Acoust. Sci. & Tech. **25**, 4 (2004)

PAPER

An improved method based on the MTF concept for restoring the power envelope from a reverberant signal

Masashi Unoki*, Masakazu Furukawa†, Keigo Sakata‡ and Masato Akagi§

School of Information Science, Japan Advanced Institute of Science and Technology, 1–1 Asahidai, Tatsunokuchi, Nomi, Ishikawa, 923–1292 Japan

(*Received 27 June 2003, Accepted for publication 13 February 2004*)

Abstract: A basic method for restoring the power envelope from a reverberant signal was proposed by Hirobayashi *et al.* This method is based on the concept of the modulation transfer function (MTF) and does not require that the impulse response of an environment be measured. However this basic method has the following problems: (i) how to precisely extract the power envelope from the observed signal; (ii) how to determine the parameters of the impulse response of the room acoustics; and (iii) a lack of consideration as to whether the MTF concept can be applied to a more realistic signal. This paper improves this basic method with regard to these problems in order to extend this method as a first step towards the development for speech applications. We have carried out 1,500 simulations for restoring the power envelope from reverberant signals in which the power envelopes are three types of sinusoidal, harmonics, and band-limited noise and the carriers are white noise, to evaluate our improved method with regard to (i) and (ii). We then have carried out the same simulations in which the carriers are two types of carrier of white noise or harmonics with regard to (iii). Our results have shown that the improved method can adequately restore the power envelope from a reverberant signal and will be able to be applied for speech envelope restoration.

Keywords: Power envelope, Reverberation time, Inverse filtering, Modulation transfer function (MTF)

PACS number: 43.72.Ew [DOI: 10.1250/ast.25.232]

1. INTRODUCTION

Restoration of the original signal from a reverberant signal is an important issue concerning not only various kinds of speech signal processing such as speech-emphasis for transmission systems (speaker to microphone) and hearing aid systems, but also regarding preprocessing for speech recognition systems. The ultimate goal of our work is to construct a blind speech dereverberation method which can restore a speech signal from reverberant speech without using useful prior information such as the impulse response of the room acoustics, and which enables less loss in speech intelligibility due to reverberation.

There are several well known inverse filtering methods which can be used to dereverberate the original signal from a reverberant signal in room acoustics. For example, Neely and Allen proposed a method that used a single microphone to remove a minimum phase component from the room effect [1]. This method, however, can only be used for room acoustics with minimum phase characteristics. Miyoshi and Kaneda proposed another method that used a microphone array and constraining non-overlaps of zeros in all pairs of the impulse responses between the sources and the microphones [2]. Wang and Itakura proposed a method of acoustic inverse filtering through multi-microphone sub-band processing that selects the best invertible microphone in each sub-band and reconstructs the full-band signal by summing up the inverse filtered sub-band signals of the best microphones [3]. These methods can be applied to room acoustics with non-minimum phase characteristics. However, for all of these methods the impulse response of the room acoustics must be precisely measured to determine the inverse filtering before the dereverberation. Moreover, the impulse response temporally varies with various environmental factors (temperature, etc.), so the room acoustics have to be precisely measured each time these methods are used. This is a significant drawback with regard to the use of these methods for various speech applications.

On the other hand, temporal envelope inverse filtering

*e-mail: unoki@jaist.ac.jp
†Currently with Fujitsu Prime Software Technologies Limited
‡Currently with DENON, Ltd.
§e-mail: akagi@jaist.ac.jp

methods have been proposed to restore the envelope information of original speech from reverberant speech and improve speech intelligibility that is degraded by reverberation. Most of these methods are based on the modulation transfer function (MTF) concept [4–6] and use temporal envelope deconvolution through high-pass filtering to remove the effect of reverberation (such as the low-pass filtering). For example, Langhans and Strube proposed an enhancement method for speech signals corrupted by reverberation or noise where they appropriately filtered the envelope signals in critical frequency bands based on short-term Fourier transform (STFT) and linear prediction [7]. They used theoretically derived inverse MTF as high-pass filtering. Avendano and Hermansky proposed a data designed filterbank technique to treat reverberant speech [8]. This technique consisted of data-derived filtering of the power spectrum trajectories of speech based on the STFT. Both methods used the temporal power spectrum deconvolved through linear or nonlinear filtering and the FFT-OverLap-Adding (OLA) reverberant phase spectrum to resynthesize the dereverberated signal.

Mourjopoulos and Hammond proposed another method to enhance reverberant speech by using multi-band processing for the envelope deconvolution [9]. Hirobayashi *et al.* proposed the power envelope inverse filtering method [10]. Both methods are based on a single- or multi-channel filterbank rather than on the STFT, so they can directly deal with the temporal envelope fluctuation based on the MTF concept. These methods differ, though, in their signal definition with regard to the envelope (amplitude or power) and the carrier (sine-wave or white noise) based on the amplitude modulation (AM) representation.

These last two methods ([9,10]) represent attempts to restore the temporal envelope from reverberant speech while the first two methods ([7,8]) attempted to restore the modulation index related to the modulation frequency of the reverberant speech to suppress the degradation of speech intelligibility caused by reverberation. These methods can restore the temporal envelope information (fluctuation or modulation index) [7–10], and provide two benefits — restoration can be done without measuring the impulse response of the room acoustics, and restoration of the amplitude information related to important features of speech recognition systems can be done. Therefore, it will be a useful preprocessing method for such applications.

