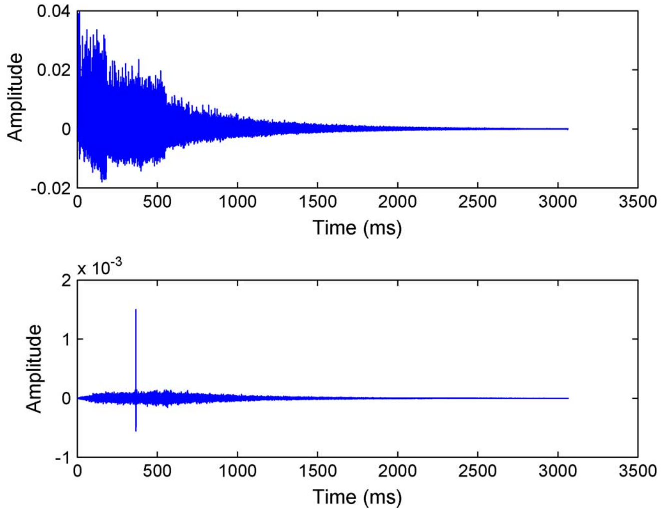


Fig. 15. Actual RIR with $\mathrm{RT60} = 3000$ ms (upper plot), and the equalized impulse response using the proposed method (lower plot).

The DRR is commonly used for performance evaluation [19], and is given by

$$\mathrm{DRR} = 10\log_{10}\left(\frac{\sum_{n=n_d-n_0}^{n_d+n_0} h^2(n)}{\sum_{n=0}^{n_d-n_0} h^2(n) + \sum_{n=n_d+n_0}^{\infty} h^2(n)}\right) \tag{32}$$

where the direct path signal arrives at sample (time) n_d. The direct path energy is computed 8 ms ($n_0 = 128$ samples) around the maximum of the impulse (or equalized) response. The DRR measures the direct path energy divided by the total reflective energy. Fig. 16 shows the results where DRR^{RIR}, DRR^{Wu} and DRR^{prop} denote, respectively, the DRR of the RIR, the equalized response using the approach of Wu and Wang [7], and our

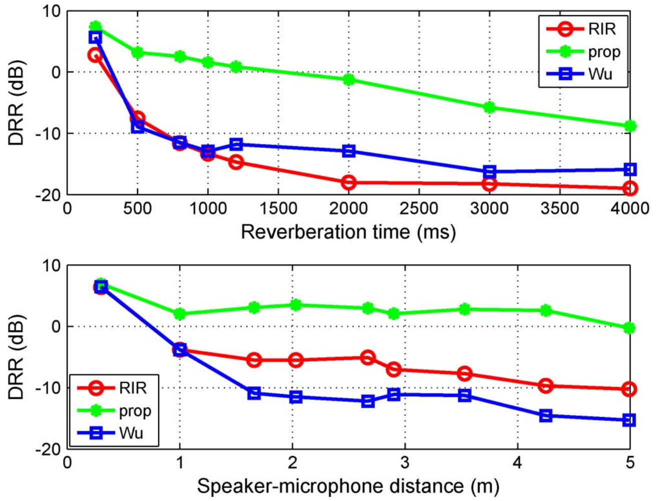


Fig. 16. DRR of the proposed and Wu and Wang algorithms for different reverberation times with $\mathrm{d} = 2$ m (upper), and for different speaker-microphone distances with $\mathrm{RT60} = 800$ ms (lower).

proposed method. It is clear that the proposed algorithm provides a significant increase in DRR and effectively estimates the inverse filter of the RIR in severe reverberant environments (i.e., with very high reverberation times and large speaker-microphone distances). The performance of the Wu and Wang method is much worse. Note that these results were obtained with only a single-microphone, which is a particularly difficult problem. To the best of our knowledge, no other algorithm based on a single-microphone can effectively estimate the inverse filter of the RIR with reverberation times higher than 1 s and arbitrary speaker-microphone distances.

