A Two-Stage Algorithm for One-Microphone Reverberant Speech Enhancement

Mingyang Wu, *Member, IEEE,* and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Under noise-free conditions, the quality of reverberant speech is dependent on two distinct perceptual components: coloration and long-term reverberation. They correspond to two physical variables: signal-to-reverberant energy ratio (SRR) and reverberation time, respectively. Inspired by this observation, we propose a two-stage reverberant speech enhancement algorithm using one microphone. In the first stage, an inverse filter is estimated to reduce coloration effects or increase SRR. The second stage employs spectral subtraction to minimize the influence of long-term reverberation. The proposed algorithm significantly improves the quality of reverberant speech. A comparison with a recent enhancement algorithm is made on a corpus of speech utterances in a number of reverberant conditions, and the results show that our algorithm performs substantially better.**

***Index Terms*—Dereverberation, inverse filtering, one-microphone algorithm, reverberant speech enhancement, reverberation time, spectral subtraction.**

I. Introduction

A MAIN cause of speech degradation in practically all listening situations is room reverberation. Although human listening is little affected by room reverberation to a considerable degree—indeed increased loudness as a result of reverberation may even enhance speech intelligibility [19]—reverberation causes significant performance decrement for current automatic speech recognition (ASR) and speaker recognition systems. Consequently, an effective reverberant speech enhancement system is essential for many speech technology applications including speech and speaker recognition. Also, hearing-impaired listeners suffer from reverberation effects disproportionally [26]. A system that enhances reverberant speech should improve intelligent hearing aid design.

In this paper, we study one-microphone reverberant speech enhancement. This is motivated by the following two considerations. First, a one-microphone solution is highly desirable for many real-world applications such as telecommunication (e.g., processing of telephone speech) and audio information retrieval (information mining from audio archives). Second, moderately reverberant speech is highly intelligible in monaural listening conditions. Hence, how to achieve this monaural capability remains a fundamental scientific question.

Many methods have been previously proposed to deal with room reverberation. Some enhancement algorithms assume that room impulse response functions are known. For instance, delay-sum beamformers [13] and matched filters [14] have been employed to reduce reverberation effects. One idea to remove reverberation effects is by passing the reverberant signal through a second filter that inverts the reverberation process and recover the original signal. A perfect reconstruction of the original signal exists, however, only if the room impulse response function is a minimum-phase filter. However, as pointed out by Neely and Allen [28], room impulse responses are often not minimum-phase. Another solution is to use multiple microphones. By assuming no common zeros among the room impulse responses, an exact inverse filtering can be realized using finite-impulse response (FIR) filters [25]. In the one-microphone case, methods, such as linear least-square equalizers, have been suggested that partially reconstruct the original signal [17].

A number of reverberant speech enhancement algorithms have been designed to perform in unknown acoustic environments but utilize more than one microphone. For example, microphone-array based methods [10], such as beamforming techniques, attempt to suppress the sound energy coming from directions other than that of the direct source and, therefore, enhance target speech. As pointed out by Koenig *et al.* [23], the reverberation tails of the impulse responses, characterizing the reverberation process in a room with multiple microphones and one speaker, are uncorrelated. Several algorithms are proposed to reduce the reverberation effects by removing the incoherent parts of received signals (for example, see [3]). Blind deconvolution algorithms aim to reconstruct the inverse filters without the prior knowledge of room impulse responses (for example, see [16], [18]). Brandstein and Griebel [9] utilize the extrema of wavelet coefficients to reconstruct the linear prediction (LP) residual of original speech.

With multiple sound sources in a room, the signals received by microphones can be viewed as convolutive mixtures of original signals emitted by the sources. Several methods (for example, see [7]) have been proposed to achieve blind source separation (BSS) of convolutive mixtures, estimating the original signals using only the information of the convolutive mixtures received by the microphones. Some methods consider unmixing systems as FIR filters, while others convert the problem into the frequency domain and solve an instantaneous BSS for every frequency channel. The performance of frequency-domain BSS algorithms, however, is quite poor in a realistic acoustic environment with moderate reverberation time [4].

Manuscript received November 6, 2003; revised March 29, 2005. This work was supported in part by the National Science Foundation under Grant IIS-0081058 and by the Air Force Office of Scientific Research under Grant F49620-01-1-0027. A brief version of this paper was included in the Proceedings of ICASSP 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bayya Yegnanarayana.

M. Wu is with the Fair Isaac Corporation, San Diego, CA 92130 USA (e-mail: MingyangWu@fairisaac.com).

D.L. Wang is with the Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TSA.2005.858066

Reverberant speech enhancement using one microphone is significantly more challenging than that using multiple microphones. Nonetheless, a number of one-microphone algorithms have been proposed. Bees *et al.* [6] employs a cepstrum-based method to estimate the cepstrum of reverberation impulse response, and its inverse is then used to dereverberate the signal. Several dereverberation algorithms (for example, see [5]) are motivated by the effects of reverberation on modulation transfer function (MTF) [21]. Yegnanarayana and Murthy [36] observed that LP residual of voiced clean speech has damped sinusoidal patterns within each glottal cycle, while that of reverberant speech is smeared and resembles Gaussian noise. With this observation, LP residual of clean speech is estimated and then the enhanced speech is resynthesized. Nakatani and Miyoshi [27] proposed a system capable of blind dereverberation by employing the harmonic structure of speech. Good results are obtained but this algorithm requires a large amount of reverberant speech produced using the same room impulse response function.

Despite these studies, existing reverberant speech enhancement algorithms, however, do not reach a performance level demanded by many practical applications. Motivated by the observation that reverberation leads to perceptual components: coloration and long-term reverberation, we present a novel two-stage algorithm for one-microphone reverberant speech enhancement. In the first stage, an inverse filter is estimated in order to reduce coloration effects so that signal-to-reverberant energy ratio (SRR) is increased. The second stage utilizes spectral subtraction to minimize the influence of long-term reverberation. Our two-stage algorithm has been systematically evaluated, and the results show that the algorithm achieves substantial improvements on reverberant speech. We have also carried out a quantitative comparison with a recent one-microphone speech enhancement algorithm on a corpus of reverberant speech and our algorithm yields significantly better performance.

This paper is organized as follows. In the next section, we give the background that motivates our two-stage algorithm. Section III presents the first stage of the algorithm—inverse filtering. The second stage of the algorithm—spectral subtraction—is detailed in Section IV. Section V discribes evaluation experiments and shows the results. Finally, we discuss related issues and conclude the article in Section VI.

II. Background

Reverberation causes a noticeable change in speech quality [8]. Berkley and Allen [8] identified that two physical variables, reverberation time T_{60} and the talker–listener distance, are important for reverberant speech quality. Consider the impulse response as a combination of three parts, the direct signal, early and late reflections, where the direct signal corresponds to the direct path from a speech source to a listener. While late reflections smear the speech spectra and reduce the intelligibility and quality of speech signals, early reflections cause a different kind of distortion called coloration: the nonflat frequency response of the early reflections distorts the speech spectrum. The coloration can be characterized by a spectral deviation defined as the standard deviation of room frequency response.

Allen [1] reported a formula derived from a nonlinear regression to predict the quality of reverberant speech as measured by subjective preference

$$\frac{P}{P_{\mathrm{MAX}}} = 1 - 0.3\sigma T_{60} \tag{1}$$

where P_{MAX} is the maximum preference, σ is the spectral deviation in decibels, and T_{60} is the reverberation time in seconds. According to this formula, increasing either spectral deviation or reverberation time results in decreased reverberant speech quality. Jetzt [22] shows that spectral deviation is determined by SRR. Furthermore, within the same room, the relative reverberant energy—the total reflection energy normalized by the direct signal energy—is approximately constant regardless of the locations of the source and the listener. Therefore, in the same room spectral deviation is determined by the talker-to-microphone distance, which determines the strength of the direct signal. Shorter talker-to-microphone distance results in higher SRR and less spectral deviation, hence, less distortion or coloration.

Consequently, we propose a two-stage model to deal with two types of degradation—coloration and long-term reverberation—in a reverberant environment. In the first stage, our model estimates an inverse filter in order to reduce coloration effects or to increase SRR. The second stage employs spectral subtraction to minimize the influence of long-term reverberation. Detailed description of the two stages of our algorithm is given in the following two sections.

III. Inverse Filtering

As described in Section I, inverse filtering can be utilized to reconstruct the original signal. In the first stage of our algorithm, we derive an inverse filter to reduce reverberation effects. For this stage we apply a multimicrophone inverse filtering algorithm proposed by Gillespie *et al.* [18] to the one-microphone arrangement. Their algorithm estimates an inverse filter of the room impulse response by maximizing the kurtosis of the LP residual of speech utilizing multiple microphones. A detailed formulation of the kurtosis maximization is given next.