We think that this kind of temporal inverse filtering method can be developed as a blind dereverberation method. We also think that AM-representation in the filterbank is better than that of the STFT in order to deal with the temporal envelope and the carrier separately, based on the MTF concept. Thus, a basic method proposed by Hirobayashi *et al.* [10] will be used as a reasonable model in our work.

In this paper, as a first step towards the construction of a blind speech dereverberation model, we reconsider the power envelope inverse filtering method proposed by Hirobayashi *et al.* [10] and point out three problems: how to precisely extract the power envelope, how to determine the model parameters, and a lack of consideration as to whether the MTF concept can be applied to a more realistic signal since Hirobayashi *et al.* applied their basic method to a speech signal without considering speech characteristics. We then improve their method to enable general temporal power envelope restoration for speech applications.

This paper is organized as follows. In Sec. 2, the concept of the power envelope inverse filtering method based on the MTF is described and its problems are pointed out. Section 3 describes the improved method that we use of overcome these problems and evaluates the results from use of this method. Section 4 gives our conclusions.

2. POWER ENVELOPE INVERSE FILTERING METHOD

2.1. The MTF Concept

The MTF concept was proposed by Houtgast and Steeneken [4] to account for a relation between a transfer function of frequency in an enclosure in terms of the envelopes of input and output signals and characteristics of the enclosure such as reverberation. This concept was introduced as a measure in room acoustics for assessing the effect of the enclosure on speech intelligibility [4–6]. The complex modulation transfer function, $M(\omega)$, is defined as

$$M(\omega) = \frac{\int_0^\infty h(t)^2 \exp(j\omega t)dt}{\int_0^\infty h(t)^2 dt}, \tag{1}$$

where $h(t)$ is the impulse response of the room acoustics and ω is the radian frequency [11]. This equation means the complex Fourier transform of the squared impulse response is divided by its total energy. Here, let us consider the impulse response of a room acoustic:

$$h(t) = \exp\left(-\frac{6.9t}{T_R}\right)n(t), \tag{2}$$

where the response has an envelope of exponential decay and white noise $n(t)$. This is the well-known stochastic-approximated impulse response in the room acoustics [11]. The MTF, $m(\omega)$, can be obtained as

$$m(\omega) = |M(\omega)| = \left[1 + \left(\omega \frac{T_R}{13.8}\right)^2\right]^{-1/2}, \tag{3}$$

where T_R is the reverberant time; that is, the time required for the power of $h(t)$ to decay by 60 dB [4–6,11].

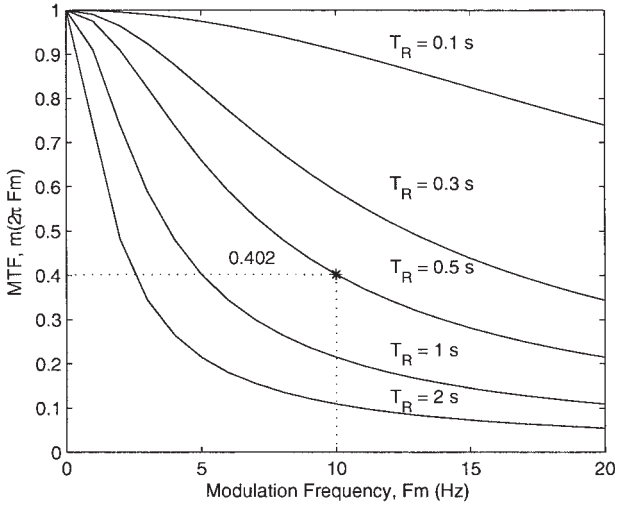

Fig. 1 Theoretical curves representing the modulation transfer function, $m(2\pi F_m)$, for various conditions with $T_R = 0.1$, 0.3, 0.5, 1.0, and 2.0 s.

Figure 1 shows the MTF, $m(\omega)$, as a function of the modulation frequency F_m (that is, the dominant frequency in the temporal envelope). These theoretical curves were calculated by substituting five reverberation times — $T_R = 0.1$, 0.3, 0.5, 1.0, and 2.0 s — and $\omega = 2\pi F_m$ into Eq. (3). Here, $m(\omega)$ can also be regarded as the modulation index (that is, a degree of relative fluctuation in the normalized amplitude) with respect to F_m. These curves show how much the modulation index of the envelope will be reduced from 1 to 0 depending on the reverberation time T_R at a specific F_m. In other words, T_R can be predicted from a specific $m(2\pi F_m)$ at a specific F_m. This is one advantage of using the MTF concept.

Based on the MTF concept, we can know how much reverberation affects a reduction of the modulation index, and then we can predict a reduced speech intelligibility using the MTF. Most of the temporal deconvolution approach is aimed at restoring the reduced MTF and then enhancing speech intelligibility using the restored MTF.

