Harmonicity Based Blind Dereverberation with Time Warping

Tomohiro Nakatani, Keisuke Kinoshita, Masato Miyoshi, Parham S. Zolfaghari

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
{nak,kinoshita,miyo,zparham}@cslab.kecl.ntt.co.jp

ABSTRACT

Speech dereverberation is desirable in applications such as robust automatic speech recognition (ASR) in the real world. Although a number of dereverberation methods have been exploited, dereverberation is still a challenging problem especially when using a single microphone. To overcome this problem, a harmonicity based dereverberation method (HERB) has recently been proposed. HERB can blindly estimate the inverse filter of a room impulse response based on harmonicity of speech signals and dereverberate the signals. However, HERB uses an imprecise assumption that hinders the dereverberation performance, that is, the fundamental frequency (F_0) of a speech signal is assumed to be constant within a short time frame when extracting the features of harmonic components. In this paper, we introduce time warping analysis into HERB to remove this bottleneck. Time warping analysis expands and contracts the time axis of a signal in order to make the F_0 of the signal constant, and makes it possible to estimate harmonic components precisely even when their frequencies change rapidly. We show that time warping analysis can effectively improve the dereverberation effect of HERB when the reverberation time is longer than 0.1 sec.

1. INTRODUCTION

Harmonicity has long been studied as a robust feature of speech signals in the real world. It is cited as a major clue in relation to a person's ability to extract a desired speech from other sounds [4]. Many speech enhancement methods employ a harmonicity-based sound segregation scheme, and have improved the performance of automatic speech recognition (ASR) [10, 16]. However, these methods have not succeeded in extracting the precise harmonic structure of speech signals in the presence of long reverberation. This is because different fundamental frequencies (F_0) in different time regions are mixed into the reverberation, and thus the harmonic structure is severely degraded. Therefore, harmonicity has not been taken into account as a primary cue for enhancing or dereverberating reverberant speech signals.

Long reverberation, on the other hand, has a severe detrimental effect on ASR. Although several adaptation techniques, such as cepstral mean normalization (CMN) [3] and maximum likelihood linear regression (MLLR) [8], have been proposed for recognizing reverberant speech signals, they can only deal with short reverberation. It is reported that the recognition performance cannot be improved sufficiently when the reverberation time is longer than 0.5 sec even if the acoustic models are used that are trained with a matched reverberation condition [7]. Therefore, the dereverberation of speech signals is essential for ASR in a reverberant environment.

Several blind dereverberation techniques have been exploited that use microphone array systems. A typical technique involves estimating the directions of arrival (DOAs) of a direct speech signal, and enhancing signal components coming from that direction. The delay-and-sum beamformer is often used for this purpose [5]. However, it requires a large number of microphones to achieve a large dereverberation gain. By contrast, another technique based on inverse filtering can suppress reverberation using a small number of microphones. Theoretically, the reverberation is completely eliminated by arranging microphones so that the transfer functions from N signal sources to $N + 1$ microphones have no common zeros [9]. Several blind techniques for estimating the inverse filter have also been proposed based on the assumption that a source signal is a statistically independent and identically distributed (i.i.d.) sequence [2, 6]. These methods can precisely dereverberate an observed signal if the source signal is actually an i.i.d. sequence. However, they cannot appropriately deal with speech signals because speech signals have inherent properties, such as harmonicity and formant structure, making their sequences statistically dependent. This approach inevitably destroys such essential properties of speech signals.

To overcome these problems, a new dereverberation principle has recently been proposed based on the harmonicity of speech signals, and a single channel blind dereverberation method, known as *Harmonicity based dEReverBeration* (*HERB*) was presented [12, 14]. According to this principle, a filter that enhances the harmonic structure of observed reverberant signals approximates the inverse filter. HERB estimates this filter by calculating the average transfer function that transforms observed signals into their direct harmonic components estimated by an adaptive harmonic filter. The experiments showed that HERB can effectively dereverberate speech signals when sufficiently long observed signals are given. However, it is found that HERB does not have as much dereverberation effect for male speech signals as for female ones. In addition, HERB cannot appropriately deal with higher frequency components when their frequencies change rapidly with time.

In this paper, we present an extended version of HERB, referred to as *HERB with TimE wArping* (*HERB-TEA*), to improve the preciseness of the dereverberation. We believe the above problems with HERB to be caused by its imprecise treatment of harmonic components, that is, F_0 is dealt with as a constant within a short time frame. To overcome this problem, we introduce time warping analysis that expands and contracts the time axis of the signals to make their F_0s approximately constant. Time warping analysis allows us to extract the precise features of harmonic components [1]. Our experiments show that this extension successfully improves the performance of HERB, especially for male speech signals.

In section 2, we describe HERB and its problems. Time warping analysis is incorporated in HERB in section 3. Experiments and concluding remarks are presented in sections 4 and 5, respectively.