Assuming that $\hat{\mathbf{g}} = [g(1), g(2), \ldots, g(L)]$ is an inverse filter of length L, the inverse-filtered speech is

$$z(t) = \hat{\mathbf{g}}\hat{\mathbf{y}}(t) \tag{2}$$

where $\hat{\mathbf{y}}(t) = [y(t-L+1), \ldots, y(t-1), y(t)]^T$ and y is the reverberant speech, sampled at 16 kHz.

The LP residual of clean speech has higher kurtosis than that of reverberant speech [36]. Consequently, an inverse filter can be sought by maximizing the kurtosis of LP residual signal of the inverse-filtered signal [18]. A schematic diagram of a direct implementation of such a system is shown in Fig. 1(a). However, due to the LP analysis in the feedback loop, the optimization problem is not trivial. As a result, an alternative system is employed for inverse filtering [18] and shown in Fig. 1(b). Here, the LP residual of the processed speech is approximated by the inverse-filtered LP residual of the reverberant speech $\widetilde{z}(t)$. Consequently, we have

$$\widetilde{z}(t) = \hat{\mathbf{g}}\hat{\mathbf{y}}_r(t) \tag{3}$$

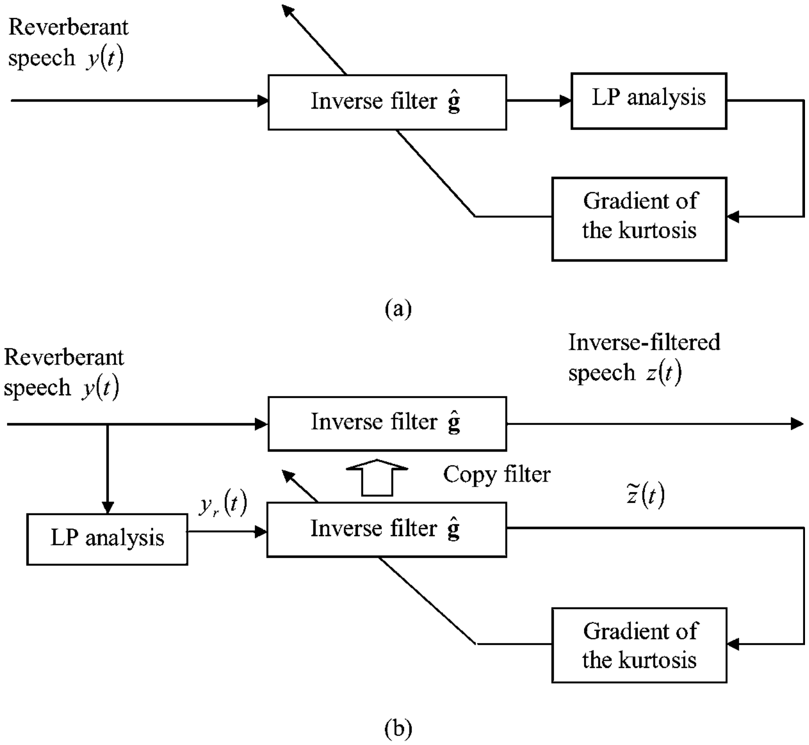
Reverberant speech $y(t)$
Inverse filter $\hat{\mathbf{g}}$
LP analysis
Gradient of the kurtosis
(a)
Reverberant speech $y(t)$
Inverse-filtered speech $z(t)$
Inverse filter $\hat{\mathbf{g}}$
Copy filter
$y_r(t)$
$\tilde{z}(t)$
LP analysis
Inverse filter $\hat{\mathbf{g}}$
Gradient of the kurtosis
(b)

Fig. 1. (a) Schematic diagram of an ideal one-microphone dereverberation algorithm maximizing the kurtosis of LP residual of inverse-filtered signal. (b) Diagram of the algorithm employed in the first stage of our algorithm.

where $\hat{\mathbf{y}}_r(t) = [y_r(t-L+1), \ldots, y_r(t-1), y_r(t)]^T$ and $y_r(t)$ is the LP residual of the reverberant speech. The optimal inverse filter $\hat{\mathbf{g}}$ is derived so that the kurtosis of $\tilde{z}(t)$, i.e., $E[\tilde{z}^4(t)]/E^2[\tilde{z}^2(t)] - 3$, is maximized. By obtaining the kurtosis gradient, the optimization problem can be formulated as a time-domain adaptive filter and the update equation of the inverse filter becomes (see [18])

$$\hat{\mathbf{g}}(t+1) = \hat{\mathbf{g}}(t) + \mu f(t)\hat{\mathbf{y}}_r(t) \tag{4}$$

where

$$f(t) = \frac{4\left(E\left[\tilde{z}^2(t)\right]\tilde{z}^3(t) - E\left[\tilde{z}^4(t)\right]\tilde{z}(t)\right)}{E^3\left[\tilde{z}^2(t)\right]} \tag{5}$$

and μ denotes the learning rate for every time step.

According to Haykin [20], however, the time-domain adaptive filter formulation is not recommended, because the large variations in the eigenvectors of the autocorrelation matrices of the input signals may lead to very slow convergence, or no convergence at all. Consequently, we use a block frequency-domain structure for optimization. In this formulation, the signal is processed block by block using fast Fourier transforms (FFTs) and the filter length L is also used as the block length. The new update equations for the inverse filter are as follows

$$\mathbf{G}'(n+1) = \mathbf{G}(n) + \frac{\mu}{M}\sum_{m=1}^{M}\mathbf{F}(m)\mathbf{Y}_r^*(m) \tag{6}$$

$$\mathbf{G}(n+1) = \frac{\mathbf{G}'(n+1)}{|\mathbf{G}'(n+1)|} \tag{7}$$

where $\mathbf{F}(m)$ and $\mathbf{Y_r}(m)$ denote, respectively, the FFT of $f(t)$ and $\hat{\mathbf{y}}_r(t)$ for the mth block. The superscript $*$ denotes complex conjugate. $\mathbf{G}(n)$ is the FFT of $\hat{\mathbf{g}}$ at nth iteration and M is the number of blocks. Equation (7) ensures that the inverse filter is normalized. Finally, the inverse-filtered speech $z(t)$ is obtained by convolving the reverberant speech with the inverse filter. Specifically, we choose $\mu = 3 \times 10^{-9}$ and use 20-s reverberant speech to derive the inverse filter. We run for 500 iterations which are needed for good results.

A typical result from the first stage of our algorithm is shown in Fig. 2. Fig. 2(a) illustrates a room impulse response function ($T_{60} = 0.3$ s) generated by the image model of Allen and Berkley [2], which is commonly used for this purpose. The equalized impulse response—the result of the room impulse response in Fig. 2(a) convolved with the obtained inverse filter—is shown in Fig. 2(b). As can be seen, the equalized impulse response is far more impulse-like than the room impulse response. In fact, the SRR value of the room impulse response is -9.8 dB in comparison with 2.4 dB for that of the equalized impulse response.

However, the above inverse filtering method does not improve on the tail part of reverberation. Fig. 3(a) and (b) show the energy decay curves of the room impulse response and the equalized impulse response, respectively. As can be seen, except for the first 50 ms, the energy decay patterns are almost identical, and, thus, the estimated reverberation times are almost the same, around 0.3 s. While the coloration distortion is reduced due to the increase of SRR, the degradation due to reverberation tails is not alleviated. In other words, the effect of inverse filtering is similar to that of moving the sound source closer to the receiver. In the next section, we introduce the second stage of our algorithm to reduce the effects of long-term reverberation.

IV. Spectral Subtraction

Late reflections in a room impulse response function smear speech spectrum and degrade speech intelligibility and quality.