2.2. Model Concept Based on the MTF

In the model of Hirobayashi *et al.* [10], the observed reverberant signal, the original signal, and the stochastic-idealized impulse response in the room acoustics [5] are assumed to be $\boldsymbol{y}(t)$, $\boldsymbol{x}(t)$, and $\boldsymbol{h}(t)$, respectively, and these are modeled based on the MTF concept as follows:

$$\boldsymbol{y}(t) = \boldsymbol{x}(t) * \boldsymbol{h}(t), \tag{4}$$

$$\boldsymbol{x}(t) = e_x(t)\boldsymbol{n}_1(t), \tag{5}$$

$$\boldsymbol{h}(t) = e_h(t)\boldsymbol{n}_2(t), \tag{6}$$

$$e_h(t) = a\exp(-6.9t/T_R), \tag{7}$$

$$\langle \boldsymbol{n}_k(t)\boldsymbol{n}_k(t-\tau)\rangle = \delta(\tau), \tag{8}$$

where "$*$" denotes the operation of the convolution, $e_x(t)$ and $e_h(t)$ are the envelopes of $\boldsymbol{x}(t)$ and $\boldsymbol{h}(t)$, and $\boldsymbol{n}_1(t)$ and $\boldsymbol{n}_2(t)$ are the mutually independent respective white noise (random variables) functions. In this paper, note that for seek of convenience the random variables and observed variables are described using bold and plain characters (ex. $\boldsymbol{x}(t)$ and $x(t)$), respectively. The parameters of the impulse response, a and T_R, are a constant amplitude term and the reverberation time, respectively [10]. In this model, the reverberant signal $\boldsymbol{y}(t)$ is the convolution of $\boldsymbol{x}(t)$ with $\boldsymbol{h}(t)$ in the time domain, so the power envelope of the reverberant signal, $e_y(t)^2$, can be determined as

$$\begin{aligned}\langle \boldsymbol{y}(t)^2\rangle &= \left\langle \left\{ \int_{-\infty}^{\infty} \boldsymbol{x}(\tau)\boldsymbol{h}(t-\tau)d\tau \right\}^2 \right\rangle \\ &= \int_{-\infty}^{\infty} e_x(\tau)^2 e_h(t-\tau)^2 d\tau \\ &= e_y(t)^2, \end{aligned} \tag{9}$$

where $\langle\cdot\rangle$ is the ensemble average operation [12] (see Appendix for a detailed derivation of Eq. (9)).

Based on this result, $e_x(t)^2$ can be (restored) by deconvoluting $e_y(t)^2$ with $e_h(t)^2$. To cope with these signals in a computer simulation, these variables are transformed from a continuous signal to a discrete signal based on the sampling theorem, such as $e_x[n]^2$, $e_h[n]^2$, $e_y[n]^2$, $x[n]$, $h[n]$, and $y[n]$. Here, n is the sample number and f_s is the sampling frequency. In this paper, f_s is set to 20 kHz. The transfer functions of power envelopes $E_x(z)$, $E_h(z)$, and $E_y(z)$ are then assumed to be the z-transforms of $e_x[n]^2$, $e_h[n]^2$, and $e_y[n]^2$, respectively. Also, the transfer function of the power envelope of the impulse response, $E_h(z)$, can be represented as [10]:

$$E_h(z) = \frac{a^2}{1 - \exp\left(-\dfrac{13.8}{T_R \cdot f_s}\right) z^{-1}}. \tag{10}$$

Thus, the transfer function of the power envelope of the original signal, $E_x(z)$, can be determined from

$$\begin{aligned} E_x(z) &= \frac{E_y(z)}{E_h(z)} \\ &= \frac{E_y(z)}{a^2}\left\{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}\right\}. \end{aligned} \tag{11}$$

Finally, the power envelope $e_x[n]^2$ can be obtained from the inverse z-transform of $E_x(z)$ [10].

Figure 2 shows an example of how the power envelope inverse filtering method is related to the MTF concept. Figure 2(a) shows a sinusoidal power envelope as the original power envelope $e_x(t)^2$ ($= 0.5(1 + \sin(2\pi F_m t))$; the modulation frequency F_m was 10 Hz and the modulation index m was 1). Figure 2(b) shows the original signal $x(t)$ calculated from $e_x(t)^2$ and a white noise carrier $n_1(t)$ using Eq. (5). Figure 2(c) shows the power envelope $e_h(t)^2$