B. Performance in Additive Noise

As discussed in Section III-B, the skewness-based approach is more robust to additive Gaussian noise than the kurtosis-based approach. However, the method based on skewness is still sensitive to additive noise due to the normalization. In this section, the performance of our proposed method is investigated in additive noise including recorded noise from real environments such as babble noise and background noise in a restaurant. For this purpose, the reverberant speech signals were synthesized by convolving the utterances from four female and four male speakers from the TIMIT database with the RIR ($\mathrm{RT60} = 250$ ms and speaker-microphone distance $\mathrm{d} = 1$ m). Then the noisy reverberant speech signals were obtained by adding these signals to three types of background noise: 1) white computer-generated Gaussian noise; 2) real babble noise;[4] and 3) real background noise recorded in a restaurant.[5] The estimated inverse filters for the proposed method (based on skewness) and the Wu and Wang method [7] (based on kurtosis) were evaluated using the DRR. Fig. 17 shows the results where white, res, and bab indicate white Gaussian noise, restaurant noise, and babble noise, respectively. The subscripts s and k denote skewness-based and kurtosis-based approaches, respectively. The DRR values for

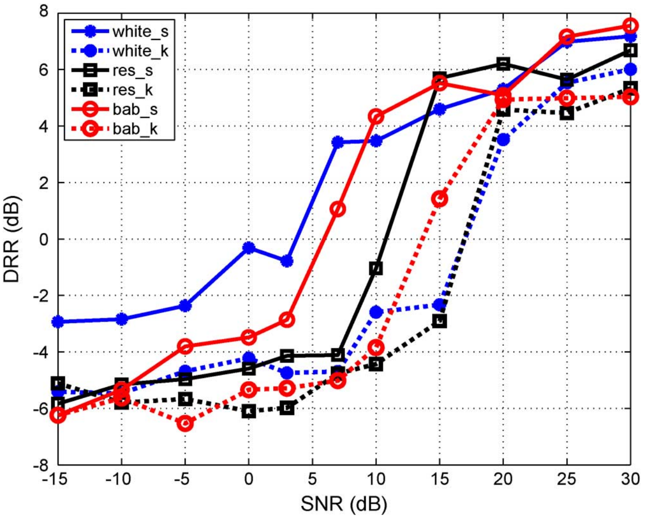


Fig. 17. DRR of the proposed and Wu and Wang algorithms for different noisy conditions with $\mathrm{RT60} = 250$ ms and d=1 m. white, res, and bab denote white Gaussian noise, noise recorded in a restaurant, and babble noise, respectively. s denotes the skewness-based approach (proposed method) and k denotes the kurtosis-based approach (Wu and Wang method).

the RIR and the equalized impulse response using the skewness-based and kurtosis-based approaches without noise are 1.52, 7.39, and 5.52, respectively. It is clear from these values that both methods provide suitable estimates of the inverse filter, but the estimate with the skewness-based approach is better than with the kurtosis-based approach. However, Fig. 17 shows that the performance is degraded as the noise intensity increases. Although the degradation for a realistic noisy environment (background noise in a restaurant) is more than with white Gaussian noise, our proposed method is effective for SNR values as low as 12 dB in this case. Moreover, our inverse filtering method is more robust in noise compared to the method proposed by Wu and Wang.

C. Performance in Time-Varying Environments

Although the basic assumption of the proposed inverse filtering is that the RIR is time-invariant, this can be violated in real environments. In order to evaluate the performance of our method in time-varying situations, we use two sets of RIRs with reverberation times of 500 and 1000 ms. For each set, the microphone is located at position [1 1.5 1] m. The speaker begins at position [2 1 3] m and moves along the x-axis with a step size of 100 cm to position [3 1 3] m. The RIRs were obtained by simulation using the image method [1]. The inverse filter is first estimated using 24 s of speech data. Each time the speaker moves, the filter is updated using different amounts of input data, with the previous filter used as the initial filter. The DRR values for the equalized impulse responses were calculated and are shown in Fig. 18. In this figure, RIR denotes the DRR of the RIR and for example $\mathrm{Data} = 2$ s denotes the DRR of the equalized impulse response updated using 2 s of input speech data. The best estimated inverse filters correspond to input speech data lengths of 24 s ($\mathrm{Data} = 24$ s), and these are compared to the filters updated with less input speech data. The results obtained indicate that the updated inverse filters using only 5 s of input speech data for a reverberation time of 500 s (upper plot), and only 6 s of

[4] [Online]. Available: http://www.ee.columbia.edu/~dpwe/sounds/noise/babble.wav.

[5] [Online]. Available: http://www.ee.columbia.edu/~dpwe/sounds/noise/restaurant.wav.