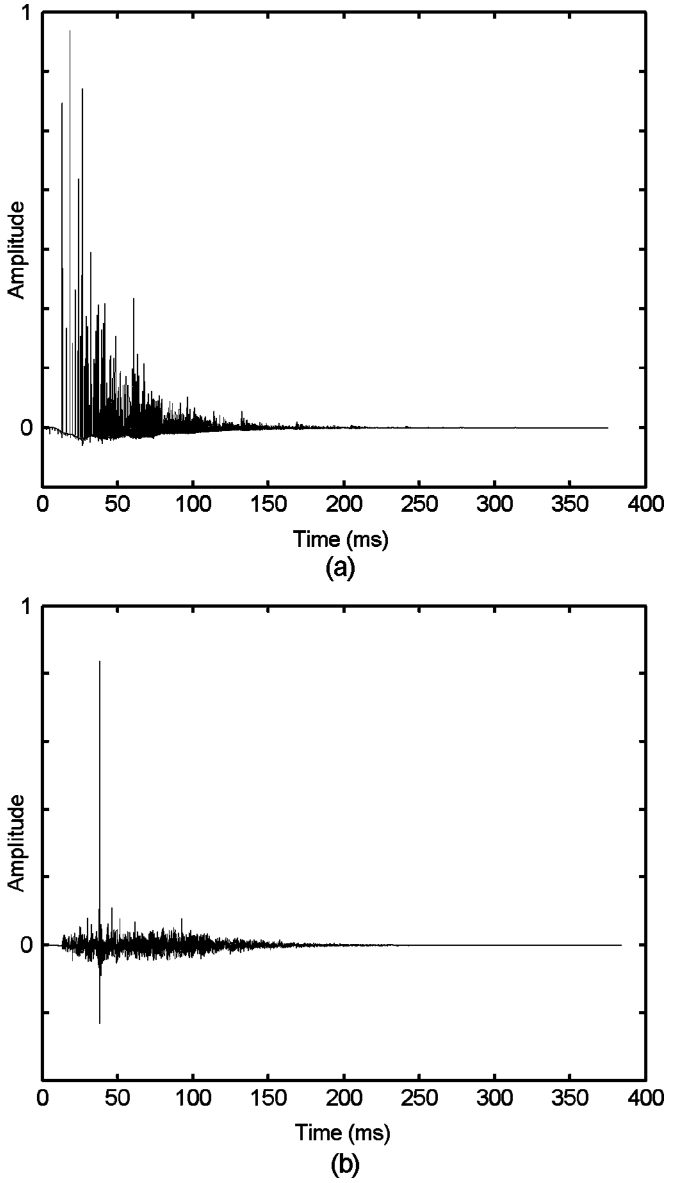

Fig. 2. (a) Room impulse response function generated by the image model in an office-size room of the dimensions 6 by 4 by 3 m (length by width by height). Wall reflection coefficients are 0.75 for all walls, ceiling, and floor. The loudspeaker and the microphone are at (2, 3, 1.5) and (4, 1, 2), respectively. (b) The equalized impulse response derived from the reverberant speech generated by the room impulse response in (a) as the result of the first stage of our algorithm.

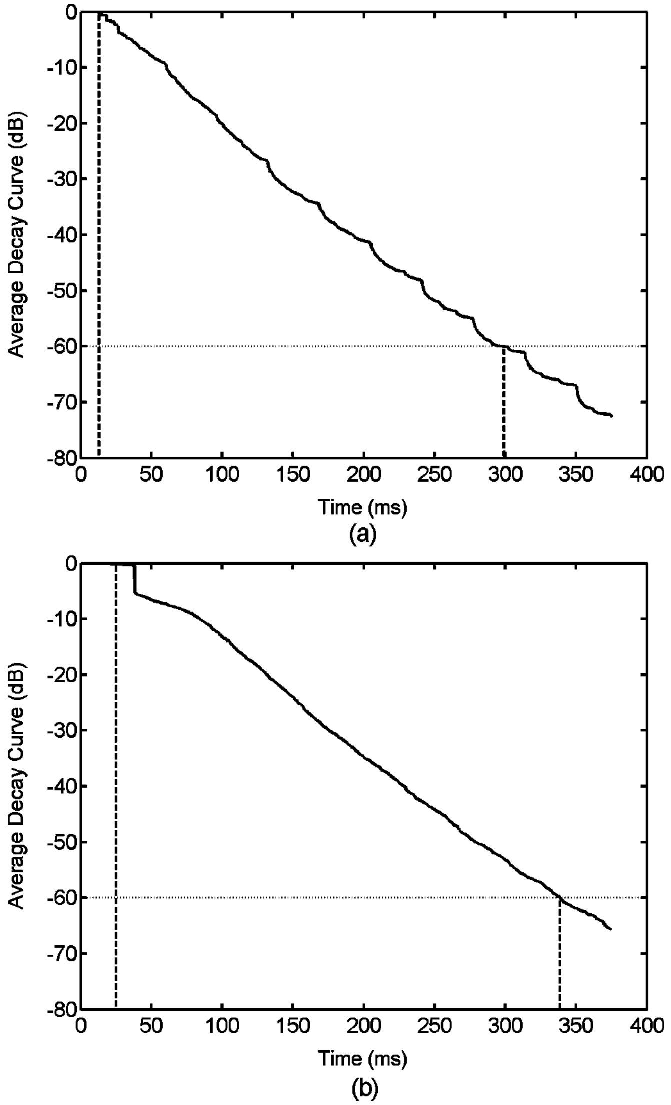

Fig. 3. Energy decay curves (a) computed from the room impulse response function in Fig. 2(a), and (b) from the equalized impulse response in Fig. 2(b). Each curve is calculated using the Schroeder integration method. The horizontal dot line represents -60-dB energy decay level. The left dash lines indicate the starting times of the impulse responses and the right dash lines the times at which decay curves cross -60 dB.

Likewise, an equalized impulse response can be decomposed into two parts: early and late impulses. Resembling the effects of the late reflections in a room impulse response, the late impulses have deleterious effects on the quality of inverse-filtered speech; by estimating the effects of the late impulses and subtracting them, we can expect to enhance the speech quality.

Several methods have been proposed to reduce the effects of late reflections in a room impulse response. Palomäki *et al.* [29] employ a robust speech recognition technique in reverberant environments by utilizing only the least reverberation-contaminated time-frequency regions. These regions are determined by applying a reverberation masking filter to estimate the relative strength of reverberant and clean speech. Wu and Wang [35] propose a one-stage algorithm to enhance the reverberant speech by estimating and subtracting effects of late reflections. Reverberation causes the elongation of harmonic structure in voiced speech and, therefore, produces elongated pitch tracks. In order to obtain more accurate pitch estimation in reverberant environments, Nakatani and Miyoshi [27] employ a filter $f_p = (1, -e, -e, \ldots, -e)$ to prefilter the amplitude spectrum in the time domain and, thus, reduces some elongated pitch tracks in reverberant speech.

The smearing effects of late impulses lead to the smoothing of the signal spectrum in the time domain. Therefore, we assume that the power spectrum of late-impulse components is a smoothed and shifted version of the power spectrum of the inverse-filtered speech $z(t)$

$$|S_l(k;i)|^2 = \gamma w(i-\rho) * |S_z(k;i)|^2 \tag{8}$$

where $|S_z(k;i)|^2$ and $|S_l(k;i)|^2$ are, respectively, the short-term power spectra of the inverse-filtered speech and the late-impulse components. Indexes k and i refer to frequency bin and time frame, respectively. The symbol $*$ denotes convolution in the time domain and $w(i)$ is a smoothing function. The short-term speech spectrum is obtained by using hamming windows of length 16 ms with 8-ms overlap for short-term Fourier analysis. The shift delay ρ indicates the relative delay of the late-impulse components. The distinction of early and late reflections for speech is commonly set at a delay of 50 ms in a room impulse response function [24]. This delay reflects speech properties and is independent of reverberation characteristics. The delay translates to approximately 7 frames for the chosen shift interval of 8 ms. Consequently we choose $\rho = 7$. Finally, the scaling factor γ specifies the relative strength of the late-impulse components after inverse filtering. We set γ to 0.32, although its detailed values do not matter (see Section V for discussions).

Considering the shape of the equalized impulse response, we choose an asymmetrical smoothing function as the Rayleigh distribution[1]

$$\begin{cases} w(i) = \frac{i+a}{a^2}\exp\left(\frac{-(i+a)^2}{2a^2}\right), & \text{if } i > -a \\ w(i) = 0, & \text{otherwise.} \end{cases} \qquad (9)$$

As shown in Fig. 4, this smoothing function peaks at $i = 0$ and goes down to 0 on the left side at $i = -a$ but drops off more slowly on the right side; the right side of the smoothing function resembles the shape of reverberation tail in an equalized impulse response. The parameter a controls the overall spread of the function. Given $\rho = 7$, a needs to be smaller than ρ, and we choose $a = 5$ (frames) which gives a reasonable match to the shape of the equalized impulse response (see Fig. 4); more discussions are given in Section V.

The inverse-filtered speech $z(t)$ can be expressed as the convolution of the clean speech $s(t)$ and the equalized impulse response $h_e(t)$

$$z(t) = \int_0^\infty s(t-\tau)h_e(\tau)d\tau. \qquad (10)$$

By separating the contributions from early and late impulses in the equalized impulse response, we rewrite (10) as

$$z(t) = \int_0^{T_l} s(t-\tau_1)h_e(\tau_1)d\tau_1 + \int_{T_l}^\infty s(t-\tau_2)h_e(\tau_2)d\tau_2 \qquad (11)$$

where T_l indicates the separation between early and late impulses. The first and the second terms in (11) represent the early- and late-impulse components, respectively, and are computed from different segments of original clean speech.