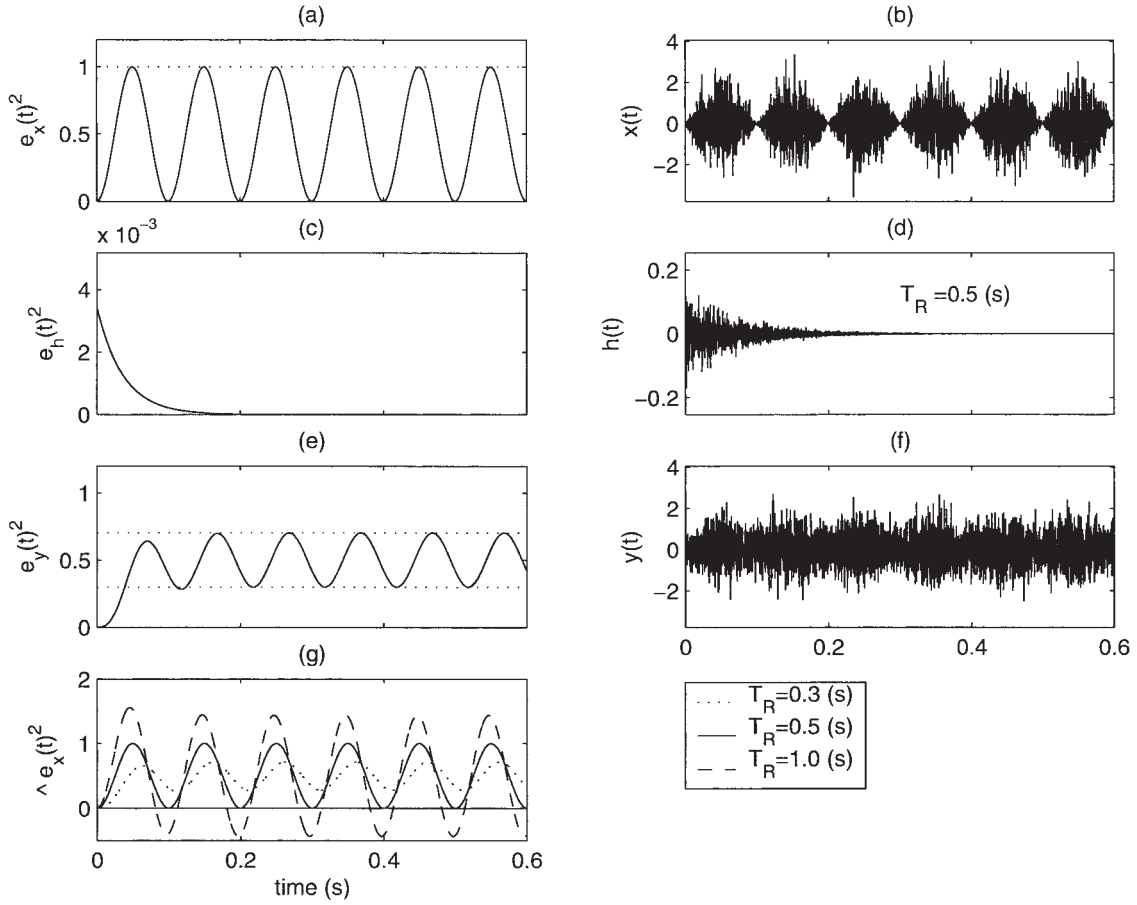

Fig. 2 Example of the relationship between the power envelopes of a system based on the MTF concept: (a) power envelope $e_x(t)^2$ of (b) original signal $x(t)$, (c) power envelope $e_h(t)^2$ of (d) impulse response $h(t)$, (e) power envelope $e_y(t)^2$ derived from $e_x(t)^2 * e_h(t)^2$, (f) reverberant signal $y(t)$ derived from $x(t) * h(t)$, and (g) restored power envelope $\hat{e}_x(t)^2$.

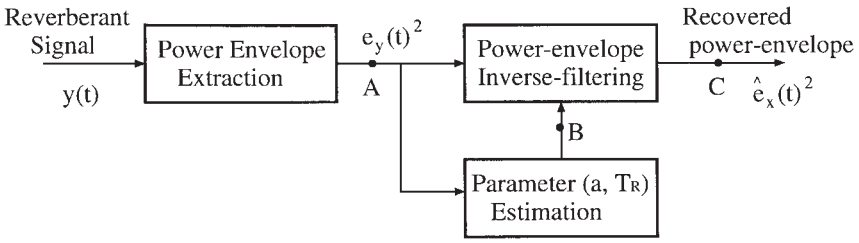

Fig. 3 Block diagram of the power envelope inverse filtering method.

calculated using Eq. (7) with $T_R = 0.5$ s. Figure 2(d) shows the impulse response $h(t)$ of Eq. (6), calculated from $e_h(t)^2$ and a white noise carrier $n_2(t)$. Figures 2(e) and (f) show the power envelope $e_y(t)^2$ obtained from a convolution of $e_x(t)^2$ with $e_h(t)^2$ and the observed reverberant signal $y(t)$ obtained from a convolution of $x(t)$ with $h(t)$, respectively. The left panels ((a), (c), and (e)) show the power envelopes of the signals and the right panels ((b), (d), and (f)) show the corresponding signals. In this figure, the modulation index decreased from 1.0 (in Fig. 2(a)) to 0.404 (maximum deviation of the envelope between the dotted lines in Fig. 2(e) relative to that in Fig. 2(a)). Since the MTF concept shows the modulation index as a function of F_m and T_R [6], it can also be shown that the decreased modulation index is derived from $m(2\pi F_m) = 0.402$ using Eq. (3) by substituting $T_R = 0.5$ s and $F_m = 10$ Hz into Eq. (3).

The solid line in Fig. 2(g) shows the restored power envelope $\hat{e}_x(t)^2$ obtained from the reverberant power envelope $e_y(t)^2$ (Fig. 2(e)) using Eq. (11) with $T_R = 0.5$ s. It is shown that the power envelope inverse filtering method can precisely restore the power envelope from a reverberant signal in terms of the shape and the magnitude.

