A TWO-STAGE ALGORITHM FOR ENHANCEMENT OF REVERBERANT SPEECH

Mingyang Wu and *DeLiang Wang*
Department of Computer Science and Engineering
and Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
Email: mwu@fairisaac.com, dwang@cse.ohio-state.edu

ABSTRACT

Room reverberation causes two perceptual distortions on clean speech: Coloration and long-term reverberation. These two effects correspond to two physical variables: Signal-to-reverberant energy ratio (SRR) and reverberation time, respectively. Based on this observation, we propose a two-stage algorithm that enhances reverberant speech from one-microphone recordings. 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 one-microphone enhancement algorithm shows that our system produces significantly better results.

1. INTRODUCTION

A main cause of speech degradation in practically all listening situations is room reverberation. Although a person with normal hearing is little affected by room reverberation to a considerable degree, hearing-impaired listeners suffer from reverberation effects disproportionally [12]. Also, reverberation causes significant performance decrement for current automatic speech recognition (ASR) and speaker recognition systems. Consequently, an effective reverberant speech enhancement system can be used for improving intelligent hearing aids design and is essential for many speech technology applications.

In this article 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 hand-free audio communication and audio information retrieval. Second, moderately reverberant speech is highly intelligible in monaural listening conditions. Hence how to achieve this monaural capability remains a fundamental scientific question.

A number of reverberant speech enhancement algorithms have been designed utilizing more than one microphone. For example, microphone-array based methods [6], 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. [10], 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. Blind deconvolution algorithms aim to reconstruct the inverse filters without the prior knowledge of room impulse responses (for example, see [8]). Brandstein and Griebel [5] utilize the extrema of wavelet coefficients to reconstruct the linear prediction (LP) residual of original speech.

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. [3] 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 [2]) are motivated by the effects of reverberation on Modulation Transfer Function (MTF). Yegnanarayana and Murthy [16] 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 [13] 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.

2. BACKGROUND

Reverberation causes a noticeable change in speech quality. Berkley and Allen [4] identified that two physical variables, reverberation time T_{60} and spectral deviation, are important for reverberant speech quality. Consider the impulse response as a combination of three parts, the direct, early, and late reflections. While late reflections smear the speech spectra and reduce the intelligibility and quality of speech signals, early reflections cause another distortion of speech signal called coloration; the non-flat 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. Increasing either spectral deviation or reverberation time results in decreased reverberant speech quality. Moreover, Jetzt [9] shows that spectral deviation is

determined by signal-to-reverberant energy ratio (SRR), which is the ratio between the energy traveling directly from a source to a listener and the energy of all acoustic reflections reaching the listener, and in turn, it is determined by talker-to-microphone distance. Shorter talker-to-microphone distance results in higher SRR and less spectral deviation, hence, less coloration.

Consequently, we propose a two-stage model to deal with two types of degradations – coloration and long-term reverberation – in a reverberant environment. In the first stage, our model estimates an inverse filter to reduce coloration effects in order to increase SRR. The second stage employs spectral subtraction to minimize the influence of long-term reverberation.

3. INVERSE FILTERING

In the first stage of our algorithm, we derive an inverse filter to reduce the reverberation effects and this stage is adapted from a multi-microphone inverse filtering algorithm proposed by Gillespie at el. [8]. An FIR inverse filter of the room impulse response is estimated by maximizing the kurtosis of the linear prediction (LP) residual of speech utilizing a block frequency-domain adaptive filter. Then, inverse-filtered speech is obtained by convolving the inverse filter with reverberant speech.

A typical result from the first stage of our algorithm is shown in Fig. 1. Fig. 1(a) illustrates a room impulse response function (T_{60} = 0.3 s) generated by the image model of Allen and Berkley [1]. The equalized impulse response – the result of the room impulse response in Fig. 1(a) convolved with the obtained inverse filter – is shown in Fig. 1(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. 1(c) and (d) 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.