To justify the use of spectral subtraction, we now show that early- and late-impulse components are approximately uncorrelated. If we consider that the clean speech $s(t)$ and the equalized impulse response $h_e(t)$ are independent random processes, we have

$$E\left[\int_0^{T_l} s(t-\tau_1)h_e(\tau_1)d\tau_1 \times \int_{T_l}^\infty s(t-\tau_2)h_e(\tau_2)d\tau_2\right]$$
$$= \int_0^{T_l}\int_{T_l}^\infty E\left[s(t-\tau_1)s(t-\tau_2)\right]E\left[h_e(\tau_1)h_e(\tau_2)\right]d\tau_2 d\tau_1. \qquad (12)$$

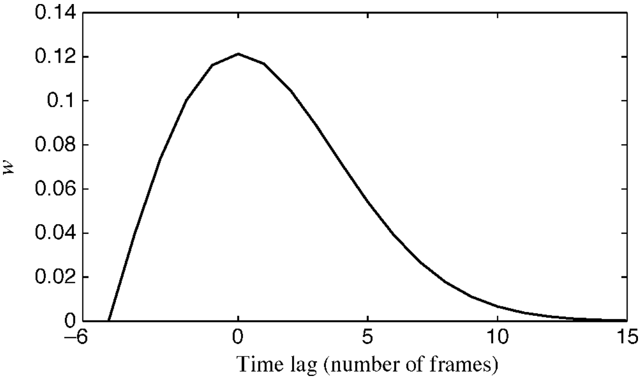

Fig. 4. Smoothing function ((9) in the text) for approximating late-impulse components. In the figure, $a = 5$.

τ_1 and τ_2 cover different ranges in their respective integrations. Due to the long-term uncorrelation of speech signal, $E[s(t-\tau_1)s(t-\tau_2)] \approx 0$ when the time difference $\tau_1 - \tau_2$ is relatively large. As a result, the correlation shown in (12) is relatively small and we assume the two components mutually uncorrelated. To further verify this, we have computed the normalized correlation coefficients between early- and late-impulse components from natural speech utterances and these coefficients are indeed very small [34].

Consequently, the power spectrum of the early-impulse components can be estimated by subtracting the power spectrum of the late-impulse components from that of the inverse-filtered speech. The results are further used as an estimate of the power spectrum of original speech. Specifically, spectral subtraction [11] is employed to estimate the power spectrum of original speech $|S_{\tilde{x}}(k;i)|^2$

$$|S_{\tilde{x}}(k;i)|^2$$
$$= |S_z(k;i)|^2 \max\left[\frac{|S_z(k;i)|^2 - \gamma w(i-\rho) * |S_z(k;i)|^2}{|S_z(k;i)|^2}, \varepsilon\right] \qquad (13)$$

where $\varepsilon = 0.001$ is the floor and corresponds to the maximum attenuation of 30 dB.

Spectral subtraction is originally proposed to enhance speech against uncorrelated background noise, and the main issue to apply spectral subtraction is how to produce a good spectral estimate of background noise, which is different for different kinds of noise. Our use of spectral subtraction to enhance reverberant speech is motivated by the consideration that long-term reverberation, corresponding to late reflections in our two-stage formulation, may be treated as uncorrelated noise. This then leads to the specific estimation in (8), which differs from the estimate in our previous one-stage algorithm [35].

Natural speech utterances contain silent gaps, and reverberation fills some of the gaps right after high-intensity speech sections. To further improve system performance, we employ a simple method to identify and then attenuate these silent gaps. First, even with reverberation filling, the energy of a silent frame in inverse-filtered speech is relatively low. Consequently, a threshold ϑ_1 is established to identify the possibility of a silent frame. Secondly, for a silent frame, the energy is substantially reduced after the spectral subtraction process described earlier

[1] Rayleigh distribution is defined as: $f(x) = (x/a^2)\exp(-x^2/2a^2)$ for $x \geq 0$ and $f(x) = 0$ otherwise.

in this section. As a result, a second threshold ϑ_2 is established for the energy reduction ratio. Specifically, the signal is first normalized so that the maximum frame energy is 1. A time frame i is identified as a silent frame only if $E_z(i) < \vartheta_1$ and $E_z(i)/E_{\tilde{x}}(i) > \vartheta_2$, where $E_z(i)$ and $E_{\tilde{x}}(i)$ are the energy values in frame i for the inverse-filtered speech $z(t)$ and the spectral-subtracted speech $\widetilde{x}(t)$. We choose $\vartheta_1 = 0.0125$ and $\vartheta_2 = 5$. For identified silent frames, all frequency bins are attenuated by 30 dB. Finally, the short-term phase spectrum of enhanced speech is set to that of inverse-filtered speech and the processed speech is reconstructed from the short-term magnitude and phase spectrum.

Note that reliable silence detection in a reverberant environment is far from trivial. The above silence detection and attenuation method is intended to deal with those silent gaps that are relatively easy to detect. This simple method leads to a small but noticeable improvement on the output from spectral subtraction. Further improvement may be possible with a comprehensive treatment of silence detection for reverberant speech.

V. Results and Comparisons

To measure progress, it is important to quantitatively assess reverberant speech enhancement performance. Ideally, an objective speech quality measure should replicate human performance. In reality, however, different objective measures are used for different conditions.

Many objective speech quality measures (for example, see [30]) are solely based on the magnitude spectrum of speech, in part motivated by the study of Wang and Lim [33] showing that phase distortion is not important for enhancing speech mixed with a moderate level of white noise. In this situation, the phases of strong spectral components of speech are not distorted significantly since these components are much stronger than the masking noise. As a result, ignoring phase information is appropriate for noisy speech enhancement. However, this may be inappropriate for enhancing reverberant speech since the phases of strong spectral components are greatly distorted in a reverberant environment. We have conducted an informal experiment by substituting the phase of clean speech with that of reverberant speech while retaining the magnitude of clean speech. Clear reduction of speech quality is heard in comparison with original speech. Consequently, we utilize frequency-weighted segmental signal-to-noise ratio ($\mathrm{SNR}_{\mathrm{fw}}$) [32] to measure performance, which takes into account of phase information. Specifically,

$$\mathrm{SNR}_{fw} = \frac{1}{M}\sum_{j=1}^{M}\left[\frac{1}{K}\sum_{k=1}^{K}\sum_{n=m_j-N+1}^{m_j}\frac{s_k^2(n)}{[s_k(n)-\hat{s}_k(n)]^2}\right] \tag{14}$$

where $s(n)$ is the original noise- and reverberation-free signal, and $\hat{s}(n)$ is the processed signal. m_j is the end-time of the jth frame and the summation is over M frames, each of length N (we use a length of 30 ms). The signals are first filtered into K frequency bands corresponding to 20 classical articulation bands [15]. These bands are unequally spaced and have varying bandwidths. However they contribute equally to the intelligibility of a processed speech. Experiments show that frequency-weighted segmental SNR is highly correlated with subjective speech quality and is superior to conventional SNR or segmental SNR [30].

A corpus of speech utterances from eight speakers, four females and four males, randomly selected from the TIMIT database [12] is used for system evaluation. Informal listening tests show that the proposed algorithm achieves substantial reduction of reverberation and has little audible artifacts. To illustrate typical performance, we show the enhancement result of a speech signal corresponding to the sentence "she had your dark suit in greasy wash water all year" from the TIMIT database in Fig. 5. Fig. 5(a) and (c) show the clean and the reverberant signal and Fig. 5(b) and (d), the corresponding spectrograms, respectively. The reverberant signal is produced by convolving the clean signal and the room impulse response function in Fig. 2(a) with $T_{60} = 0.3$ s. As can be seen, while the clean signal has fine harmonic structure and silence gaps between the words, the reverberant speech is smeared and its harmonic structure is elongated. The inverse-filtered speech, resulting from the first stage of our algorithm, and its spectrogram are shown in Fig. 5(e) and (f), respectively. Compared with the reverberant speech, inverse filtering restores some detailed harmonic structure of the original speech, although the smearing and silence gaps are not much improved. This is consistent with our understanding that coloration mostly degrades the detailed spectrum and phase information. Finally, the processed speech using the entire algorithm and its spectrogram are shown in Fig. 5(g) and (h), respectively. As can be seen, the effects of reverberation have been significantly reduced in the processed speech. The smearing is lessened and many silence gaps are clearer.

