A ONE-MICROPHONE ALGORITHM FOR REVERBERANT SPEECH ENHANCEMENT

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

ABSTRACT

We present an algorithm for reverberant speech enhancement using one microphone. We first propose a novel pitch-based reverberation measure for estimating reverberation time (RT60) based on the distribution of relative time lags. This measure of pitch strength correlates with reverberation and decreases systematically as detrimental effects of reverberation on harmonic structure increase. Then a reverberant speech enhancement method is developed to estimate and subtract later echo components. The results show that our approach appreciably reduces reverberation effects.

1. INTRODUCTION

Two causes of degradation in speech exist in practically all listening situations: background noise and room reverberation. Many techniques such as spectral subtraction, adaptive noise cancellation, and comb filtering have been developed to improve the perceived quality of speech degraded by background noise, and are effective in low to moderate noise level [7]. Alternatively, computational auditory scene analysis systems treat background noise as distinct sound sources and segregate acoustic waveforms into different streams representing different sources, therefore are capable of segregating speech from noise interference and speech utterances from each other (for example, see [10]).

Most algorithms developed to enhance reverberant speech utilize more than one microphone. Microphone array based methods [5] attempt to suppress the sound energy coming from directions other than the direct source and therefore enhance target speech. Other methods, such as the system developed by Gillespie et al. [6], employ prior knowledge of speech signal distribution to estimate an inverse filter of the room impulse response. These approaches require the source (loudspeaker) and the microphones to be static. Brandstein [3] simulates the effect of moving a source within a few centimeters range and concludes that effective systems applying inverse filters have to update the filters on a frame-by-frame basis.

Reverberant speech enhancement using one microphone has also been studied. A cesptrum-based method is employed by Bees et al. [2] to estimate reverberation impulse response, and then its inverse is used to dereverberate the signal. Yegnanarayana and Murthy [12] develop a reverberant speech enhancement system by manipulating LP residual signals based on the residual characteristics of clean speech. Single microphone approaches, however, only achieve moderate success on dereverberation.

In this paper, we propose a robust algorithm for reverberant speech enhancement using one microphone. A pitch-based measure is employed for estimating the reverberation time, and a method based on estimating and subtracting later echo components is developed for enhancement of reverberant speech.

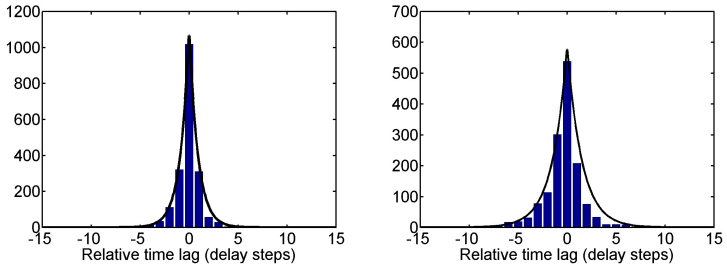
2. MODEL DESCRIPTION

The proposed model consists of two stages. In the first stage, described in Section 2.1, a pitch-based reverberation measure is developed for estimating the reverberation time. Then, in the second stage of the model, described in Section 2.2, we develop a method of reverberant speech enhancement using the reverberation time estimated in the first stage.

Many tasks require a robust measure on degraded speech indicating the degree of reverberation. For example, Yegnanarayana and Murthy [12] employ the kurtosis of LP residual signal as a measure to estimate signal-to-reverberation component ratio in a time frame. Extending this idea, Gillespie et al. [6] utilize the kurtosis as an optimization criterion to derive an inverse filter and therefore to dereverberate the degraded signal.

Reverberation corrupts harmonic structure in voiced speech, and we find that the degree of corruption can be used as a measure of reverberation. Brandstein [4] employs a criterion of signal periodicity for time-delay estimation using microphone arrays. The criterion, indicating the degree of speech signal influenced by the detrimental effect of noise and reverberation, is used to weight generalized cross-correlations across all time frames. As a result, the weights of time frames with less degradation are increased relatively and the robustness of the system is improved. However, this criterion measures the influence from both noise and reverberation. Our goal is to develop a pitch-based measure on the degree of reverberation. It is robust to noise and can be used for estimating key parameters of room impulse response such as the reverberation time (RT60).