2. HERB – HARMONICITY BASED DEREVERBERATION METHOD

In this section, we first briefly describe the property of the inverse filter, referred to as the dereverberation filter, estimated by HERB,

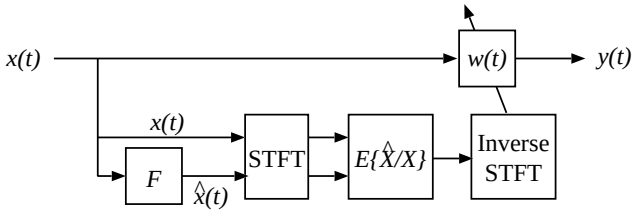


Figure 1: Diagram of HERB: each observed signal, $x(t)$, is first filtered by an adaptive harmonic filter, F, to obtain $\hat{x}(t)$. The average ratio of $\hat{x}(t)$ to $x(t)$ in the frequency domain over different observations is calculated to determine the dereverberation filter, $w(t)$. Finally, $x(t)$ is dereverberated by being convolved with $w(t)$.

and then discuss the problems with HERB.

2.1. Dereverberation Filter Estimated by HERB

Figure 1 shows a diagram of HERB. HERB estimates the dereverberation filter as an average filter that transforms observed reverberant signals into the output of an adaptive harmonic filter roughly estimating the direct harmonic components in the observed signals. The dereverberation filter, $W(\omega)$, is calculated as follows[1]:

$$W(\omega) = E\left\{\frac{\hat{X}(\omega)}{X(\omega)}\right\}, \tag{1}$$

where X and $\hat{X}$ are the observed reverberant signal and the output of an adaptive harmonic filter, respectively. $E\{\cdot\}$ is an average function that calculates the average value of $\hat{X}/X$ for different observed signals from a sound source.

This filter has been proven to approximate the inverse filter of a room transfer function for speech signals[2]. Here, we briefly interpret the property of this filter after introducing a speech signal model.

2.1.1. Speech Signal Model

A speech signal, $S(\omega)$, can be modeled by the sum of the harmonic components, $S_h(\omega)$, derived from the glottal vibration, and non-harmonic components, $S_n(\omega)$, such as fricatives and plosives as eq. (2).

$$S(\omega) = S_h(\omega) + S_n(\omega). \tag{2}$$

The observed reverberant signal, $X(\omega)$, is then obtained by multiplying a room transfer function $H(\omega)$ by $S(\omega)$ as eq. (3). This transfer function can also be divided into two functions, D and R. The former transforms S into the direct signal, DS, and the latter into the reverberation part, RS, as shown in eq. (4).

$$X(\omega) = H(\omega)S(\omega), \tag{3}$$

$$X(\omega) = D(\omega)S(\omega) + R(\omega)S(\omega). \tag{4}$$

Then, the observed signal, X, is also represented as eq. (5) using eqs. (2) and (4).

$$X(\omega) = D(\omega)S_h(\omega) + (R(\omega)S_h(\omega) + H(\omega)S_n(\omega)). \tag{5}$$

The first term on the right side of eq. (5), DS_h, is the direct signal of the harmonic components, and is as highly periodic as the harmonic components in the source signal. By contrast, RS_h in the second term on the right side is the reverberation part of the harmonic components, and thus has degraded harmonicity. HS_n is not harmonic because S_n is originally a non-harmonic part. Therefore, the second term on the right side represents the non-harmonic parts in the observed signal.

Of these components, DS_h can approximately be extracted from X by an adaptive harmonic filter. This approximated direct signal $\hat{X}(\omega)$ can be modeled as follows:

$$\hat{X}(\omega) = D(\omega)S_h(\omega) + (\hat{R}_h(\omega) + \hat{H}_n(\omega)), \tag{6}$$

where $\hat{R}_h(\omega)$ and $\hat{H}_n(\omega)$, respectively, are part of the reverberation of S_h and part of the direct signal and reverberation of S_n, which unexpectedly remain in $\hat{X}$ after the harmonic filtering. We assume that all the estimation errors in $\hat{X}$ are caused by $\hat{R}_h$ and $\hat{H}_n$ in eq. (6).

2.1.2. Interpretation of Dereverberation Filter

By substituting X and $\hat{X}$ in eq.(1) with eqs. (3) and (6), we can derive the following equation [14]:

$$W(\omega) \simeq \frac{D(\omega) + \hat{R}(\omega)}{H(\omega)} P\{|S_h(\omega)| > |S_n(\omega)|\}, \tag{7}$$

where

$$\hat{R}(\omega) = E\left\{\frac{\hat{R}_h(\omega)}{S_h(\omega)}\right\}_{|S_h(\omega)|>|S_n(\omega)|}, \tag{8}$$

where $P\{\cdot\}$ is a probability function, and $E\{\cdot\}_A$ represents an average function under a condition where A holds.