Table I shows the systematic results for the utterances from the eight speakers. $\mathrm{SNR}_{fw}^{\mathrm{rev}}$, $\mathrm{SNR}_{fw}^{\mathrm{inv}}$, and $\mathrm{SNR}_{fw}^{\mathrm{processed}}$ denote the frequency-weighted segmental SNRs for reverberant speech, inverse-filtered speech, and processed speech, respectively. The SNR gains for inverse-filtered speech and the processed speech are represented by $\mathrm{SNR}_{fw}^{\mathrm{inv-rev}} = \mathrm{SNR}_{fw}^{\mathrm{inv}} - \mathrm{SNR}_{fw}^{\mathrm{rev}}$ and $\mathrm{SNR}_{fw}^{\mathrm{processed-rev}} = \mathrm{SNR}_{fw}^{\mathrm{processed}} - \mathrm{SNR}_{fw}^{\mathrm{rev}}$, respectively. As can be seen, the quality of the processed speech is substantially improved, with an average SNR gain of 4.82 dB over reverberant speech. We note that some slight processing artifacts can be heard as a result of the second stage processing. Such distortions are commonly observed from the processing of spectral subtraction. Nonetheless, the second stage provides a significant SNR increase and cleans inverse-filtered speech.

To put our performance in perspective, we compare with a recent one-microphone reverberant speech enhancement algorithm proposed by Yegnanarayana and Murthy [36]. We refer to this algorithm as the YM algorithm. The YM algorithm first applies gross weights to LP residual so that more severely reverberant speech segments are attenuated. Then, fine weights are applied to the residual so that they resemble more closely the damped sinusoidal patterns of LP residual from clean speech. Observing that the envelop spectrum of clean speech is flatter than that of reverberant speech, the authors modify LP coefficients to flatten the spectrum. Since the YM algorithm is implemented for speech signals sampled at 8 kHz, we downsample the speech signals from 16 kHz and adapt our algorithm to perform at 8 kHz. The results of processing the downsampled signal from Fig. 5 are shown in

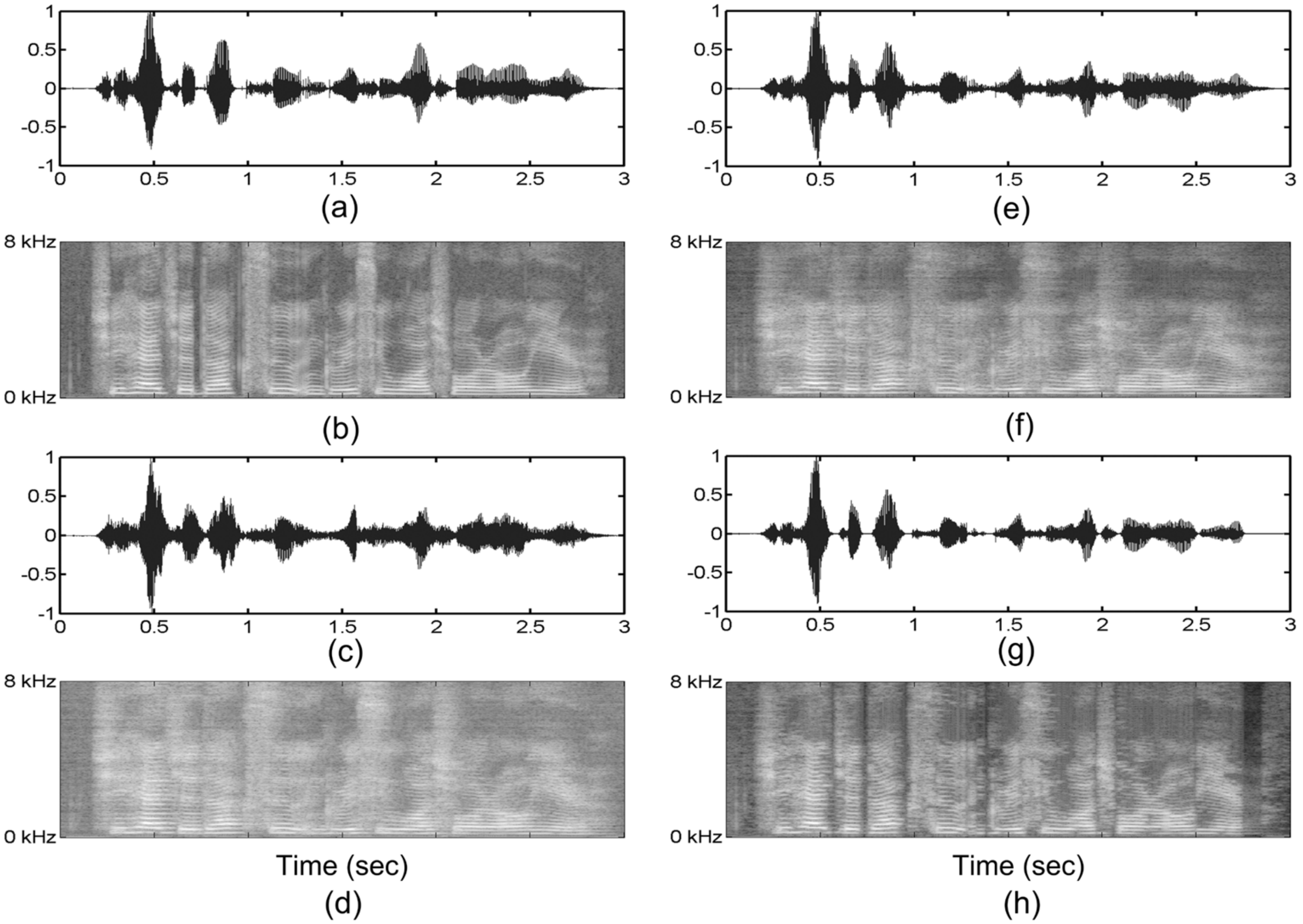

Fig. 5. Results of reverberant speech enhancement. (a) Clean speech. (b) Spectrogram of clean speech. (c) Reverberant speech. (d) Spectrogram of reverberation speech. (e) Inverse-filtered speech. (f) Spectrogram of inverse-filtered speech. (g) Speech processed using our algorithm. (h) Spectrogram of the processed speech. The speech is a female utterance "she had your dark suit in greasy wash water all year," sampled at 16 kHz.

TABLE I
SYSTEMATIC RESULTS OF REVERBERANT SPEECH ENHANCEMENT FOR SPEECH UTTERANCES OF FOUR FEMALE AND FOUR MALE SPEAKERS RANDOMLY SELECTED FROM THE TIMIT DATABASE

Speaker/Gender	SNR_{fw}^{rev}	SNR_{fw}^{inv}	$SNR_{fw}^{processed}$	$SNR_{fw}^{inv-rev}$	$SNR_{fw}^{processed-rev}$
	(dB)	(dB)	(dB)	(dB)	(dB)
Female#1	-2.62	0.01	1.84	2.63	4.46
Female#2	-2.07	0.01	1.56	2.17	3.63
Female#3	-4.28	-1.69	0.74	2.60	5.02
Female#4	-3.02	-0.90	1.07	2.12	4.09
Male#1	-4.47	-0.30	1.74	4.17	6.21
Male#2	-4.42	-0.50	1.07	3.92	5.49
Male#3	-3.23	0.66	2.01	3.90	5.24
Male#4	-3.04	-0.06	1.41	2.99	4.45
Average	-3.39	-0.33	1.43	3.06	4.82

Fig. 6. Fig. 6(a) and (c) show the clean and the reverberant signal sampled at 8 kHz and Fig. 6(b) and (d), the corresponding spectrograms, respectively. Fig. 6(e) and (f) show the processed speech using the YM algorithm and its spectrogram, respectively. As can be seen, spectral structure is clearer and some silence gaps are attenuated. The processed speech using our algorithm and its spectrogram are shown in Fig. 6(g) and (h). The figure clearly shows that our algorithm enhances the reverberant speech more than does the YM algorithm.

Quantitative comparisons are also obtained from the speech utterances of the eight speakers separately and presented in Table II.[2] $\mathrm{SNR}_{fw-8k}^{rev}$, $\mathrm{SNR}_{fw-8k}^{\mathrm{YM}}$, and $\mathrm{SNR}_{fw-8k}^{\mathrm{processed}}$ represent the frequency-weighted segmental SNR values of reverberant speech, the processed speech using the YM algorithm, and the processed speech using our algorithm, respectively. The SNR gains by employing the YM algorithm and our algorithm are denoted by $\mathrm{SNR}_{fw-8k}^{\mathrm{YM-rev}}$ and $\mathrm{SNR}_{fw-8k}^{\mathrm{processed-rev}}$, respectively. As can be seen, the YM algorithm obtains an average SNR gain of 0.74 dB compared to that of 4.15 dB by our algorithm.