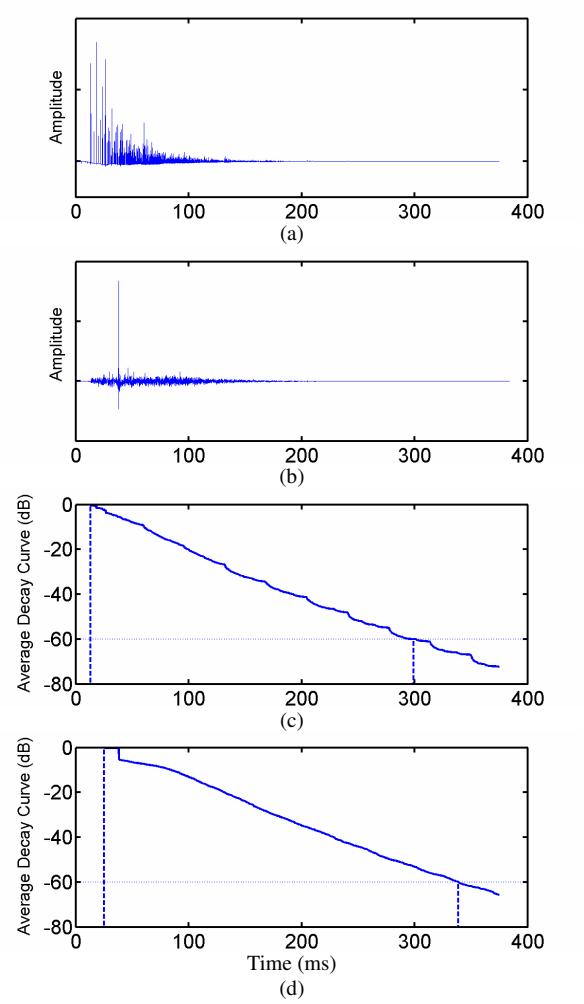

Fig. 1. (a) A room impulse response function generated by the image model in an office-size room. (b) The equalized impulse response derived from the reverberant speech generated by the room impulse response in (a) as the result of the first stage of our algorithm. Energy decay curves (c) that computed from the room impulse response function in (a). (d) That from the equalized impulse response in (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.

3. SPECTRAL SUBTRACTION

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

In a previous version of this algorithm, Wu and Wang [15] propose a one-stage method to enhance the reverberant speech by estimating and subtracting effects of late reflections.

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 $z(t)$:

$$|S_l(k;i)|^2 = \gamma w(i-\rho) * |S_z(k;i)|^2, \qquad (1)$$

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.

Table I. The 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_{fw}^{rev} (dB)	SNR_{fw}^{YM} (dB)	SNR_{fw}^{proc} (dB)	SNR_{fw}^{YM-rev} (dB)	$SNR_{fw}^{proc-rev}$ (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

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 [11]. This delay reflects the properties of speech and is independent from reverberation characteristics. Consequently, it translates to approximately 7 frames for a shift interval of 8 ms, and we choose $\rho = 7$ as a result. Finally, the scaling factor γ specifies the relative strength of the late-impulse components after inverse filtering and we set it to 0.32.

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

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

where we choose $a = 5$ and it controls the span of the smoothing function. This smoothing function goes down to zero on the left side quickly but tails off slowly on the right side; the right side of the smoothing function resembles the shape of reverberation tails in equalized impulse responses.

Assuming the early- and late-impulse components are approximately uncorrelated., 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 [7] is employed to estimate the power spectrum of original speech $|S_{\tilde{x}}(k;i)|^2$:

$$|S_{\tilde{x}}(k;i)|^2 = |S_z(k;i)|^2 \max\left[\frac{|S_z(k;i)|^2 - \gamma w(i-\rho) * |S_z(k;i)|^2}{|S_z(k;i)|^2}, \varepsilon\right], \qquad (3)$$

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

Natural speech utterances contain silent gaps, and reverberation fills some of the gaps right after high-intensity speech sections. We identify these silent gaps by examine the energy of inverse-filtered speech and energy reduction radio after spectral subtraction in a time frame. 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.

3. RESULTS AND DISCUSSIONS

A corpus of speech utterances from eight speakers, four females and four males, randomly selected from the TIMIT database 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 results in Fig. 2. Fig. 2(a) and (c) show the clean and the reverberant signal and Fig. 2(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. 1(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.

To put our performance in perspective, we compare with a recent one-microphone reverberant speech enhancement algorithm proposed by Yegnanarayana and Murthy [16]. We refer to this algorithm as the YM algorithm. The YM algorithm applies weights to LP residual so that they resemble more closely the damped sinusoidal patterns of LP residual from clean speech. Fig. 2(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. 2(g) and (h). 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. The figure clearly shows that our algorithm enhances the reverberant speech more than does the YM algorithm. An audio demonstration also can be found at http://www.cse.ohio-state.edu/~dwang/demo/WuReverb.html.

Quantitative comparisons are obtained from the speech utterances of the eight speakers separately utilizing frequency-weighted segmental SNR [14] and presented in Table I. SNR_{fw}^{rev}, SNR_{fw}^{YM}, and SNR_{fw}^{proc} 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 SNR_{fw}^{YM-rev} and $SNR_{fw}^{proc-rev}$, 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.

Although our algorithm is designed for enhancing reverberant speech using one microphone, it is straightforward to extend it into multi-microphone scenarios. Many inverse filtering algorithms, such as the algorithm by Gillespie et al. [8], 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.