Equation (7) means that W approximately coincides with the product of $(D+\hat{R})/H$ and $P\{|S_h| > |S_n|\}$. The former, $(D+\hat{R})/H$, strictly equals the inverse filter, D/H, when an adaptive harmonic filter can completely reduce $\hat{R}_h$ in eq. (6) without any errors. Although it is very difficult to reduce $\hat{R}_h$ completely, a major part of RS_h can be eliminated with an adaptive harmonic filter. In addition, $\hat{R}$ is defined as an average filter that transforms S_h to $\hat{R}_h$. Therefore, $\hat{R}$ is expected to become a transformation that produces reduced reverberation. As a consequence, the signal obtained by multiplying the observed signal X by $(D+\hat{R})/H$ is expected to be the sum of the direct signal and the reduced reverberation, that is, $((D+\hat{R})/H)X = DS + \hat{R}S$. By contrast, $P\{|S_h| > |S_n|\}$ in eq. (7) is the probability that the harmonic component has a larger energy than the non-harmonic component, and has a real value between 0.0 and 1.0. This term changes the gain of eq. (1) but does not affect its dereverberation function.

2.2. Adaptive Harmonic Filter

As discussed above, the precise extraction of direct harmonic components with the adaptive harmonic filter is very important for HERB. For this purpose, HERB uses a harmonic filter based on a sinusoidal representation. Using this filter, F_0 of the observed signal at each time frame is first estimated from an observed signal, $x(t)$. Then, the amplitudes and phases of individual harmonic components are extracted from a short time Fourier transformation (STFT) of $x(t)$ as follows:

$$X_l(\omega) = \sum_n g_1(t_n - t_l)x(t_n)e^{-j\omega(t_n - t_l)}, \tag{9}$$

$$A_{k,l} = |X_l(r\{k\dot{\theta}_{t_l}\})|, \tag{10}$$

$$p_{k,l} = \angle X_l(r\{k\dot{\theta}_{t_l}\}), \tag{11}$$

[1] In this paper, time and frequency domain signals are represented by lower and upper case symbols, respectively. Arguments "(ω)" that represent the center frequencies of the short time Fourier transformation (STFT) bins are often omitted from frequency domain signals.

[2] A physical interpretation of this filter is given in [12]

where n and t_n are respectively the index of a waveform sample and its time, $A_{k,l}$ and $p_{k,l}$ are respectively the amplitude and phase of the k-th harmonic component at a time frame whose center time is t_l, $\dot{\theta}_{t_l}$ is F_0 of the frame, $g_1(t)$ is a window function, and $r\{\cdot\}$ is a function that quantizes a continuous frequency into a discrete center frequency of the nearest STFT bin. Finally, the output of the filter, $\hat{x}(t)$, is synthesized by adding sinusoids as eq. (12) and by combining them over succeeding frames based on the overlap-add synthesis as eq. (13).

$$\hat{x}_l(t_n) = \sum_k A_{k,l} \cos(k\dot{\theta}_{t_l}(t_n - t_l) + p_{k,l}), \quad (12)$$

$$\hat{x}(t_n) = \sum_m g_2(t_n - t_{l+m\triangle l})\hat{x}_{l+m\triangle l}(t_n), \quad (13)$$

where $\triangle l$ is a frame shift in samples and $g_2(t)$ is a window function.

2.3. Problems

There are, however, the following problems involved with the adaptive harmonic filter used in HERB.

- Features of harmonic components are extracted by assuming that the F_0 of speech signals is constant within a short time frame using eqs. (9), (10) and (11), although F_0 generally changes even in a local time region. This causes estimation errors in a direct signal, $\hat{x}(t)$, and thus degrades the dereverberation filter estimation. Because harmonic frequencies in higher frequency regions change more rapidly than those in lower frequency regions, the estimation errors increase with frequency.
- When the F_0 of a speech signal is small, it is difficult to distinguish its direct signal from its reverberation part using an adaptive harmonic filter. This is because the differences between adjacent harmonic frequencies at a frame are small in such cases and relatively large parts of the reverberation overlap the direct signal.

As a consequence, the dereverberation performance of HERB is consistently worse for male speech signals than for female speech signals. In certain experiments using male speech signals, HERB even increased the energies of the reverberations compared with those of room impulse responses in time regions long after the direct signals had arrived.