Reverberation corrupts the speech by blurring its temporal structure. However, due to the spectral continuity of speech, the early echoes in the reverberation mainly increase the intensity of the reverberant speech, whereas the later ones are deleterious to speech quality and intelligibility [8]. Estimating the effects of



(a) (b)

Fig. 1. Histograms and estimated distributions of relative time lags in channel 22 (center frequency = 264 Hz) of (a) clean speech, and (b) reverberant speech with reverberation time of 0.3 s. The bar graphs represent histograms and the solid lines represent the estimated distributions.

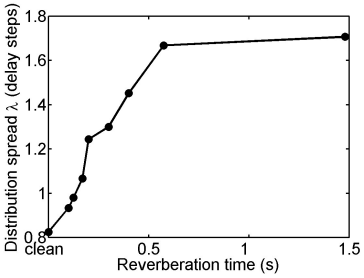


Fig. 2. Average distribution spread λ of relative time lag with respect to reverberation time.

later echoes and subtracting them from the reverberant speech should enhance the speech quality.

2.1. A pitch-based reverberation measure

Speech contains three types of time frames: voiced, unvoiced, and silence. A pitch-based measure of reverberation should be based only on voiced time frames. Moreover, in a noisy background, some frequency channels in a voiced frame may be severely corrupted by noise. This measure should be based on the signals from "clean" frequency channels.

In order to satisfy these criteria, our measure, to be detailed below, is extended from a recent multi-pitch tracking algorithm [11]. This algorithm can track pitch periods reliably and can be used to provide voiced/unvoiced labeling. Also, it gives a channel selection method for identifying weakly corrupted frequency channels from which the pitch-based measure is extracted.

The pitch-tracking algorithm consists of four stages. In the first stage, the input signals are sampled at 16 kHz and then filtered into 128 frequency channels by fourth-order gammatone filters [9]. Channels with center frequencies lower than 800 Hz (channels 1-55) are categorized as low-frequency ones, others (channels 56-128) as high-frequency ones. Envelopes are extracted in high-frequency channels. At the end of the first stage, normalized correlograms are computed using a window size of 16 ms in all channels. Channel and peak selection forms the second stage. Based on the shapes of normalization correlograms, only channels weakly corrupted by noise are selected and passed to later processing. The third stage integrates periodicity information across all channels and the final stage forms continuous pitch tracks using a hidden Markov model. A revised version of the algorithm restricted to only one pitch track is used for the present study, which deals with single speech sources.

We observe that the differences between the pitch periods determined by the pitch tracker and the time lag from the closest peaks of normalized correlograms in selected channels indicate the level of degradation in the harmonic structure. More specifically, relative time lag Δ is defined as the distance from the detected pitch period to the closest peak in correlograms. We then collect the Δ statistics from the selected channels across all voiced frames from 16 clean speech utterances chosen from the TIMIT database for every channel separately. As a typical example, the Δ histogram for channel 22 is shown in Fig. 1(a). As can be seen, the distribution is sharply centered at zero.

We propose to use the spread of the distribution as an indication of reverberation because it measures the "cleanness" of harmonic structure in speech signals. A signal composed of an ideal stationary harmonic structure is very clean. In this case, the relative time lags collected from the signal have the same value of zero, and the distribution has zero spread. Due to the nonstationary nature of speech, the distribution spread of clean speech shown in Fig. 1(a) is greater than zero.

Room reverberation corrupts harmonic structure, and echoes from natural speech tend to spread the distribution of relative time lags. To illustrate this, we collect the statistics of relative time lags from reverberant speech generated by convolving clean speech with a room impulse response function of 0.3 sec reverberation time. The histogram is shown in Fig. 1(b). The spread is wider than that of clean speech.