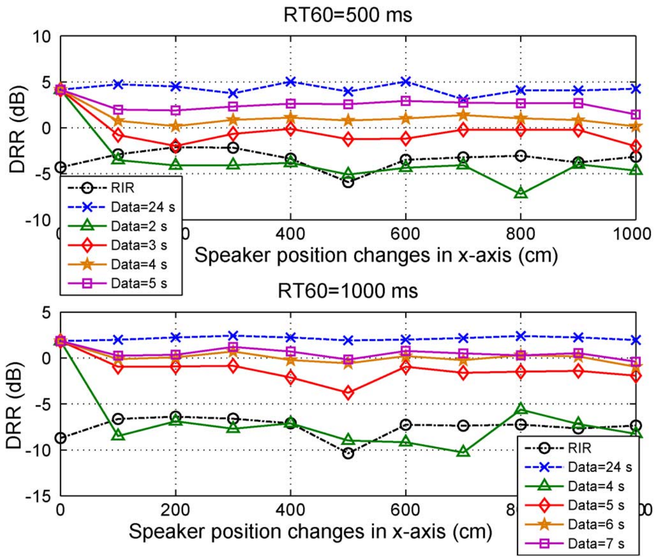


Fig. 18. Performance of the skewness-based method with a time-varying RIR. The inverse filter is updated using different lengths of input speech data. RIR denotes the DRR of the RIR, and for example Data $= 2$ s denotes the DRR of the equalized impulse response updated using 2-s input speech data.

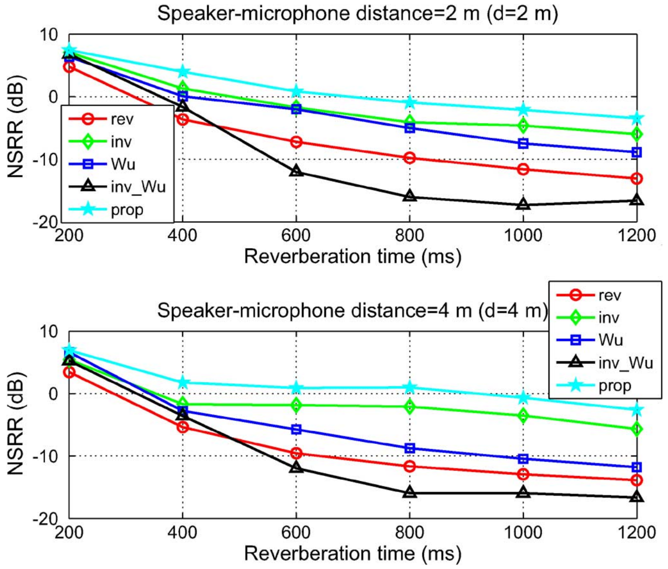


Fig. 19. NSRR of the proposed algorithm for different reverberation times when $d = 2$ m (upper plot) and $d = 4$ m (lower plot). rev, inv, inv-Wu, Wu, and prop denote the SRR of reverberant speech, the inverse-filtered speech using the skewness-based and kurtosis-based methods, and the processed speech using the Wu and Wang and proposed methods, respectively.

input speech data for a reverberation time of 1000 s (lower plot), provide performance similar to the best inverse filters. Similar results were obtained for other RIRs. This shows that the inverse filter can be updated effectively in slowly time-varying environments with small amounts of input speech data even in high reverberation conditions. Thus our method can be employed in slowly time-varying situations where the speaker pauses after moving, which can often occur in real environments. When a speaker does not pause after moving, it is not possible to accurately update the inverse filter.

D. Inverse Filtering Applications

As mentioned previously, inverse filtering finds applications in speech processing where reverberation occurs, including speech dereverberation [7], time-delay estimation (TDE) [26], and monaural speech segregation [29]. In this section, we evaluate the effectiveness of our proposed inverse filtering method for these reverberation problems.

1) Dereverberation: One of the most important applications of single-microphone inverse filtering is to reduce reverberation effects, particularly coloration caused by early reverberation [24]. Our inverse filtering method can be used in dereverberation applications as a first-stage to reduce the early reverberation effects, and combined with a second approach such as spectral subtraction to eliminate all reverberation effects as in [7]. In this section, the performance of the proposed method for reducing the effects of reverberation is investigated.

Two sets of RIRs were employed: one with RT60 in the range 200 to 1200 ms and a speaker-microphone distance of 2 m, and the other with RT60 in the same range but with a speaker-microphone distance of 4 m. The dereverberation performance was evaluated in terms of four common measures for this application. The overall reverberation reduction was evaluated using the normalized-signal-to-reverberation ratio (NSRR) [25].[6] The bark spectral distortion (BSD) [19] is a measure of speech quality. It provides the room reverberation as a perceptually weighted spectral difference between the original and reverberant signals, and thus shows the reduction in coloration and reverberation decay tail effects. As the LP residual kurtosis is a common measure to evaluate the reduction of coloration caused by early reflections [18], it is used here as the third measure. Finally, the listening quality of the dereverberated speech signals was evaluated using the perceptual evaluation of speech quality (PESQ). These four objective measures were applied to 32-ms frames overlapped by 50%. The results averaged over eight utterances (four male speakers and four female speakers) are shown in Figs. 19–22 where rev, inv, inv-Wu, prop, and Wu indicate the values for the reverberant speech signal, the inverse-filtered speech using the skewness-based method, the kurtosis-based method, the processed speech signals using the proposed two-stage reverberant speech enhancement method (skewness-based inverse filtering as the first-stage and the spectral subtraction of Wu and Wang [7] as the second-stage), and the Wu and Wang method [7], respectively. The upper plots are for a speaker-microphone distance of $\mathrm{d} = 2$ m, and the lower ones for a distance of $\mathrm{d} = 4$ m.