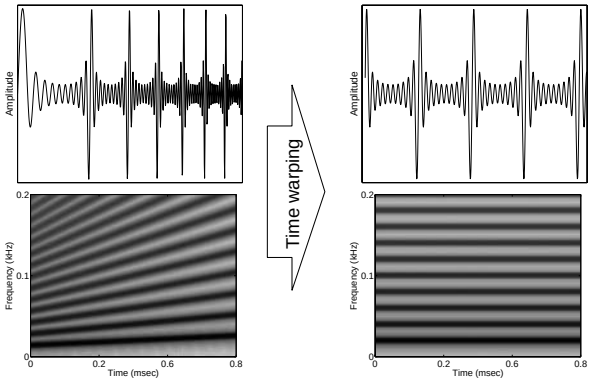
3. HERB WITH TIME WARPING

To improve the dereverberation performance of HERB, we extended it by introducing time warping analysis into its adaptive harmonic filtering. This extended method is referred to as *HERB with TimE wArping* (*HERB-TEA*) in this paper.

3.1. Adaptive Harmonic Filter with Time Warping

Figure 2 illustrates the idea of time warping and Fig. 3 shows the flow of adaptive harmonic filtering when time warping analysis is employed. The time warping analysis first uses a time-warping function that expands and contracts the time axis of a signal in the original time domain to obtain a signal with an approximately constant F_0 in a warped time domain. The amplitudes and phases of the sinusoidal components are extracted from the signals in the warped time domain. A harmonicity enhanced signal is then synthesized in the original time domain using the extracted features and the time warping function.

Let $\tau = \mathcal{W}_l(t)$ be the time-warping function that transforms $x(t)$ within a short time frame, whose center time is t_l, in the original time domain into $x_{\mathcal{W}_l}(\tau)$ in the warped time domain, then the



Original time domain Warped time domain

Figure 2: Waveforms (upper panels) and spectrograms (lower panels) of a signal before and after time warping. In this example, the fundamental frequency of the signal increases with time in the original time domain while it is constant in the warped time domain.

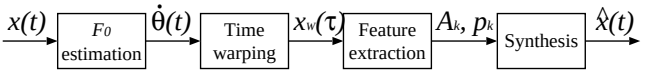


Figure 3: Processing flow of adaptive harmonic filtering with time warping

relation between $x(t)$ and $x_{\mathcal{W}_l}(\tau)$ is represented as follows:

$$x_{\mathcal{W}_l}(\mathcal{W}_l(t)) = x(t) \ \text{ for } \ |t - t_l| < \frac{T}{2}. \quad (14)$$

where T is the length of the frame. In particular, let $\theta(t)$ be the phase of the F_0 component of $x(t)$, and $\phi(\tau)$ be that of $x_{\mathcal{W}_l}(\tau)$, then the relation between $\theta(t)$ and $\phi(\tau)$ is represented as eq. (15).

$$\phi(\mathcal{W}_l(t)) = \theta(t) \ \text{ for } \ |t - t_l| < \frac{T}{2}. \quad (15)$$

In time warping analysis, we determine $\mathcal{W}_l(t)$ so that it makes $\dot{\phi}(\tau)$ constant within a time frame as eq. (16).

$$\frac{d\phi(\tau)}{d\tau} = \dot{\phi}_{\tau_l} \ \text{ for } \ |\mathcal{W}_l{}^{-1}(\tau) - t_l| < \frac{T}{2}, \quad (16)$$

where $\tau_l = \mathcal{W}_l(t_l)$. τ_l and $\dot{\phi}_{\tau_l}$ are parameters that can be set to an arbitrary number[3]. In addition, to simplify the calculation, we assume that the time derivative of F_0 is constant within a short time frame in the original time domain, that is:

$$\frac{d^2\theta(t)}{dt^2} = \ddot{\theta}_{t_l} \ \text{ for } \ |t - t_l| < \frac{T}{2}, \quad (17)$$

where $\ddot{\theta}_{t_l}$ is the derivative of F_0 at time t_l. Then, the time warping function, $\mathcal{W}_l(t)$, that satisfies eqs. (15), (16) and (17) is derived as follows:

$$\mathcal{W}_l(t) = (t - t_l)^2 \frac{\ddot{\theta}_{t_l}}{2\dot{\phi}_{\tau_l}} + (t - t_l)\frac{\dot{\theta}_{t_l}}{\dot{\phi}_{\tau_l}} + \tau_l, \quad (18)$$

[3]For example, $\tau_l = 0$ and $\dot{\phi}_{\tau_l} = \dot{\theta}_{t_l}$ are reasonable parameter settings.

$$\mathcal{W}_l^{-1}(\tau) = \begin{cases} \dfrac{(\dot{\theta}_{t_l}^2 + 2(\tau - \tau_l)\dot{\phi}_{\tau_l}\ddot{\theta}_{t_l})^{\frac{1}{2}} - \dot{\theta}_{t_l}}{\ddot{\theta}_{t_l}} + t_l, \\ \qquad\qquad\qquad\qquad \text{for } \ddot{\theta}_{t_l} \neq 0, \\ (\tau - \tau_l)\dfrac{\dot{\phi}_{\tau_l}}{\dot{\theta}_{t_l}} + t_l, \quad \text{for } \ddot{\theta}_{t_l} = 0. \end{cases} \tag{19}$$