In order to measure the distribution spread, we employ a mixture of a Laplacian and a uniform distribution for modeling the distribution in channel *c* (see [11]):

$$p_c(\Delta) = (1-q)\frac{1}{2\lambda_c}\exp\left(-\frac{|\Delta|}{\lambda_c}\right) + qU(\Delta;\eta_c), \qquad (1)$$

where $0 < q < 1$ is a partition coefficient of the mixture and λ_c is the Laplacian distribution parameter. $U(\Delta;\eta_c)$ is a uniform distribution with range η_c. In a low-frequency channel, we set the length of the range as the wavelength of the center frequency.

We also assume a linear relationship between the frequency channel index *c* and the Laplacian distribution parameter λ_c,

$$\lambda_c = a_0 + a_1 c\ . \qquad (2)$$

The maximum likelihood method is utilized to estimate the three parameters a_0, a_1, and q in low-frequency channels. The estimated distributions of relative time lags in clean and reverberant speech are also shown in Fig. 1(a) and 1(b). As can be seen, the model distributions fit the histograms very well.

Finally, the measure of distribution spread λ is defined as the average of parameters λ_c in low-frequency channels, for harmonic structure of clean speech in low-frequency channels is more stable than that in high-frequency ones. Fig. 2 shows the relationship of λ and reverberation time. Here, the reverberant signals are generated by convolving the same 16 clean speech signals with room impulse response functions of various reverberation times obtained from the image model [1]. As can

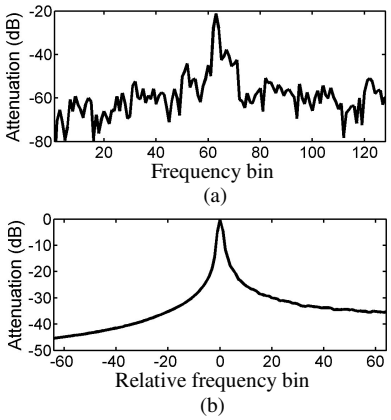


Fig. 3. (a) Magnitudes of reverberation inter-frame influence from an original signal at frequency bin 62 and a time frame *i* to time frame *i*+17, and (b) average pattern of the magnitudes of inter-frame influence.

be seen, the plot is monotonic and therefore the relative time lag spread λ can be used to estimate the reverberation time.

2.2. Enhancement of reverberant speech

The reverberant signal received at a microphone, $y(t)$, can be modeled as:

$$y(t) = h(t) * x(t), \tag{3}$$

where $x(t)$ is the original speech signal and $h(t)$ an FIR room impulse response.

Then, short-term Fourier analysis is applied to the signals using non-overlapping rectangular window of length N. Because of the linearity of Equation 3 and the causality of impulse responses, the short-term spectrum of the reverberant signal is derived as:

$$S_y(k_y;i_y) = \sum_{d=0}^{D}\sum_{k=0}^{N-1} I(k,k_y;d) S_x(k;i_y-d), \tag{4}$$

where $S_x(k;i)$ is the short-term speech spectrum of the original signal, and indexes k and i refer to frequency bins and time frames, respectively.

Reverberation inter-frame influence $I(k,k_y;d)$ represents the influence from time-frequency bin $(k;i-d)$ of the original signal to bin $(k_y;i)$ of the reverberant signal, and it can be computed from the room impulse response $h(t)$ directly. The time-frequency bin $(k_y;i_y)$ of the reverberant signal is only affected by bins of the original signal that are in time frames between time frame i_y and $i_y - D$, where D is determined by the length of room impulse response.

The reverberation inter-frame influence $I(k,k_y;d)$ has two components: magnitude and phase. The effects of moving the source with the centimeter range away are simulated also using the image model [1]. We find that the phases of inter-frame influence are highly susceptible to source positions, while the magnitudes vary only moderately. Therefore, it is impractical to use the phases of inter-frame influence since they are unstable in real environments. Here we employ only magnitude in this study.