Fig. 19 shows that the skewness-based inverse filtering method reduces the effects of reverberation as there is an improvement in the NSRR values, especially at the larger speaker-microphone distance of $\mathrm{d} = 4$ m. This is because the coloration effects are dominant at larger distances. Conversely, the kurtosis-based inverse filtering method can only be used for low reverberation conditions ($\mathrm{RT60} \leq 400$ ms). In addition, it is clear that using the spectral subtraction method with the skewness-based inverse filtering can further reduce the overall effects of reverberation compared with the two-stage speech enhancement method of Wu and Wang. Fig. 20 indicates that the

[6] The paper as well as free Matlab source code for NSRR is available at http://home.tiscali.nl/ehabets/publications/Naylor2010.html.

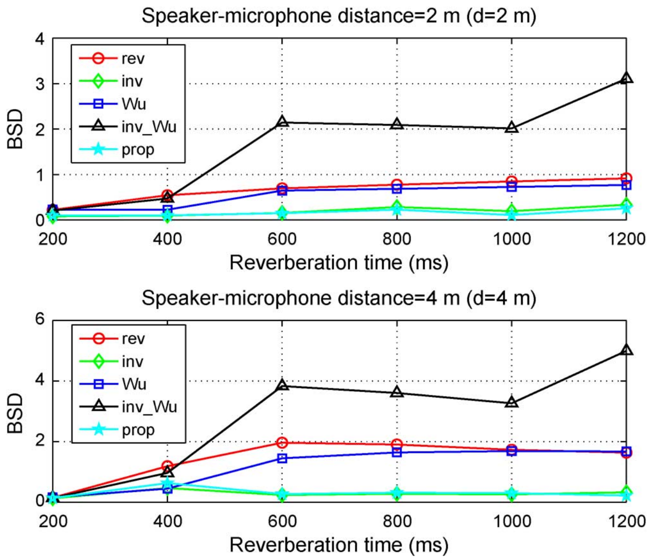


Fig. 20. BSD of the proposed algorithm for different reverberation times with d = 2 m (upper plot) and d = 4 m (lower plot). rev, inv, inv-Wu, Wu, and prop denote the BSD of reverberant speech, the inverse-filtered speech using the skewness-based and kurtosis-based methods, processed speech using the Wu and Wang method, and the proposed method, respectively.

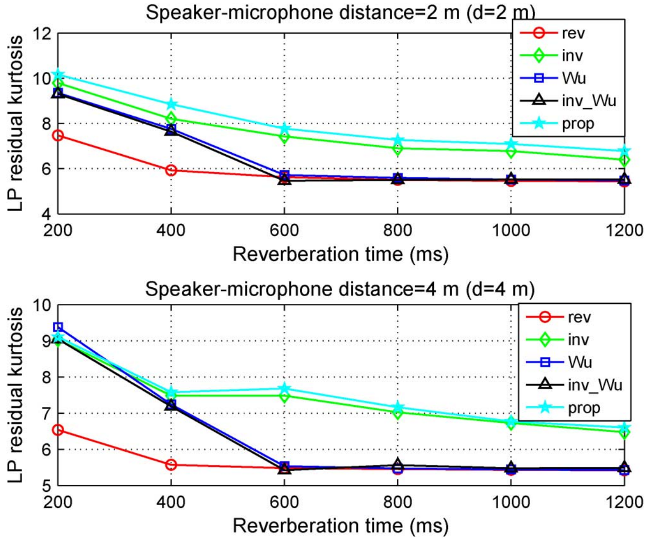


Fig. 21. LP residual kurtosis of the proposed algorithm for different reverberation times with d = 2 m (upper plot) and d = 4 m (lower plot). rev, inv, inv-Wu, Wu, and prop denote the LP residual kurtosis of reverberant speech, the inverse-filtered speech using the skewness-based and kurtosis-based methods, processed speech using the Wu and Wang method, and the proposed method, respectively.