The signal, $x_{\mathcal{W}_l}(\tau)$, in the warped time domain can then be obtained from $x(t)$ as follows:

$$x_{\mathcal{W}_l}(\tau) = x(\mathcal{W}_l^{-1}(\tau)). \tag{20}$$

The F_0 of this signal is expected to be constant because of the assumption of eq. (16), and thus, it is appropriate to model the signal with a sinusoidal representation. Let $X_{\mathcal{W}_l}(\omega)$ be the STFT of $x_{\mathcal{W}_l}(\tau)$, then the amplitude $A_{k,l}$ and phase $p_{k,l}$ of the k-th harmonic component in the warped time domain are extracted as follows:

$$X_{\mathcal{W}_l}(\omega) = \sum_n g_1(\tau_n - \tau_l) x_{\mathcal{W}_l}(\tau_n) e^{-j\omega(\tau_n - \tau_l)}, \tag{21}$$

$$A_{k,l} = |X_{\mathcal{W}_l}(r\{k\dot{\phi}_{\tau_l}\})|, \tag{22}$$

$$p_{k,l} = \angle X_{\mathcal{W}_l}(r\{k\dot{\phi}_{\tau_l}\}). \tag{23}$$

Then, the output of the harmonic filter in the original time domain at this frame can be synthesized as follows:

$$\hat{x}_l(t_n) = \sum_k A_{k,l} \cos(r\{k\dot{\phi}_{\tau_l}\}(\mathcal{W}_l(t_n) - \tau_l) + p_{k,l}), \tag{24}$$

Finally, the overlap-add synthesis is used in a similar way to eq. (13) in order to combine signals over succeeding frames.

In our implementation, we calculate the derivative of F_0, or $\ddot{\theta}_{t_l}$ in eqs. (18) and (19), by approximating each local time trajectory of F_0 with a quadratic function and by extracting the derivative of the function. The value can easily be calculated as follows:

$$\ddot{\theta}_{t_l} = \frac{\sum_{m=-p}^{p} m\dot{\theta}_{t_{l+m\,\triangle l}}}{(\triangle l / f_s)\sum_{m=-p}^{p} m^2}, \tag{25}$$

where $\triangle l$ is a frame shift, f_s is a sampling frequency, and p specifies the local time region taken into account for this approximation.

3.2. Processing Flow of HERB-TEA

HERB-TEA is implemented using the same processing flow as HERB [13], except that it uses time warping analysis with its adaptive harmonic filtering.

One of the most important issues when implementing HERB is to estimate precise F_0 values in the presence of long reverberations. This is because they directly affect the performance of the adaptive harmonic filtering. For this purpose, HERB adopts a robust recently proposed F_0 estimator [11], and introduces a complementary scheme for estimating the F_0 values and the dereverberation filter [13]. In this scheme, 1) F_0 values are first estimated directly from reverberant signals, and the dereverberation filter is calculated based on the F_0 values, 2) then, the F_0 values are calculated again but more precisely using signals dereverberated by the dereverberation filter, and the dereverberation filter is also estimated more precisely using these F_0 values, and 3) finally, the F_0 values and the dereverberation filter are gradually refined through the iteration of this complementary estimation.

The implementation is described in detail in [13].

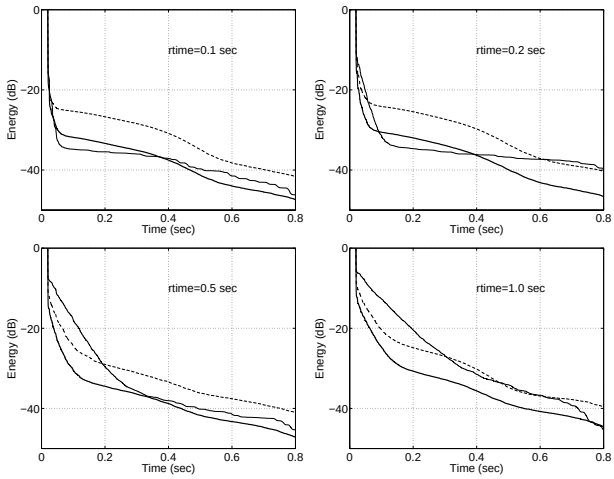


Figure 4: Energy decay curves of the room impulse responses (thin solid line) and dereverberated impulse responses (HERB: thin dashed line, HERB-TEA: thick line) for different reverberation times (rtime) when using male speech signals as training data.