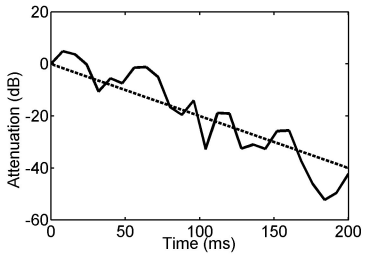


Fig. 4. Peak magnitude attenuation with respect to time delay.

In implementation, our speech enhancement system uses hamming windows of length 16 ms with 8 ms overlap for short-term Fourier analysis. Magnitudes of inter-frame influence are computed by simulating reverberation effects of sinusoids of unit energy in a time frame on a later time frame. As an example, Fig. 3(a) shows the magnitudes of reverberation inter-frame influence from an original signal at frequency bin 62 and time frame i to time frame $i+17$ (136 ms later), i.e., $|I(62,k_y;17)|$. As shown in the figure, the magnitudes of inter-frame influence have one prominent peak occupying the same frequency bin of the original signal, i.e., frequency bin 62 in this example. Also, the magnitudes decrease rapidly away from this frequency bin. This pattern is true of all scenarios. An average pattern is obtained by averaging all patterns obtained from various room impulse response functions and time frame shifts. The average pattern is shown in Fig. 3(b), and it smoothes out the variations typically shown in Fig. 3(a).

The second aspect of the magnitude curve is that the peak magnitudes are more attenuated as frame shifts increase due to the decaying pattern of room impulse response function. An example of peak magnitude attenuation is shown in Fig. 4. The solid line represents the attenuation in frequency bin 62 with different frame shifts. The decaying pattern of room impulse response can be approximated by an exponential decay function, specified by reverberation time [8]. The theoretical attenuation curve based on the reverberation time is shown as the dash line in Fig. 4. Although some variations exist, the theoretical attenuation curve approximates the real attenuation curve.

Knowing the reverberation time, the magnitude of the inter-frame influence can be estimated from the theoretical attenuation curve. More specifically, we obtain:

$$|I(k,k_y;d)| = A(d)P(k_y-k) \tag{5}$$

where $A(d)$ is the theoretical attenuation curve shown as the dash line in Fig. 4 and $P(k_0)$ is the average pattern shown in Fig. 3(b).

The distinction of early and later echoes for speech is defined as a delay of 50 ms in the room impulse response function [8]. This translates to approximately 7 time frame shifts for an inter-frame distance of 8 ms. We estimate the effects of later echo components using Equation 4. Since the phases are unknown, a frame-by-frame iterative spectral subtraction based method is employed for speech enhancement. We derive:

$$\left|S_{\bar{y}}(k_y;i)\right| = \sqrt{u\left[\left|S_y(k_y;i)\right|^2 - \alpha\sum_{d=7}^{D}\sum_{k=0}^{N-1}\left|I(k,k_y;d)\right|^2\left|S_{\bar{y}}(k;i-d)\right|^2\right]} \tag{6}$$

where $S_{\bar{y}}(k;i)$ is the short-term spectrum of enhanced speech, and $u(x)=x$ if $x\geq 0$, and $u(x)=0$ otherwise. Parameter $\alpha=0.08$ is associated with spectral subtraction method. The short-term phase spectrum of enhanced speech is set to that of reverberant speech. Finally, the processed signal is reconstructed from $S_{\bar{y}}(k;i)$.