Our algorithm has also been tested in reverberant environments with different reverberation times. The first stage of our algorithm—inverse filtering—is able to perform reliably with reverberation times ranging from 0.2 s to 0.4 s, which cover the reverberation times of typical living rooms. When reverberation times are greater than 0.4 s, the length of the inverse filter (64 ms) is too short to cover the long room impulse responses. On the other hand, when reverberation times are less than 0.2 s, the quality of reverberant speech is reasonably high even without processing. Unless the inverse filter is precisely estimated, inverse filtering may even degrade the reverberant speech rather than improve it. Fig. 7 shows the performance of our algorithm under different reverberation times. The dot, dash, and solid lines represent the frequency-weighted

[2] Sound files can be found at http://www.cse.ohio-state.edu/pnl/demo/WuReverb.html.

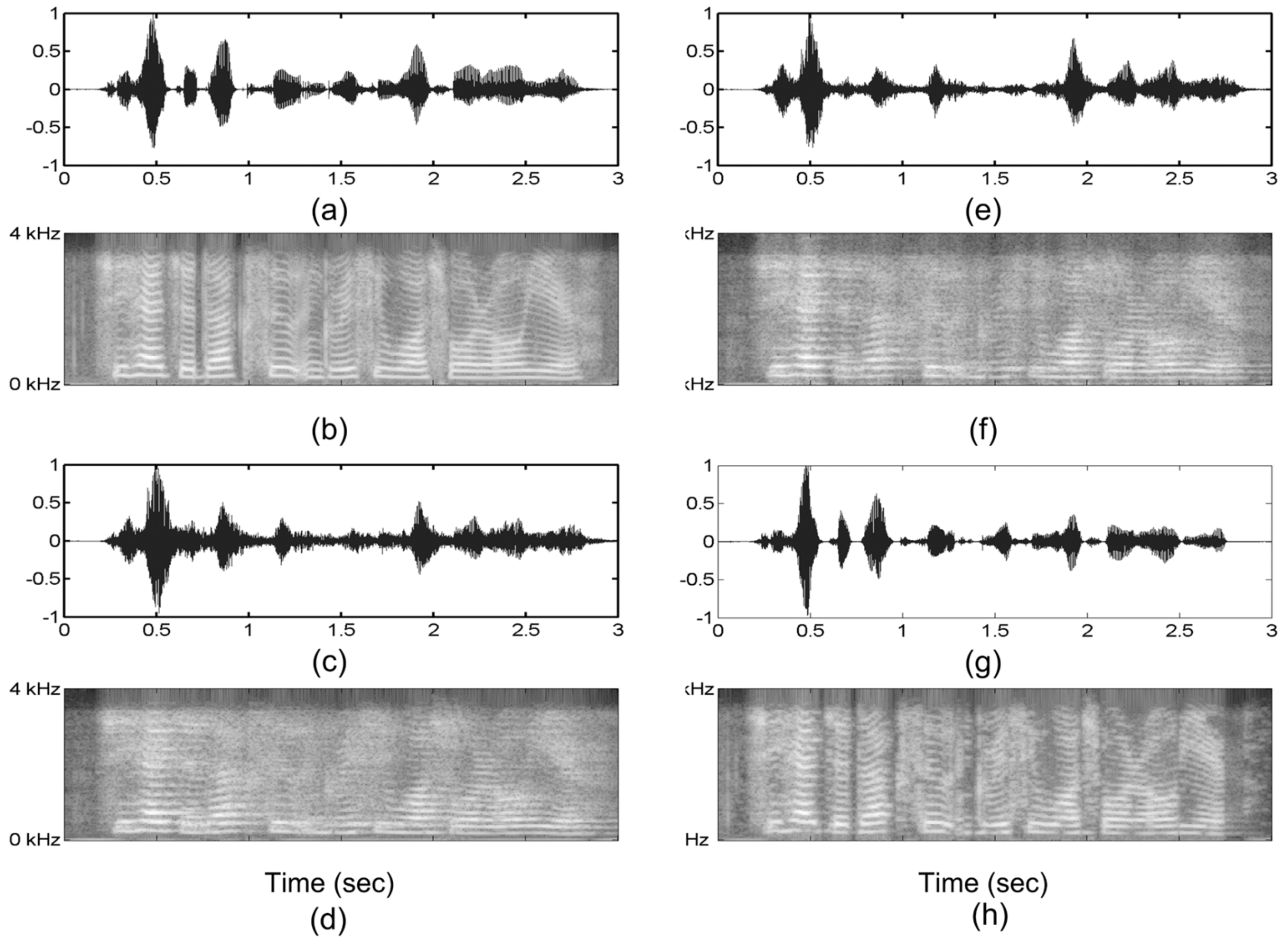

Fig. 6. Results of reverberant speech enhancement of the same speech utterance in Fig. 5 downsampled to 8 kHz. (a) Clean speech. (b) Spectrogram of clean speech. (c) Reverberant speech.(d) Spectrogram of reverberant speech. (e) Speech processed using the YM algorithm. (f) Spectrogram of (e). (g) Speech processed using our algorithm. (h) Spectrogram of (g).

TABLE II
SYSTEMATIC RESULTS OF REVERBERANT SPEECH ENHANCEMENT FOR SPEECH UTTERANCES OF FOUR FEMALE AND FOUR MALE SPEAKERS RANDOMLY SELECTED FROM THE TIMIT DATABASE. ALL SIGNALS ARE SAMPLED AT 8 kHz

Speaker/Gender	SNR^{rev}_{fw-k8}	SNR^{YM}_{fw-k8}	$SNR^{processed}_{fw-k8}$	SNR^{YM-rev}_{fw-k8}	$SNR^{processed-rev}_{fw-k8}$
	(dB)	(dB)	(dB)	(dB)	(dB)
Female#1	-3.64	-3.06	0.92	0.58	4.56
Female#2	-3.51	-3.05	0.74	0.46	4.25
Female#3	-3.86	-3.19	-0.20	0.68	3.66
Female#4	-4.12	-3.29	0.73	0.83	4.84
Male#1	-3.86	-2.65	-0.92	1.21	2.94
Male#2	-3.33	-2.68	1.77	0.65	5.10
Male#3	-3.30	-2.53	1.20	0.76	4.49
Male#4	-3.50	-2.76	-0.13	0.75	3.38
Average	-3.64	-2.90	0.51	0.74	4.15

segmental SNR values of reverberant speech, inverse-filtered speech, and the enhanced speech, respectively. As can be seen, our algorithm consistently improves the quality of reverberant speech within this range of reverberation times. Note that reverberation time can be automatically estimated by using algorithms such as the one proposed in [35].

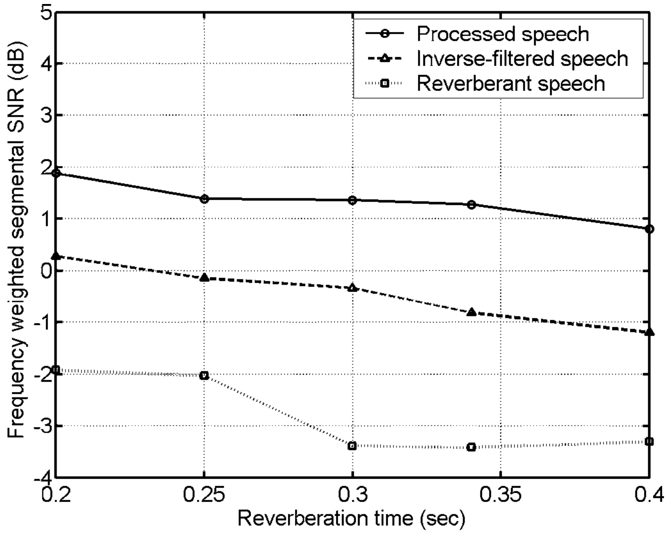

Fig. 7. Results of the proposed algorithm with respect to different reverberation times. The dot, dash, and solid lines represent the frequency-weighted segmental SNR values of reverberant speech, inverse-filtered speech, and the processed speech.