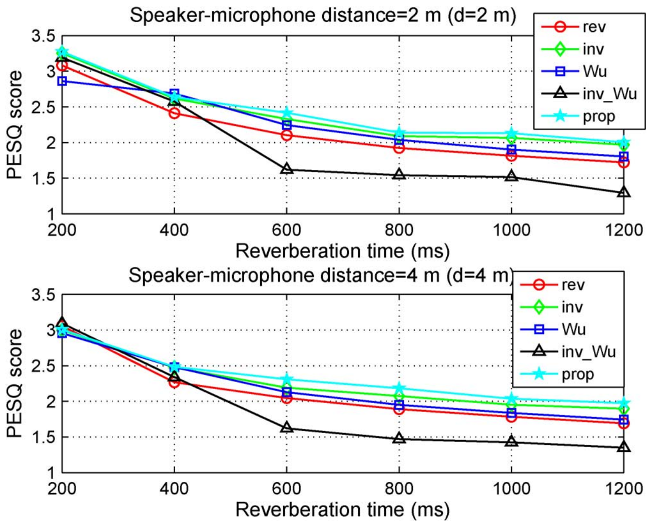


Fig. 22. PESQ of the proposed algorithm for different reverberation times with d = 2 m (upper plot) and d = 4 m (lower plot). rev, inv, inv-Wu, Wu, prop denote the *PESQ* of reverberant speech, the inverse-filtered speech using the skewness-based and kurtosis-based methods, processed speech using the Wu and Wang method, and the proposed method, respectively.

skewness-based inverse filtering method can greatly improve the BSD by reducing both the reverberation decay tail and coloration effects, whereas the kurtosis-based inverse filtering method and the two-stage method of Wu and Wang provide worse performance. As expected, our inverse filtering method essentially removes the early reverberation effects (coloration) by increasing the LP residual kurtosis, as shown in Fig. 21. The previous inverse filtering method and the two-stage method of Wu and Wang can only remove this effect in low reverberation conditions ($\mathrm{RT60} \leq 400$ ms). Our method can remove the coloration effects in highly reverberant conditions whereas, to the best of our knowledge, no other single microphone dereverberation algorithm can achieve such a result. Finally, Fig. 22 shows the PESQ for different reverberation times. This further confirms the superiority of our inverse filtering method for speech enhancement in different reverberation environments.

An audio demonstration can be found at https://sites.google.com/site/saeedmosayyebpour/experiment-data. These speech samples indicate that the processed speech signals still suffer from pre-echoes introduced by the inverse filtering method and musical noise introduced by the spectral subtraction method of Wu and Wang. Thus, the spectral subtraction method of Wu and Wang cannot effectively mitigate the remaining artifacts including the pre-echoes. For this application, inverse filtering alone should not be used because the resulting sound quality is poor and a suitable secondary technique should be used to mitigate the remaining artifacts. However, this method does reduce the effects of reverberation, particularly the early reverberation, even in highly reverberant situations.

2) Time Delay Estimation: The goal of TDE is to measure the relative time distance of arrival (TDOA) between spatially separated sensors. This plays an important role in radar, sonar, and seismology for localizing radiating sources. A variety of methods to deal with this problem have been proposed, but each has limitations in highly reverberant environments. In a highly reverberant room, none of the known methods work well and thus TDE in adverse conditions (e.g., a high RT60) is an important research problem. A two-stage algorithm using our proposed inverse filtering method and a procedure to estimate the TDOA has been developed [26].

The performance of the proposed TDE method was compared with the generalized cross-correlation (GCC) method that utilizes a weighted function of the phase transform (PHAT) [28], and the eigenvalue decomposition (EVD) method [27]. Experiments were conducted to evaluate the TDE between two microphones. Ten different RIRs with different microphone-speaker positions for RT60 values ranging from 200 to 1200 ms were employed. The three methods were used to estimate the TDOA for the 45 RIR pair combinations, and the root mean squared

TABLE IV

PERFORMANCE OF PITCH-BASED SEGREGATION [29] ALONE ($\text{output}_{\text{org}}$), AND USING THE SKEWNESS-BASED INVERSE FILTERING ($\text{output}_{\text{skew}}$) OR THE SKEWNESS-BASED ($\text{output}_{\text{kur}}$) AS THE FIRST STAGE. (input) DENOTES THE EVALUATION OF THE INPUT SIGNAL

	SNR (dB)				PESQ			
Interference	input	output_{org}	output_{skew}	output_{kur}	input	output_{org}	output_{skew}	output_{kur}
cocktail party noise	-1.7067	-0.0249	1.5121	0.312	2.611	2.686	2.927	2.332
white noise	-6.7317	-0.1174	0.551	0.2002	1.775	2.697	2.761	2.058
babble noise	-4.7278	-0.5148	1.9138	1.3626×10^{-4}	2.074	2.707	2.814	1.922