2.3. Problems

Figure 3 shows a block diagram of the power envelope inverse filtering method. With this model concept, the basic

method can restore the power envelope of an original signal from an observed reverberant signal if it can completely extract the reverberant power envelope $e_y(t)^2$ from the observed reverberant signal $y(t)$ (at the first block shown in Fig. 3) and the parameters of the room impulse response (at the lower second block in Fig. 3), and T_R and a are known before the power envelope inverse filtering.

However, this basic method might not precisely restore the power envelope $e_x(t)^2$ from a reverberant signal $y(t)$ using Eq. (11), provided that the power envelope $e_y(t)^2$ was incompletely extracted from the observed $y(t)$. In addition, the restored envelope $\hat{e}_x(t)^2$ might not be improved when using incorrect parameters a and T_R in this model even if the power envelope $e_y(t)^2$ was completely extracted from the observed $y(t)$. For example, in contrast with the solid line in Fig. 2(g), if the method is applied with an inappropriate value (for example, $T_R = 0.3\,\mathrm{s}$ or $T_R = 1.0\,\mathrm{s}$), the restored power envelope $\hat{e}_x(t)^2$ will not be precisely restored as the other lines in Fig. 2(g) show. In this paper, the former and the latter cases are called "under-restoration" and "over-restoration," respectively. In particular, $\hat{e}_x(t)^2$ was excessively restored from $e_y(t)^2$ and had a negative-envelope when $T_R = 1.0\,\mathrm{s}$ (dot-dashed line in Fig. 2(g)) although $\hat{e}_x(t)^2$ was restored as less improvement when $T_R = 0.3\,\mathrm{s}$ (dotted line in Fig. 2(g)). Therefore, their improvements are less than that of the exact restoration with $T_R = 0.5\,\mathrm{s}$.

In the basic method proposed by Hirobayashi *et al.* [10], there is no description with regard to these points. They also attempted to apply this method to speech applications without considering speech characteristics. We still have to overcome these problems associated with the basic method in order to develop this method for speech applications. Therefore, the following problems are pointed out: (i) how to precisely extract the power envelope from the observed signal, (ii) how to determine the parameters of the reverberant time and the amplitude terms (T_R and a) of the impulse response, and (iii) a lack of consideration as to whether the MTF concept can be applied to a more realistic signal.

3. IMPROVED METHOD

In this section, we point out three problems in the basic method with regard to developing it for speech applications, and we then solve these problems to improve the basic method.

3.1. Extraction of the Power Envelope

In the basic method, there is no detailed description of how to precisely extract $e_y(t)^2$ from $y(t)$. In general, there are well-known techniques for signal demodulation in AM transmission. For example, a typical amplitude demodulation method is the low-pass half-wave rectification (HWR) method [13]. This method is applied as follows: the input signal ($y(t)$) is rectified using the HWR, and then low-pass filtering (or leaky integration) of the rectified signal is used. Synchronous demodulation is also a well-known method [13]. In both methods, it is assumed that the carrier signal is sinusoidal with a single frequency. So, if either of these typical techniques is employed to extract the power envelope from an observed reverberant signal based on the MTF concept, it cannot precisely extract $e_y(t)^2$ because the carrier is a white-noise signal, but not a sinusoidal signal.

In this paper, we propose two methods that can be used to extract the power envelope. One is to use the ensemble average, which is a straightforward method based on Eq. (9). In this case, because $y(t)$ is a single observed reverberant signal, the ensemble average $\langle \boldsymbol{y}(t)^2 \rangle$ in Eq. (9) cannot be calculated directly. To cope with this problem, we introduce an assumption that a product of each white noise signal becomes the other white noise signal. Let $\hat{\boldsymbol{n}}(t)$ be a set of white noise signals constituted of a finite number of white noise elements. By assuming $\hat{\boldsymbol{y}}(t) = y(t)\hat{\boldsymbol{n}}(t)$ as a quasi-set of $\boldsymbol{y}(t)$, we can use Eq. (9) to extract the power envelope from the observed reverberant signal $y(t)$. This method is proposed as

$$\hat{e}_y(t)^2 := \mathbf{LPF}\left[\left\langle \hat{\boldsymbol{y}}(t)^2 \right\rangle\right] = \mathbf{LPF}\left[\left\langle (y(t)\hat{\boldsymbol{n}}(t))^2 \right\rangle\right]. \quad (12)$$

In this equation, we used low-pass filtering (**LPF**) as post-processing to remove the higher frequency components in the power envelope caused by the approximation of $\hat{\boldsymbol{n}}(t)$. If $\hat{\boldsymbol{n}}(t)$ can be produced completely is the same way as $\boldsymbol{n}(t)$, it is not necessary to use **LPF**.

The second method is composed of the Hilbert transform relations [14] and low-pass filtering. The Hilbert transform is also called as 90-degree shift transform from an odd or even function to an even or odd function. These relations are represented as relations between the real and the imaginary parts, or the magnitude and phase of the Fourier transform. The Hilbert transform, therefore, is often used to calculate the instantaneous amplitude of the signal. In this method, the carrier should not be a sinusoidal signal with a single frequency; instead it should be even or odd functions.