3. RESULTS AND DISCUSSIONS

Our algorithm has been evaluated with different utterances and reverberations. 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. 6. Fig. 6(a) and (b) show the clean signal and its spectrogram, respectively. The reverberant signal is produced by convolving the clean signal and a room impulse response function with a 0.3 s reverberation time. Fig. 6(c) and (d) show the reverberant signal and its spectrogram, respectively. Finally, the processed speech and its spectrogram are shown in Fig. 6(e) and (f). The figure, as well as a listening test, shows that the effects of reverberation are appreciably reduced. For example, the tail blurs in reverberant speech filling the silence gaps between energy bursts in clean speech are significantly suppressed. In some cases, they are entirely removed. From visual inspection, our results are comparable to those of Yegnanarayana and Murthy [12]. We have also tested our system on reverberant signals corrupted by white noise. Similar improvements are obtained.

Our pitch-based reverberation measure exploits a well-established notion – pitch – in psychoacoustics, and can potentially be applied to reverberant signals of multiple speech sources. Moreover, the measure may be employed as a criterion for optimization-based dereverberation methods. This paper represents a first step and further performance improvements are expected in future research.

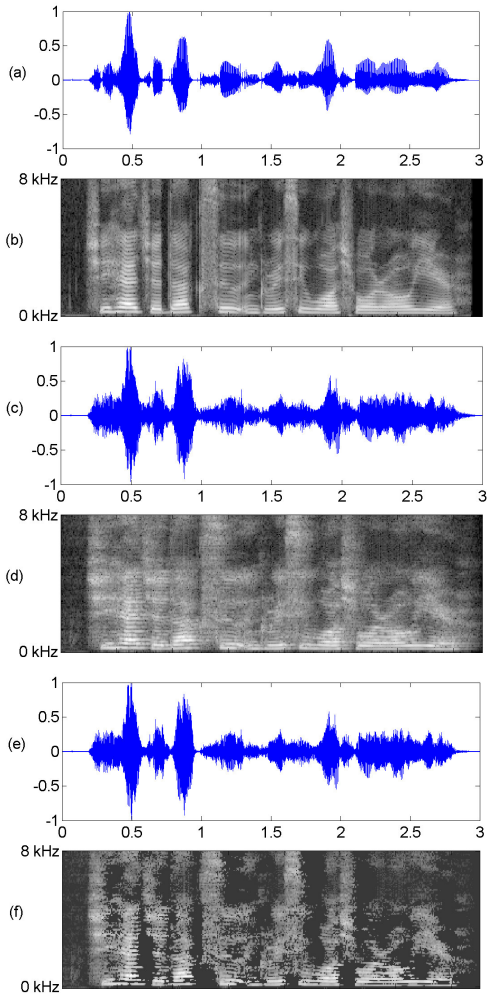


Fig. 5. Results of enhancement of reverberant speech: (a) clean speech, (b) spectrogram of clean speech, (c) reverberant speech, (d) spectrogram of reverberant speech, (e) speech processed using the proposed algorithm, and (f) spectrogram of the processed speech.

Acknowledgments This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-1-0027).

REFERENCES

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

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

[3] M.S. Brandstein, "On the use of explicit speech modeling in microphone array applications," *Proc. IEEE ICASSP*, 1998, pp. 3613-3616.

[4] M.S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.,* vol. 105, no. 5, pp. 2914-2919, 1999.

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

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

[7] X. Huang, A. Acero, and H.W. Hon, *Spoken Language Processing: a Guide to Theory, Algorithm, and system development*, Upper Saddle River, NJ: Prentice Hall, 2001.

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

[9] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rrice, *APU Report 2341: An Efficient Auditory Filterbank Based on the Gammatone Function*, Cambridge: Applied Psychology Unit, 1988.

[10] D.L. Wang and G.J. Brown, "Seperation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, no.3, pp.684-697, 1999.

[11] M. Wu, D.L. Wang, and G.J. Brown, "A multi-pitch tracking algorithm for noisy speech," to appear in *IEEE Trans. Speech Audio Processing*. An earlier version also appears in "A multi-pitch tracking algorithm for noisy speech," *Proc. IEEE ICASSP*, 2002, pp. 369-372.

[12] B. Yegnanarayana and P.S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech Audio Processing*, vol. 8, no.3, pp. 267-281, 2000.