Many factors, such as reverberation time and the quality of inverse filtering, contribute to the relative strength of the late-impulse components after inverse filtering. Consequently, one expects that the scaling factor γ in (8), indicating the relative strength, should vary with respect to these factors in order to yield optimal subtraction. To study the effect of varying γ

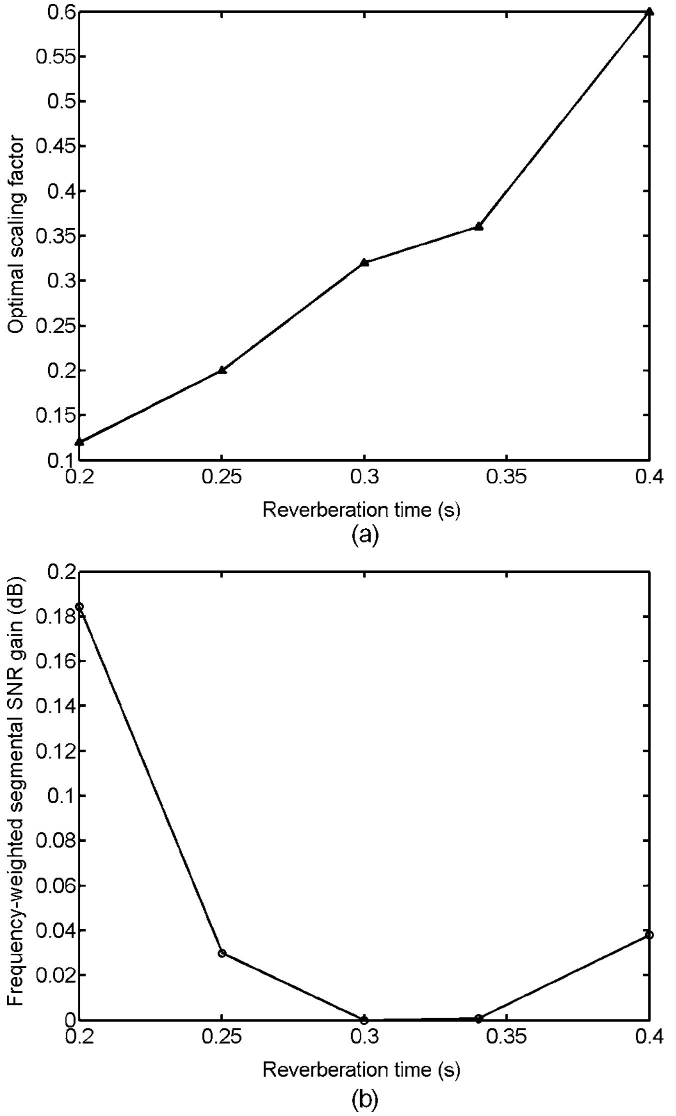

Fig. 8. (a) Optimal scaling factors with respect to reverberation times. (b) Frequency-weighted segmental SNR gains by using the optimal scaling factors instead of a fixed scaling factor.

values, the optimal scaling factors are identified by finding the maxima of frequency-weighted segmental SNR values for eight speech utterances mentioned before in different reverberant conditions. Fig. 8(a) shows these optimal values with respect to reverberation time. The optimal frequency-weighted segmental SNR gains in comparison to those derived by using the fixed scaling factor of 0.32 are shown in Fig. 8(b). As can be seen, the optimal scaling factor is positively correlated to reverberation time and ranges from 0.1 to 0.6. However, the performance gain by using the optimal factor is no greater than 0.2 dB. Informal listening tests also show that the speech quality improvement by using the optimal scaling factor is negligible. We think that the main reason is the nonstationarity of speech signal, whose energy varies widely in both spectral and temporal domains. Comparing the spectrograms of inverse-filtered and clean speech, clean speech exhibits much more pronounced energy valleys (gaps). The second stage of our system is designed to restore such valleys. The reverberant energy that fills clean speech valleys tends to originate from earlier energy peaks of clean speech, and a range of scaling factors can attenuate these valleys to the energy floor as specified in (13). As a result, the system performance is not very sensitive to specific values of γ.

It is well known that speech signal is short-term stationary but long-term nonstationary. Late reflections of reverberation have delays that exceed the period during which speech can be reasonably considered as stationary, and as a result, they smear speech spectra as discussed in Section II. Early reflections, on the other hand, have delays within this period. Because of the short-term stationarity of speech, early reflections and direct-path signal have similar magnitude spectra. Consequently, early reflections cause coloration distortion and increase the intensity of reverberant speech. The time delay that separates early from late reflections is, hence, not a property of room impulse response; instead, it is a property of the source signal and indicates the boundary between short-term stationarity and long-term nonstationarity. For instance, music signal tends to change less rapidly than speech and, as a result, the delay that separates early and late reflections is longer for music signal. Considering average properties of speech, the delay separating early and late reflections is commonly set at 50 ms [24]. This translates to $\rho = 7$ specified in Section IV. This explanation implies that the choice of ρ should not depend on room reverberation time.

The selection of the parameter a in the Rayleigh smoothing function of (9) is subject to two primary constraints. On the left side (see Fig. 4), the function needs to quickly drop to 0 with $a < \rho$. On the right side, the smoothing function should follow the reverberation tail and therefore reflect the reverberation time. Under these constraints, a is set to five as specified before. We observe little improvement by adjusting the value of a.

If the reverberation time is outside the range of 0.2 to 0.4 s, the reverberant speech should be handled differently. For reverberation time from 0.1 s to 0.2 s, the second stage of our algorithm—estimating and subtracting the late-impulse components—can be applied directly without passing through the first stage. Speech utterances from eight speakers described before are employed for evaluation. Our experiments show that, under reverberation times of 0.12 and 0.17 s, the second stage of our algorithm with a scaling factor of 0.05 improves the average frequency-weighted segmental SNR values from 3.89 and 1.36 dB of reverberant speech to 4.38 and 2.55 dB of the processed speech, respectively. For reverberation times lower than 0.1 s, the reverberant speech already has very high quality and no enhancement is necessary. For reverberation times greater than 0.4 s, one could also directly use the second stage of our algorithm. To see its effects, we perform further experiments using a scaling factor of 2.0 and employing the speech utterances used before. Utilizing the utterances from the same eight speakers, our experiments show that, with $T_{60} = 0.58$ s, average frequency-weighted segmental SNR improves from -5.7 dB of reverberant speech to -1.4 dB of the processed speech.

VI. Discussion and Conclusion

Many algorithms for reverberant speech enhancement utilize FIR filters for inverse filtering. The length of an FIR inverse

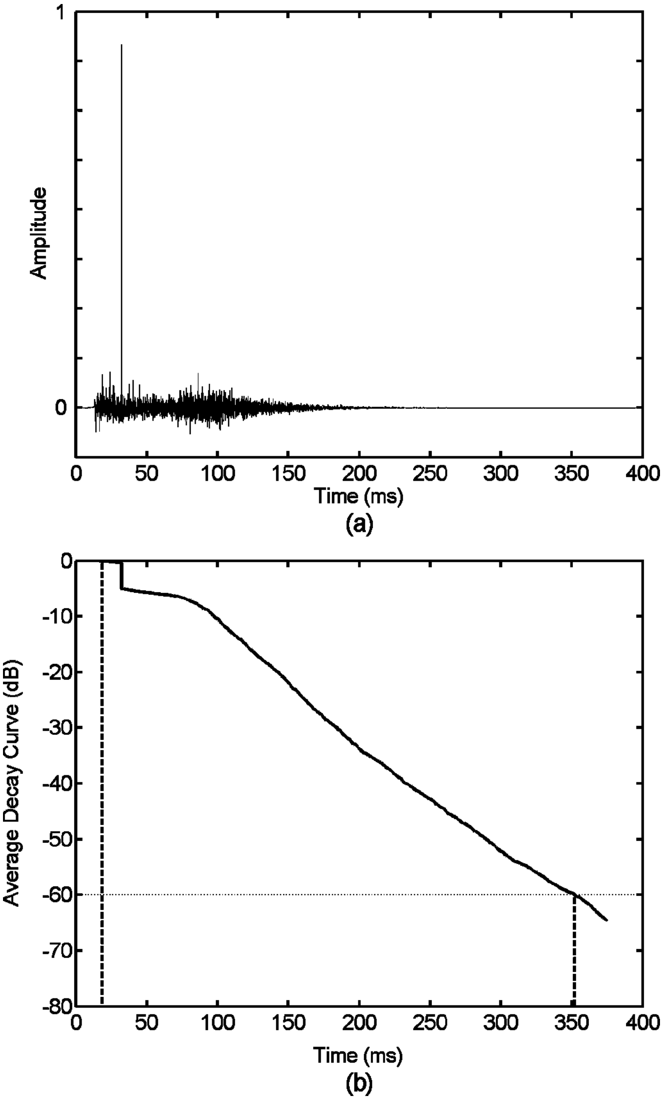

Fig. 9. (a) The equalized impulse response derived from the room impulse response in Fig. 2(a) using linear least-square inverse filtering of length 1024 (64 ms). (b) Its energy decay curve computed using the Schroeder integration method. The horizontal dot line represents -60 dB energy decay level. The left dash line indicates the starting time of the impulse responses and the right dash line the time at which decay curves crosses -60 dB.