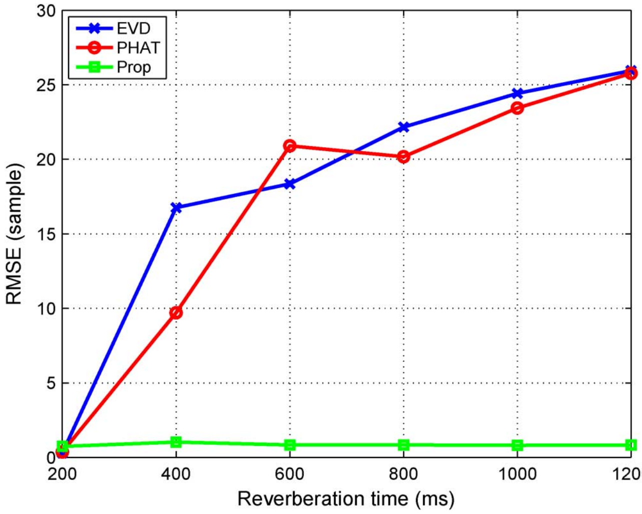


Fig. 23. RMSE of the estimated TDOA for different reverberation times using the EVD method [27] (EVD), PHAT method [28] (PHAT), and proposed method [26] (Prop).

error (RMSE) of the results was calculated. An acceptable estimate is defined as one that satisfies

$$|\text{TDOA}| \leq f_s \frac{R}{c} \tag{33}$$

where R is the distance between microphones, $f_s = 16000$ Hz is the sampling rate, and $c = 340$ m/s is the velocity of sound. The RMSE values for the acceptable estimations are shown in Fig. 23. It is clear from this figure that the proposed TDE method using our inverse filtering algorithm as the first-stage estimates the TDOA in different reverberation conditions better than the two conventional methods. The key point is that even in high reverberation conditions, our algorithm can accurately estimate the inverse filter. Note that only one microphone signal is used to estimate the inverse filter of each microphone. Since speech data from two microphones is available, this redundancy could be used to improve TDOA estimation performance.

3) Monaural Speech Segregation: Generally, there are two problems to overcome for monaural speech segregation: additive background noise and reverberation. While the problem of monaural speech segregation in additive noise conditions has been extensively investigated, little research has been devoted to reverberant environments. Pitch-based segregation algorithms have achieved considerable success in additive noise conditions. They have recently been extended to the case with both additive noise and reverberation by utilizing an inverse filtering method as a first stage [29]. As reverberation smears the harmonic structure of speech signals, it degrades the performance of pitch-based segregation algorithms. This degradation is intensified as the reverberation increases [29]. Using the inverse filtering in [29] as a first stage improves the harmonic structure of the signal while smearing signals originating at other locations (interference). This first stage provides a better input signal for pitch-based segregation and thus improves the performance in reverberant environments.

The performance of the pitch-based segregation method was evaluated with and without inverse filtering.[7] Both skewness-based and kurtosis-based inverse filtering were employed as the first stage of the pitch-based segregation method. The segregation was performed with a microphone located at position [1 1.5 2] m in a rectangular room with a reverberation time of 200 ms, and the source at position [0.5 1 1] m. The interference originates from position [1 0.5 2] m. We used three different interference signals: 1) cocktail party noise; 2) white computer-generated Gaussian noise; and 3) babble noise. The performance was evaluated using the signal-to-noise ratio (SNR) and PESQ as measures. The results are shown in Table IV where input, $\text{output}_{\text{org}}$, $\text{output}_{\text{skew}}$ and $\text{output}_{\text{kur}}$ denote the input signal, the output of the pitch-based segregation method, and the output of the pitch-based segregation method with skewness-based inverse filtering as the first stage and kurtosis-based inverse filtering as the first stage. It is clear that the use of skewness-based inverse filtering improves the segregation performance more than the other techniques.

V. CONCLUSION

In this paper, we proposed a single-microphone adaptive gradient-ascent algorithm for estimating the inverse filter of the RIR. This HOS-based approach maximizes the normalized third-order moment (skewness) of the LP residual to estimate the inverse filter. It was proven that a sufficiently long LP residual signal has an asymmetric pdf with high skewness, making skewness an effective metric to construct a score function for deconvolution. We also proved that skewness (as a measure of asymmetry), is more effective than the previously used kurtosis (as a measure of peakedness), in terms of the degradation caused by noise and reverberation. The proposed algorithm was optimized for implementation by using a procedure for initialization of the inverse filter and estimating the expected value of the feedback function. Performance results were presented which clearly show that the resulting equalized impulse response has improved DRR, and that the inverse filter of the RIR can be effectively estimated in highly reverberant conditions (very high reverberation times with various speaker-microphone distances). The performance in noisy conditions, including recorded background noise, shows that the

[7] The free source code available at http://www.cse.ohio-state.edu/dwang/pnl/shareware/hu-chapter06/voiced06.zip was used to implement the pitch-based segregation method.

proposed method is effective in moderate noise conditions. In slowly time-varying environments, it was demonstrated that the inverse filter can be updated appropriately with small amounts of input speech, so the proposed method can be employed in realistic situations.