In this paper, we assume that carriers are composed of odd or even functions, so that the Hilbert transform relations can easily obtain the instantaneous amplitude of the observed signal. Thus, we can extract the power envelope of $y(t)$ using

$$\hat{e}_y(t)^2 := \mathbf{LPF}\left[\left| y(t) + j \cdot \mathbf{Hilbert}(y(t)) \right|^2\right]. \quad (13)$$

In this equation, we also used LPF as post-processing to extract the power envelope from the instantaneous amplitude.

In this paper, we use an LPF cut-off frequency of 20 Hz in both equations because an important modulation region for speech perception [15] and speech recognition is from 1 to 16 Hz [16,17].

3.2. Determination of the Impulse Response Parameters

In this method, parameters T_{R} and a must be adequately determined before the power envelope inverse filtering, so that it can become an envelope blind-deconvolution method. However, the basic method proposed by Hirobayashi *et al.* uses the known T_{R} before processing. As a result, their method has to know the exact T_{R} instead of the exact impulse response in the room acoustics before processing, and this restricts the application of their model.

In this paper, we consider the possibility of determining the reverberation time T_{R} and the amplitude term a from the observed reverberant signal to restore the power envelope. For example, it may be thought that T_{R} can be estimated from the relationship between the modulation frequency F_m and the MTF of $m(2\pi F_m)$ using Eq. (3). If F_m is a monotone frequency (i.e., $e_x(t)^2$ is a sinusoidal power envelope), it will be easy to determine T_{R} by substituting F_m and the observed modulation index into Eq. (3). However, in general, the frequency components of the power envelope do not take a single value (F_m), therefore it is difficult to precisely determine an exact T_{R} using Eq. (3).

Next, we consider over- and/or under-restoration of the power envelope with T_{R} as shown in Fig. 2(g). The inverse filtering of Eq. (9) is a type of differentiation (high-pass filtering) because Eq. (9) represents a function of the integration (or low-pass filtering). This inverse filtering produces higher frequency components in the power envelope in which peaks and dips will be emphasized. We found that the modulation index of the restored power envelope matched that of the original power envelope when the exact power envelope was obtained using a specific T_{R}.

Thus, we assume that the modulation index of the original power envelope is 1 because the power envelope has one zero-point (dip) or silence, at least, and then we define that a matching-condition between the original and the restored power envelope is to restore the modulation index reduced by reverberation. This condition can be examined by detecting a timing-point where the maximum dip of the power envelope will be 0 or the negative area of the restored power envelope will be 0. T_{R} can be estimated using

$$\hat{T}_{\mathrm{R}} = \max\left(\underset{T_{\mathrm{R,min}} \le T_{\mathrm{R}} \le T_{\mathrm{R,max}}}{\arg\min} \int_0^T \left| \min\left(\hat{e}_{x,T_{\mathrm{R}}}(t)^2, 0 \right) \right| dt \right), \qquad (14)$$

where T is signal duration and $\hat{e}_{x,T_{\mathrm{R}}}(t)^2$ is the set of candidates of the restored power envelope as a function of T_{R}. Note that the operation of "max(arg min{·})" means to determine the maximum argument of T_{R} from a timing point where the negative area of $\hat{e}_{x,T_{\mathrm{R}}}(t)^2$ approximately equals zero or a particular minimum area. This equation means the restored power envelope is constrained to prevent it being a negative power envelope. Here, $T_{\mathrm{R,min}}$ and $T_{\mathrm{R,max}}$ are the lower limited region and the upper limited region of T_{R}, respectively.

For example, three candidates of $\hat{e}_{x,T_{\mathrm{R}}}(t)^2$ for $T_{\mathrm{R}} = 0.3$, 0.5, and 1.0 s are shown in Fig. 2(g). Here, we assume that $T_{\mathrm{R,min}} = 0.0$ and $T_{\mathrm{R,max}} = 1.0$. One candidate of $\hat{e}_{x,T_{\mathrm{R}}}(t)^2$, when $T_{\mathrm{R}} = 0.3$ s, is an under-restoration of the power envelope and the other candidate of $\hat{e}_{x,T_{\mathrm{R}}}(t)^2$, when $T_{\mathrm{R}} = 1.0$ s, is an over-restoration. If we use Eq. (14) to estimate $\hat{T}_{\mathrm{R}}$, we can obtain the optimal $\hat{T}_{\mathrm{R}}$ of 0.5 from three candidates.

In the model of Hirobayashi *et al.*, they did not describe how to determine the parameter of a. In their model, however, we find that a is given the same value for both Eqs. (7) and (10), so this may not be a critical problem. In general, the effect of reverberation in the room acoustics creates a signal transmission delay rather than increasing a gain. We therefore assume that a gain of the impulse response can be approximated as the total power of the impulse response. Since a is related to the gain of the room acoustics, in this paper, the value of a is determined from the summarized $e_h(t)^2$ as follows:

$$a = \sqrt{1 \Big/ \int_0^T \exp(-13.8t/T_{\mathrm{R}}) dt}. \qquad (15)$$

In practice, if we have to set an appropriate value of a for applications in a real environment, we think we can obtain the optimal a to fit the various $e_h(t)^2$ to many power envelopes of the observed impulse responses in real environments.