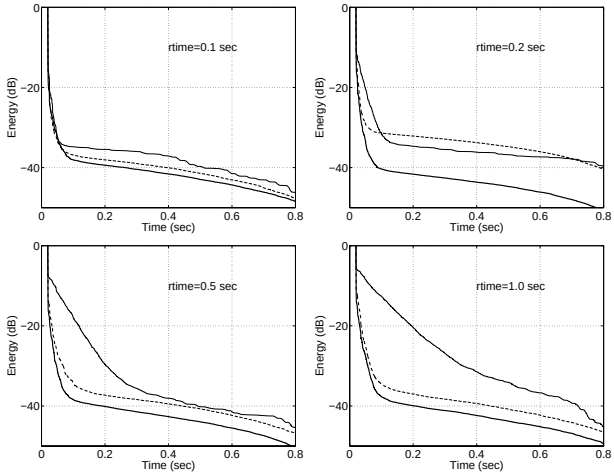


Figure 5: Energy decay curves of the room impulse responses (thin solid line) and dereverberated impulse responses (HERB: thin dashed line, HERB-TEA: thick line) for different reverberation times (rtime) when using female speech signals as training data.

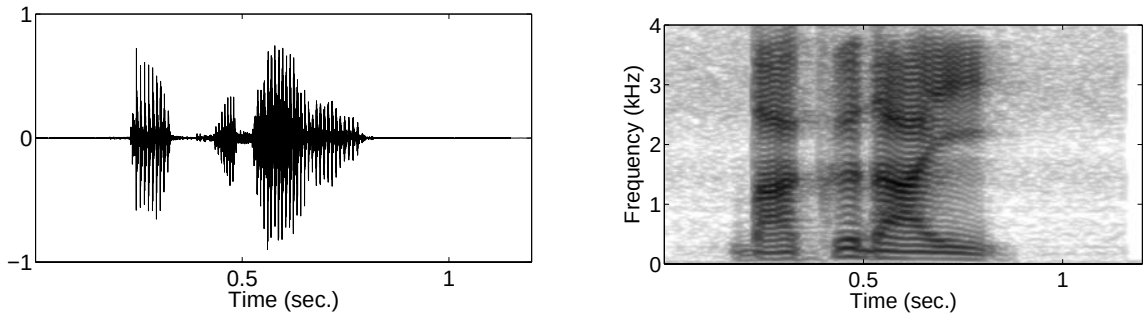
4. EXPERIMENTS

We evaluated the performance of HERB-TEA using the dereverberation task described in section 4.1 in terms of the energy decay curves of the impulse responses and ASR.

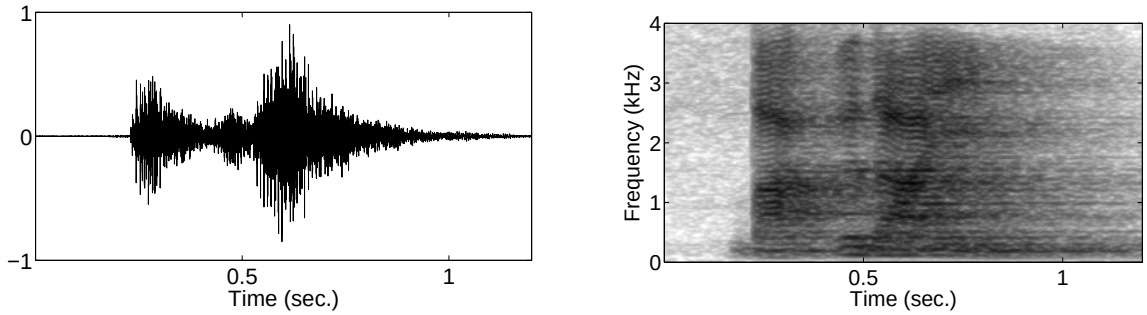
4.1. Task: Dereverberation of Word Utterances

The task used in our experiments was the dereverberation of reverberant word utterances. We used 5240 Japanese word utterances provided by a male and a female speaker (MAU and FKM) included in the ATR database as source signals, $s(t)$. We used four impulse responses measured in a reverberant room whose reverberation times were about 0.1, 0.2, 0.5, and 1.0 sec. Reverberant signals, $x(t)$, were obtained by convolving $s(t)$ with the impulse responses. Each dereverberation filter was estimated using all male word utterances or all female word utterances.

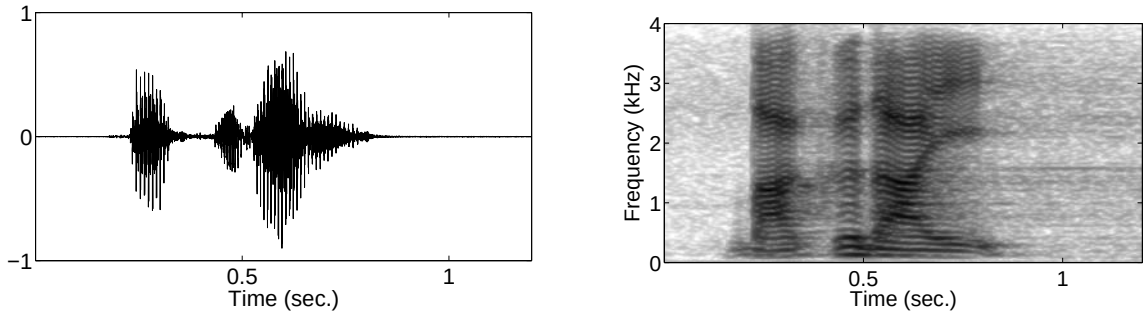
In the experiments, we assumed that each word utterance