The effectiveness of the proposed inverse filtering method was investigated in three different applications. It was used as a first stage in a single-microphone dereverberation application to reduce the coloration effect. In conjunction with spectral subtraction, the reverberation effects were reduced more effectively than with the two stage method proposed by Wu and Wang [7] based on four standard measures (NSRR, BSD, LP residual kurtosis, and PESQ), but the spectral subtraction method of Wu and Wang cannot effectively reduce the effect of the pre-echoes on the equalized impulse response which remains from the first stage of inverse filtering. Thus, the processed speech suffers from annoying audible artifacts. Further improvement may be possible by combining a suitable technique with our inverse filtering method to combat both the late-impulses and the pre-echoes remaining after inverse filtering. Regardless of this, our inverse filtering method provides a significant reduction in early reverberation effects even in highly reverberant rooms which to the best of our knowledge is the best single-microphone based dereverberation method for this purpose. It was shown that a two stage TDE method employing the proposed inverse filtering method can successfully estimate the TDOA even in high reverberation conditions. The last application considered was monaural speech segregation. The performance of pitch-based segregation was improved using our proposed inverse filtering method.

REFERENCES

[1] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[2] S. T. Neely and J. B. Allen, “Invertibility of a room impulse response,” *J. Acoust. Soc. Amer.*, vol. 66, no. 1, pp. 165–169, Jul. 1979.

[3] J. Mourjoupolous, P. M. Clarkson, and J. K. 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.*, 1982, pp. 1858–1861.

[4] D. Bees, M. Blostein, and P. Kabal, “Reverberant speech enhancement using cepstral processing,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1991, pp. 977–980.

[5] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, “Speech dereverberation via maximum-kurtosis subband adaptive filtering,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 3701–3704.

[6] M. Tonelli and N. Mitianoudis, “A maximum likelihood approach to blind audio dereverberation,” in *Proc. Int. Conf. Digital Audio Effects*, Oct. 2004, pp. 254–261.

[7] M. Wu and D. L. Wang, “A two-stage algorithm for one-microphone reverberant speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, May 2006.

[8] A. J. Bell and T. J. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[9] D. Erdogmus, K. E. Hild, J. C. Principe, M. Lazaro, and I. Santamaria, “Adaptive blind deconvolution of linear channels using Renyi’s entropy with Parzen window estimation,” *IEEE Trans. Signal Process.*, vol. 52, no. 6, pp. 1489–1498, Jun. 2004.

[10] P. A. Naylor and N. D. Gaubitch, “Speech dereverberation,” in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sep. 2005.

[11] B. Yegnanarayana and P. S. Murthy, “Enhancement of reverberant speech using LP residual signal,” *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 267–281, May 2000.

[12] R. Grimmet and D. R. Stirzaker, *Probability and Random Processes*, 3rd ed. Oxford, U.K.: Oxford Univ. Press, 2001.

[13] P. Pääjärvi and J. P. LeBlanc, “Skewness maximization for impulsive sources in blind deconvolution,” in *Proc. IEEE Nordic Signal Process. Symp.*, Jun. 2004, pp. 304–307.

[14] P. Pääjärvi and J. P. LeBlanc, “Online adaptive blind deconvolution based on third-order moments,” *IEEE Signal Process. Lett.*, vol. 12, no. 12, pp. 863–866, Dec. 2005.

[15] P. Pääjärvi and J. P. LeBlanc, “Blind equalization of PPM signals using third-order moments,” in *Proc. IEEE Signal Process. Adv. Wireless Commun. Workshop*, Dec. 2007, pp. 1–5.

[16] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, “Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 534–545, May 2009.

[17] A. Keshavarz, S. Mosayyebpour, M. Biguesh, A. Gulliver, and M. Esmaeili, “Speech-model based accurate blind reverberation time estimation using an LPC filter,” *IEEE Trans. Audio, Speech, Lang. Process.*, to be published.

[18] E. A. P. Habets, N. Gaubitch, and P. A. Naylor, “Temporal selective dereverberation of noisy speech using one microphone,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4577–4580.