3.3. Evaluation

In this section, we evaluate the improved method as to whether it can resolve problems (i)–(iii). The values of $x(t)$ consisted of the white noise multiplied by three types of power envelope:

(1) Sinusoidal: $e_x(t)^2 = 1 - \cos(2\pi F t)$;

(2) Harmonics: $e_x(t)^2 = 1 + \dfrac{1}{K}\sum_{k=1}^{K} \sin(2\pi k F_0 t + \theta_k)$;

(3) Band-limited noise: $e_x(t)^2 = \mathbf{LPF}[n(t)]$.

Here, $F = 10$ Hz, $F_0 = 1$ Hz, $K = 20$, θ_k is a random phase, and the cut-off frequency of $\mathbf{LPF}[\cdot]$ is 20 Hz. We used these artificial power envelopes to investigate the relation between each envelope and the reverberation with regard to complexity for signal contents, as related in the MTF concept, and to simultaneously evaluate the power

envelope extraction methods and the improvement in the power envelope restoration. The impulse responses, $h(t)$, consisted of five types of envelope: $e_h(t)$ in Eq. (7) with $T_R = 0.1$, 0.3, 0.5, 1.0, and 2.0 s in which a was set using Eq. (15) with each T_R, multiplied by 100 white noise carriers. All stimuli, $y(t)$, were composed through 1,500 ($= 3 \times 5 \times 100$) convolutions of $x(t)$ with $h(t)$.

As an evaluation measure, Hirobayashi *et al.* used the improvement index of the power envelope distortion [10]. This can be regarded as the improvement of SNR (where S is the original power envelope and N is the difference between the original and the estimated/restored power envelope) between the original envelope and the extracted/restored envelope. This is one of the better evaluation measures for measuring the restoration error between temporal envelopes (with magnitude), but cannot be used to judge the similarity between temporal envelopes (with shape).

In this paper, to evaluate both the error and similarity in terms of the power envelopes, we thus used the correlation (Corr) as well as the SNR as follows:

$$\mathrm{Corr}(e_x{}^2, \hat{e}_x{}^2) = \frac{\int_0^T \left(e_x(t)^2 - \overline{e_x(t)^2}\right)\left(\hat{e}_x(t)^2 - \overline{\hat{e}_x(t)^2}\right)dt}{\sqrt{\left\{\int_0^T \left(e_x(t)^2 - \overline{e_x(t)^2}\right)^2 dt\right\}\left\{\int_0^T \left(\hat{e}_x(t)^2 - \overline{\hat{e}_x(t)^2}\right)^2 dt\right\}}}, \quad (16)$$

$$\mathrm{SNR}(e_x{}^2, \hat{e}_x{}^2) = 20\log_{10}\frac{\int_0^T e_x(t)^2 dt}{\int_0^T \left(e_x(t)^2 - \hat{e}_x(t)^2\right)dt}, \quad \text{(dB)} \quad (17)$$

where the notation $\overline{e_x(t)^2}$ means the averaged $e_x(t)^2$, and $e_x(t)^2$ and $\hat{e}_x(t)^2$ are the original and the restored power envelopes, respectively.

3.3.1. Power envelope extraction

First, we compared the extraction accuracy for the power envelopes using three types of method: the ensemble average (Eq. (12)), the Hilbert transform relations (Eq. (13)), and the HWR method. Figure 4 shows the extraction accuracy for the power envelopes of sinusoidal stimuli when using the three methods (at point A in Fig. 3), with $T_R = 0$, 0.1, 0.3, 0.5, 1.0 and 2.0 s. Each point and the error bar show the mean and the standard deviation of the results. Figures 5 and 6 show the extraction accuracy for the power envelopes of the other two types of stimuli when using the three methods, with $T_R = 0$, 0.1, 0.3, 0.5, 1.0 and 2.0 s. Comparing the evaluations of all methods (Figs. 4–6), we found that the ensemble average and the Hilbert transform relations methods were far superior to the HWR method. Under all conditions with T_R and with power envelopes, we found that both of our proposed methods could precisely extract the power envelope from the observed reverberant signal, as correlation was over 0.95

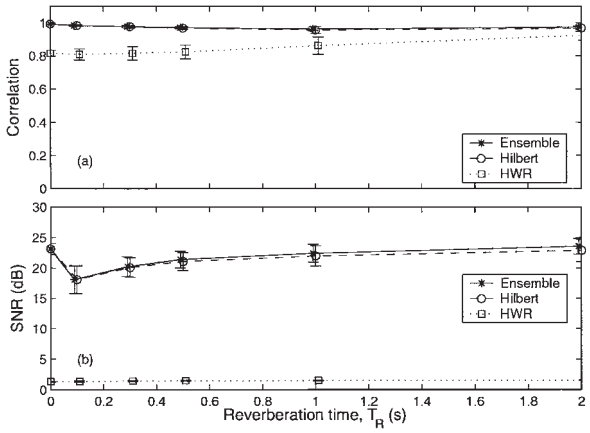

Fig. 4 Extraction accuracy of the power envelope for sinusoid stimuli: (a) correlation and (b) SNR.