To conclude, we have presented a two-stage reverberant speech enhancement algorithm using one microphone, and the stages correspond to inverse filtering and spectral subtraction. The evaluations show that our algorithm enhances the quality of reverberant speech effectively and performs significantly better than a recent reverberant speech enhancement algorithm.

Acknowledgments This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (FA9550-04-1-0117).

REFERENCES

[1] 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.

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

[3] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Proc. IEEE ICASSP*, 1991, pp. 977-980.

[4] 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*, G. A. Studebaker and I. Hochberg, Eds., 2nd ed., Needham Heights, MA: Allyn and Bacon, 1993, pp. 3-14.

[5] 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, NY: Springer Verlag, 2001, pp. 133-153.

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

[7] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*, Upper Saddle River, NJ: Prentice-Hall, 1987.

[8] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE ICASSP*, 2001, pp. 3701-3704.

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

[10] 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.

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

[12] A. K. Nábelek, "Communication in noisy and reverberant environments," in *Acoustical factors affecting hearing aid performance*, G. A. Stubebaker and I. Hochberg, Eds., 2nd ed., Needham Height, MA: Allyn and Bacon, 1993.

[13] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. IEEE ICASSP*, 2003, pp. 92-95.

[14] J. M. Tribolet, P. Noll, and B. J. McDermott, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE ICASSP*, Tulsa, OK, 1978, pp. 586-590.

[15] M. Wu and D. L. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. IEEE ICASSP*, 2003, pp. 844-847.

[16] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 267-281, 2000.

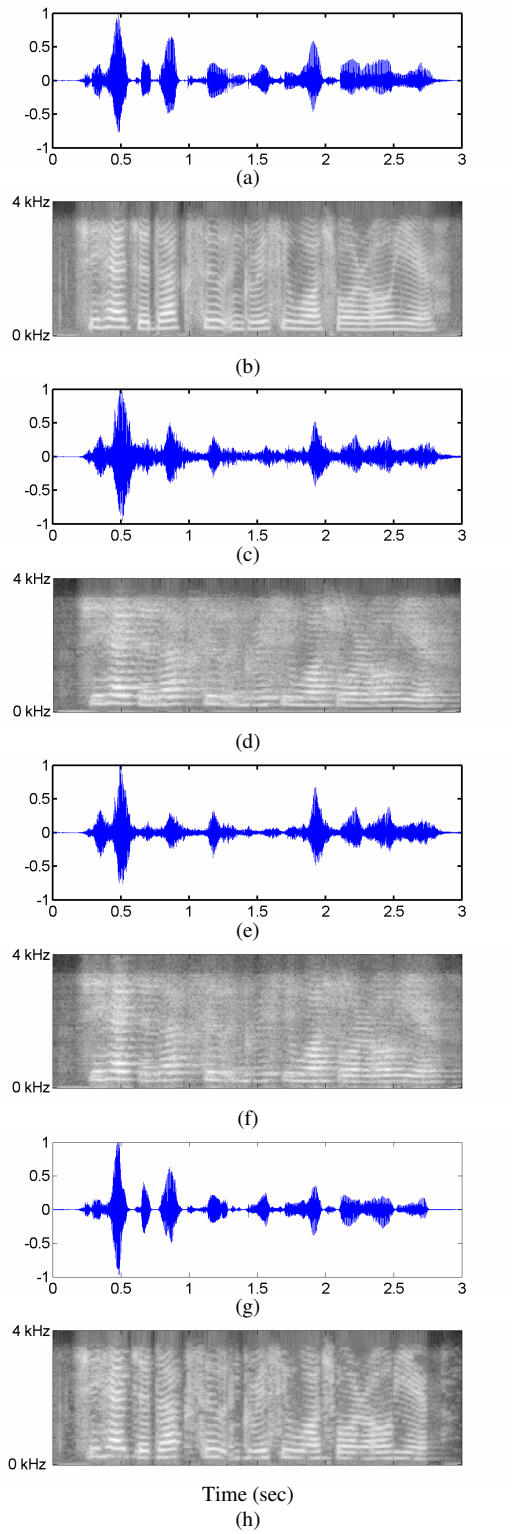

Fig. 2. Results of reverberant speech enhancement: (a) clean speech, (b) spectrogram of clean speech, (c) reverberant speech, (d) spectrogram of reverberant speech, (e) speech processed using the YM algorithm, (f) spectrogram of (e), (g) speech processed using our algorithm, and (h) spectrogram of (g). The speech is a female utterance "She had your dark suit in greasy wash water all year," sampled at 8 kHz.