(a) Source signal



(b) Observed reverberant signal



(c) Dereverberated signal

Figure 6: Waveforms (left panels) and spectrograms (right panels) of (a) source signal, (b) observed signal, and (c) dereverberated signal, for the utterance "Ba-Ku-Dai." (Reverberation time: 1.0 sec)

with reverberation was recorded separately, and that there is no time-overlap between utterances including reverberation durations. When we estimated the dereverberation filter using eq. (1), we calculated the STFT of $x(t)$ and $\hat{x}(t)$ with a short time frame whose length was long enough to contain each whole word utterance with zero padding. The length of the dereverberation filter was 131,072 taps; that is, we used a 10.9 sec rectangle window for the X and $\hat{X}$ calculations. By contrast, we used a much shorter time frame, that is, a 42 msec hanning window and 1 msec window shift for the F_0 estimation and adaptive harmonic filtering in order to extract the time-varying features of the harmonic components. We used signals sampled at 12 kHz.

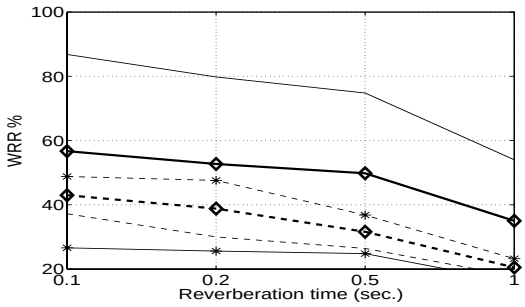
4.2. Energy Decay Curves of Impulse Responses

Figures 4 and 5 show energy decay curves of room impulse responses and dereverberated impulse responses obtained by HERB and HERB-TEA while controlling the reverberation time. Each dereverberated impulse response was obtained by convolving a room impulse response with its dereverberation filter, and each decay curve was calculated using Schroeder's method [15].

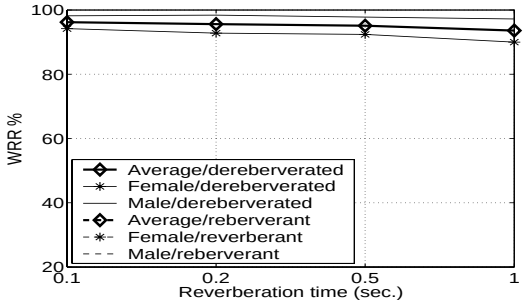
These figures show that HERB-TEA could effectively reduce the reverberation energy when the reverberation time was longer than 0.1 sec. HERB-TEA reduced the energy more successfully than HERB in all cases. This improvement was especially clear with male speech signals, that is, the energies of the dereverberated impulse responses in higher time regions were, in certain cases, increased compared with energies of the room impulse responses when using HERB, while they were effectively reduced when using HERB-TEA. In addition, HERB-TEA also reduced the energy just after the direct signal more successfully than HERB in most cases. Because this part of the reverberation energy has the largest effect on speech intelligibility [17], HERB-TEA is expected to improve it[4].

Figure 6 shows waveforms and spectrograms of a source signal, an observed reverberant signal, and a signal dereverberated by HERB-TEA. The source signal is a Japanese word "Ba-Ku-Dai" uttered by a male speaker. The reverberation time is 1.0 sec. It shows that HERB-TEA could effectively restore the time and frequency

[4]The effectiveness of HERB-TEA can clearly be confirmed by listening to the dereverberated signals included in the proceedings CD.



(a) WRRs obtained with clean model



(b) WRRs obtained with HERB model

Figure 7: Word recognition rates (WRRs) of reverberant and dereverberated signals when using (a) a clean model and (b) a HERB model under different reverberation time conditions.

structure of the source signal.

4.3. Speaker Dependent Word Recognition Rate

We evaluated the speaker dependent word recognition rate (WRR) of speech signals dereverberated by HERB-TEA. For this purpose, we prepared two types of acoustic monophone model. One was a model trained on source signals, referred to as a clean model, and the other was a model trained on signals obtained by applying HERB-TEA to the source signals, referred to as a HERB model. We used the clean model to recognize both reverberant and dereverberated signals, and the HERB model to recognize dereverberated signals. 4740 words randomly selected from 5240 words were used as training data, and the remaining 500 words were used as test data. 12-th order MFCCs, 12-th order delta MFCCs, three state HMMs, five mixture Gaussian distributions, 25 msec frame length, and 5 msec frame shift were adopted as the analysis conditions.