filter, however, puts limitation on the system performance. For example, Fig. 9(a) shows the equalized impulse response derived from the room impulse response in Fig. 1 ($T_{60} = 0.3$ s) using linear least-square inverse filtering [17]. This technique derives an optimal FIR inverse filter in the least-square sense for length 1024 (64 ms) with the perfect knowledge of the room impulse response. The corresponding energy decay curve computed according to the Schroeder integration method [31] is shown in Fig. 9(b). As can be seen, the impulses after 70 ms from the starting time of the equalized impulse response are not much attenuated. Some remedies have been investigated. For example, Gillespie and Atlas proposed a binary-weighted linear-least-square equalizer [17], which attenuates more long-term reverberation at the expense of lower SRR values. However, because the length of the inverse filter is shorter than the length of reverberation, the reverberation longer than the filter cannot be effectively reduced in principle. In theory, longer FIR inverse filters may achieve better performance. However, long inverse filters introduce many more free parameters that are often difficult to estimate in practice. Sometimes, it leads to instability of convergence and often requires a large amount of training data. A few algorithms have been proposed to derive long FIR inverse filters. For example, Nakatani and Miyoshi [27] proposed a system capable of blind dereverberation of one-microphone speech using long FIR filters (2 s, personal communication, 2003). To configure this long FIR filter, a large amount of training data (5240 Japanese words) are needed for good results and the room impulse response cannot change during the entire time period. This implies that the listener and the speech source are fixed for a very long period of time, which is hardly realistic. In many practical situations, however, only relatively short FIR inverse filters can be derived. In this case, the second stage of our algorithm can be used as an add-on to many inverse-filtering based algorithms.

Although our algorithm is designed for enhancing reverberant speech using one microphone, it is straightforward to extend it into multimicrophone scenarios. Many inverse filtering algorithms, such as the algorithm by Gillespie *et al.* [18], are originally proposed using multiple microphones. After inverse filtering using multiple microphones, the second stage of our algorithm—the spectral subtraction method—can be utilized for reducing long-term reverberation effects.

Araki *et al.* [4] point out a fundamental performance limitation of the frequency domain BSS algorithms. When a room impulse response is long, the frame length of FFT used for frequency domain BSS needs to be long in order to cover the long reverberation. However, when a mixture signal is short, the lack of data in each frequency channel caused by the longer frame size triggers the collapse of the assumption of independence of source signals. Under these constraints, one can identify a frame length of FFT to achieve the optimal performance of a frequency domain BSS system. This optimal length, however, is comparatively short with a long room impulse response. For example, in one of their experiments, the optimal frame length is 1024 (64 ms) for a convolutive BSS system in a room with the reverberation time of 0.3 s. Consistent with the argument we offered earlier, a BSS system employing the optimal frame length is unable to attenuate long-term reverberation effects of either target or interfering sound sources. On the other hand, the second stage of our algorithm can be extended to deal with multiple sound sources by applying a convolutive BBS system and then reducing long-term reverberation effects.

Our algorithm is also robust to modest levels of background noise. We have tested our algorithm on reverberant utterances mixed with white noise so that the SNRs of reverberant speech, where the reverberant speech is treated as signal, are 20 dB. The results show that our method consistently reduces reverberation effects and yields an average SNR gain similar to that without background noise [34].

To conclude, we have presented a new two-stage reverberant speech enhancement algorithm using one microphone, and the stages correspond to inverse filtering and spectral subtraction. The first-stage aims to reduce coloration effects caused by early reflections, and inverse filtering helps to improve the magnitude spectrum of reverberant speech and reduce phase distortions especially in the strong spectral components. The second-stage aims to reduce long-term reverberation, and spectral subtraction helps to further improve the magnitude spectrum. The evaluations show that our algorithm enhances the quality of reverberant speech effectively and performs significantly better than a recent reverberant speech enhancement algorithm.

Acknowledgment

The authors would like to thank the anonymous reviewers for their useful suggestions.

References

[1] J. B. Allen, "Effects of small room reverberation on subjective preference," *J. Acoust. Soc. Amer.*, vol. 71, p. S5, 1982.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[3] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *J. Acoust. Soc. Amer.*, vol. 62, pp. 912–915, 1977.

[4] S. Araki, R. Mukai, S. Makino, T. Nishikara, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[5] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. ICSLP*, 1996, pp. 889–892.

[6] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, 1991, pp. 977–980.

[7] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind source separation and blind deconvolution," *Neur. Computation*, vol. 7, pp. 1129–1159, 1995.

[8] D. A. Berkley and J. B. Allen, "Normal listening in typical rooms: the physical and psychophysical correlates of reverberation," in *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed, G. A. Studebaker and I. Hochberg, Eds. Needham Heights, MA: Allyn and Bacon, 1993, pp. 3–14.

[9] M. S. Brandstein and S. Griebel, "Explicit speech modeling for microphone array applications," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. New York: Springer Verlag, 2001, pp. 133–153.

[10] M. S. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer Verlag, 2001.

[11] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Upper Saddle River, NJ: Prentice-Hall, 1987.

[12] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Speech Recognition Workshop*, 1986, pp. 93–99.

[13] J. L. Flanagan, J. D. Johnson, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, pp. 1508–1518, 1985.

[14] J. L. Flanagan, A. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Commun.*, vol. 13, pp. 207–222, 1993.

[15] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, pp. 90–119, 1947.

[16] K. Furuya and Y. Kaneda, "Two-channel blind deconvolution for non-minimum phase impulse responses," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, 1997, pp. 1315–1318.

[17] B. W. Gillespie and L. E. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, 2002, pp. 557–560.

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

[19] B. Gold and N. Morgan, *Speech and Audio Signal Processing*. New York: Wiley, 2000.

[20] S. Haykin, *Adaptive Filter Theory*, 4th ed. Upper Saddle River, N.J.: Prentice-Hall, 2002.

[21] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Amer.*, vol. 77, pp. 1069–1077, 1985.

[22] J. J. Jetzt, "Critical distance measurement of rooms from the sound energy spectral response," *J. Acoust. Soc. Amer.*, vol. 65, pp. 1204–1211, 1979.

[23] A. H. Koenig, J. B. Allen, D. A. Berkley, and T. H. Curtis, "Determination of masking level differences in an reverberant environment," *J. Acoust. Soc. Amer.*, vol. 61, pp. 1374–1376, 1977.

[24] H. Kuttruff, *Room Acoustics*, 4th ed. New York: Spon, 2000.

[25] M. Miyoshi and Y. Kaneda, "Inverse filtering of room impulse response," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[26] A. K. Nábelek, "Communication in noisy and reverberant environments," in *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed, G. A. Stubebaker and I. Hochberg, Eds. Needham Heights, MA: Allyn and Bacon, 1993.

[27] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, 2003, pp. 92–95.

[28] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, pp. 165–169, 1979.

[29] K. J. Palomäki, G. J. Brown, and J. Barker, "Missing data speech recognition in reverberant conditions," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, Orlando, FL, 2002, pp. 65–68.

[30] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[31] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, pp. 409–412, 1965.

[32] J. M. Tribolet, P. Noll, and B. J. McDermott, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, Tulsa, OK, 1978, pp. 586–590.

[33] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 30, no. 4, pp. 679–681, Aug. 1982.

[34] M. Wu, "Pitch tracking and speech enhancement in noisy and reverberant environments," Ph.D. dissertation, Dept. Comput. Inf. Sci., Ohio State Univ., Columbus, 2003.

[35] M. Wu and D. L. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, 2003, pp. 844–847.

[36] 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.

Mingyang Wu received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 1995, and the M.S. and Ph.D. degrees in computer science and engineering from The Ohio State University, Columbus, in 1999 and 2003, respectively.

He is currently with the Fair Isaac Corporation, San Diego, CA. His research interests include machine learning, neural networks, speech processing, and computational auditory scene analysis.

DeLiang Wang (M'90–SM'01–F'04) received the B.S. and M.S. degrees from Peking (Beijing) University, Beijing, China, in 1983 and 1986, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, in 1991, all in computer science.

From July 1986 to December 1987, he was with the Institute of Computing Technology, Academia Sinica, Beijing. Since 1991, he has been with the Department of Computer Science and Engineering and the Center for Cognitive Science at Ohio State University, Columbus, where he is currently a Professor. From October 1998 to September 1999, he was a Visiting Scholar in the Department of Psychology at Harvard University, Cambridge, MA. His research interests include machine perception and neurodynamics.

Dr. Wang is the President of the International Neural Network Society. He is a recipient of the 1996 U.S. Office of Naval Research Young Investigator Award.