[19] E. A. P. Habets, “Single- and multi-microphone speech dereverberation using spectral enhancement,” Ph.D. dissertation, Eindhoven Univ. of Tech., Eindhoven, The Netherlands, 2007.

[20] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York: Wiley, 1987.

[21] S. Mosayyebpour, A. Sayyadiyan, M. Zareian, and A. Shahbazi, “Single channel inverse filtering of room impulse response by maximizing skewness of LP residual,” in *Proc. IEEE Int. Conf. Signal Acq. Process.*, Feb. 2010, pp. 130–134.

[22] S. Gazor and W. Zhang, “Speech probability distribution,” *IEEE Signal Process. Lett.*, vol. 10, no. 7, pp. 204–207, Jul. 2003.

[23] E. W. Weisstein, *CRC Concise Encyclopedia of Mathematics*. London, U.K.: Chapman & Hall/CRC, 1999.

[24] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Berlin, Germany: Springer, 2010.

[25] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets, “Signal-based performance evaluation of dereverberation algorithms,” *J. Elect. Comput. Eng.*, vol. 2010, 2010, Article ID 127513 .

[26] S. Mosayyebpour, A. Sayyadiyan, E. Soltan Mohammadi, A. Shahbazi, and A. Keshavarz, “Time delay estimation using one microphone inverse filtering in highly reverberant room,” in *Proc. IEEE Int. Conf. Signal Acq. Process.*, Feb. 2010, pp. 140–144.

[27] J. Benesty, “Adaptive eigenvalue decomposition algorithm for passive acoustic source localization,” *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.

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

[29] N. Roman and D. L. Wang, “Pitch-based monaural segregation of reverberant speech,” *J. Acoust. Soc. Amer.*, vol. 120, no. 1, pp. 458–469, Jul. 2006.

Saeed Mosayyebpour (S’11) received the B.Sc. degree in electrical engineering from Amirkabir University of Technology, Tehran, Iran, in 2007, with majors in electronics and communication engineering and the M.Sc. degree in electrical engineering from Amirkabir University, in 2010. He is currently pursuing the Ph.D. degree at the University of Victoria, Victoria, BC, Canada.

His research interests include speech processing, in particular speech enhancement in noisy reverberant environments, source localization, speech separation, and statistical signal processing.

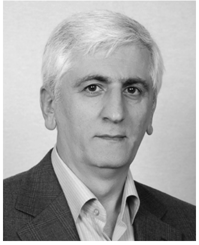
Hamid Sheikhzadeh (M'03–SM'04) received the B.S. and M.S. degrees in electrical engineering from Amirkabir University of Technology, Tehran, Iran, in 1986 and 1989, respectively, and the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994.

He was a faculty member in the Electrical Engineering Department, Amirkabir University of Technology, until September 2000. From 2000 to 2008, he was a Principle Researcher with ON semiconductor, Waterloo, ON, Canada. During this period, he developed signal processing algorithms for ultra-low-power and implantable devices leading to many international patents. Currently, he is a faculty member in the Electrical Engineering Department of Amirkabir University of Technology. His research interests include signal processing and speech processing, with particular emphasis on speech recognition, speech enhancement, auditory modeling, adaptive signal processing, subband-based approaches, and algorithms for low-power DSP.

T. Aaron Gulliver (M'86–SM'96) received the Ph.D. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 1989.

From 1989 to 1991, he was a Defence Scientist at Defence Research Establishment Ottawa, Ottawa, ON, Canada. He has held academic positions at Carleton University, Ottawa, and the University of Canterbury, Christchurch, New Zealand. He joined the University of Victoria in 1999 and is a Professor in the Department of Electrical and Computer Engineering.

Dr. Gulliver became a Fellow of the Engineering Institute of Canada in 2001. He is currently an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. From 2000 to 2003, he was Secretary and a member of the Board of Governors of the IEEE Information Theory Society. His research interests include information theory and communication theory, algebraic coding theory, MIMO systems, and ultrawideband communications.



Morteza Esmaeili (M'09) received the M.Sc. degree in mathematics from the Teacher Training University of Tehran, Tehran, Iran, in 1988 and the Ph.D. degree in mathematics (coding theory) from Carleton University, Ottawa, ON, Canada, in 1996.

After the Ph.D. degree, he was a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. He joined Isfahan University of Technology, Isfahan, Iran, in September 1998, where he is now a Professor in the Department of Mathematical Sciences. He has been an Adjunct Professor in the Department of Electrical and Computer Engineering, University of Victoria, Victoria, B.C., Canada, since July 2009. His current research interests include coding and information theory, cryptography, and combinatorics.