Figure 7 (a) shows the WRRs we obtained using the clean model. The average WRRs of dereverberated male and female speech signals were improved compared with those of reverberant signals. However, the average WRRs of the dereverberated signals were at most 60%. This result means that HERB-TEA cannot restore the precise spectral shapes of the original source signals although it can greatly reduce the reverberation. We consider this result to be caused by the limitation of HERB-TEA, that is, the probability function $p\{\cdot\}$ in eq. (7) modifies the spectral shapes of the dereverberated signals. By contrast, Fig. 7 (b) shows the WRRs obtained using the HERB model. The WRRs of the dereverberated signals were more than 90% under all reverberation conditions although only clean source signals were used for the acoustic model training. This means that speech signals dereverberated by HERB-TEA have similar spectral shapes independent of the reverberation time. In other words, HERB-TEA can successfully reduce the spectral variations in speech signals produced by reverberation without losing the speech features essential for ASR.

5. CONCLUSION

This paper proposed a method for improving the dereverberation effect of the harmonicity based dereverberation method (HERB) by introducing time warping analysis into its adaptive harmonic filtering. The time warping analysis allows us to extract features of harmonic components precisely even when their frequencies change within a short time frame. Experimental results showed that HERB with time warping (HERB-TEA) provided better dereverberation performance than HERB in terms of the energy decay curves of the impulse responses under various reverberation conditions. In addition, speaker dependent word recognition rates could be increased to more than 90% even under a 1.0 sec reverberation time condition when using the HERB model as the acoustic model. This means that HERB-TEA can effectively reduce spectral variations produced by different impulse responses without losing essential features for ASR. Future work will include an investigation of how such high quality speech dereverberation can be achieved with fewer speech data.

6. REFERENCES

[1] Abe, T., and Honda, M., "Sinusoidal modeling based on instantaneous frequency attractors," *Proc. ICASSP-2003*, vol. 6, pp. 133–136, 2003.

[2] Amari, S., Douglas, S.C., Cichocki, A., and Yang, H.H., "Multichannel blind deconvolution and equalization using the natural gradient," *Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications*, Paris, pp. 101–104, April 1997.

[3] Atal, B.S., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *JASA*, 55(6), pp. 1304-1312, 1974.

[4] Bregman, A. S., *Auditory scene analysis - the perceptual organization of sound*, MIT Press, 1990.

[5] Flanagan, J.L., "Computer-steered microphone arrays for sound transduction in large rooms," *JASA*, 78(11), pp. 1508–1518, 1985.

[6] Furuya, K., and Kaneda, Y., "Noise reduction and dereverberation using correlation matrix based on the multiple-input/output inverse-filtering theorem (MINT)," *Proc. International Workshop on Hands-Free Speech Communication*, pp. 59–62, 2001.

[7] Kingsbury, B., and Morgan, N., "Recognizing reverberant speech with rasta-plp," *Proc. ICASSP-97*, vol. 2, pp. 1259–1262, 1997.

[8] Leggetter, C.J., and Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[9] Miyoshi, M., and Kaneda, Y., "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, 36(2), pp. 145–152, 1988.

[10] Nakatani, T., *Computational auditory scene analysis based on residue-driven architecture and its application to mixed speech recognition*, Ph.D. thesis, Dept. of Applied Analysis & Complex Dynamical Systems, Kyoto Univ., Mar., 2002. (Online available: http://www.kecl.ntt.co.jp/icl/signal/nakatani/papers/story.pdf)

[11] Nakatani, T., and Irino, T., "Robust fundamental frequency estimation against background noise and spectral distortion," *Proc. ICSLP-2002*, vol. 3, pp. 1733–1736, Denver, Sep., 2002.

[12] Nakatani, T., and Miyoshi, M., "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. ICASSP-2003*, vol. 1, pp. 92–95, Apr., 2003.

[13] Nakatani, T., Miyoshi, M., and Kinoshita, K., "Implementation and effects of single channel dereverberation based on the harmonic structure of speech," *Proc. IWAENC-2003*, pp. 91–94, Sep., 2003.

[14] Nakatani, T., Miyoshi, M., and Kinoshita, K., "One microphone blind dereverberation based on quasi-periodicity of speech signals," Advances in Neural Information Processing Systems 16 (NIPS 16), MIT Press, 2004 (in press).

[15] Schroeder, M.R., "New method of measuring reverberation time," *JASA*, 37, pp. 409–412, 1965.

[16] Weintraub, M., "A computational model for separating two simultaneous talkers," *Proc. ICASSP-86*, vol. 11, pp. 81–84, 1986.

[17] Yegnanarayana, B., and Ramakrishna, B.S., "Intelligibility of speech under nonexponential decay conditions," *JASA*, vol. 58, pp. 853–857, Oct. 1975.