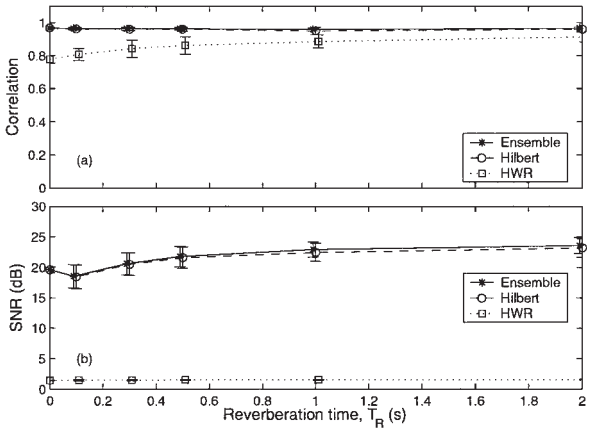

Fig. 5 Extraction accuracy of the power envelope for harmonics stimuli: (a) correlation and (b) SNR.

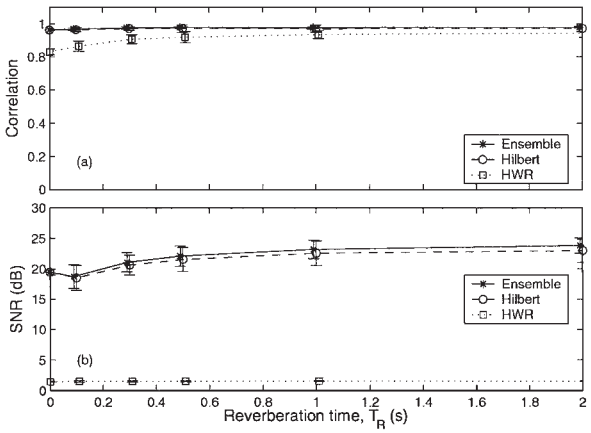

Fig. 6 Extraction accuracy of the power envelope for band-limited noise: (a) correlation and (b) SNR.

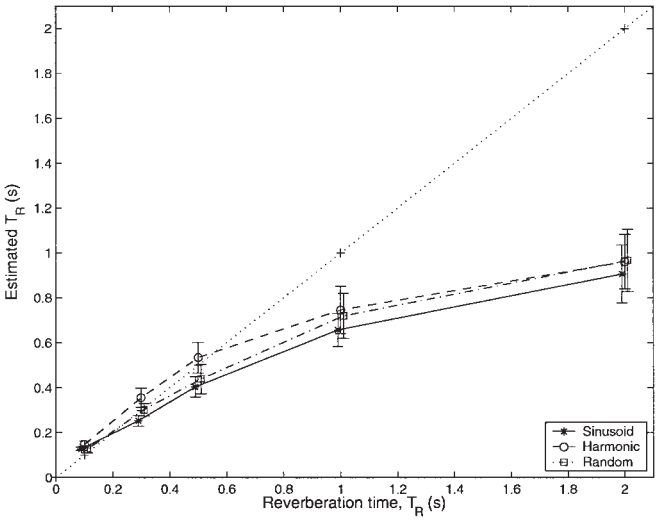

Fig. 7 Estimated reverberation time. The dotted line shows the idealized reverberation time.

and the SNR was about 25 dB, but the HWR method could not. We also found that the ensemble average was somewhat superior to the Hilbert transform method because the ensemble method used signal definition based on the MTF concept (Eqs. (4)–(8)). For the following evaluations in this paper, we mainly used the ensemble average method as the power envelope extraction method.

3.3.2. Estimation of T_R

Next, we evaluated the determination of the reverberation time $\hat{T}_R$ using Eq. (14) and the power envelope restored using the ensemble average. Figure 7 shows $\hat{T}_R$ the estimated reverberation time (at point B in Fig. 3). Each point and the error bar show the mean and the standard deviation for $\hat{T}_R$. The dotted line shows the original T_R. For sinusoidal and band-limited noise power envelopes, we found $\hat{T}_R$ matched the idealized value from 0 to about 0.5, but there were discrepancies with the idealized value above about 0.5. For the harmonics power envelope, we found $\hat{T}_R$ exceeded the idealized value from 0 to about 0.5, but there were discrepancies when the idealized value was above about 0.5. These discrepancies are discussed in the next section.

3.3.3. Restoration of the power envelope

In this paper, improved Corr and improved SNR are used to show the improvement in restoration accuracy achieved through the improved method. Improved Corr is calculated from $\mathrm{Corr}(e_x{}^2, \hat{e}_x{}^2) - \mathrm{Corr}(e_x{}^2, e_y{}^2)$ and improved SNR is calculated from $\mathrm{SNR}(e_x{}^2, \hat{e}_x{}^2) - \mathrm{SNR}(e_x{}^2, e_y{}^2)$. As shown in Figs. 2(a) and (e), the modulation index and/or the power envelope fluctuations (peaks and dips in the temporal envelope) are reduced by reverberation as a function of the reverberation time T_R. $\mathrm{Corr}(e_x{}^2, e_y{}^2)$ and $\mathrm{SNR}(e_x{}^2, e_y{}^2)$ are also reduced with increasing T_R. Therefore, if the power envelope was restored from a reverberant signal, both improved Corr and SNR should have positive values. If either measure had a negative value and the other had a positive value, it indicated that the power envelope was not adequately restored.

Figure 8 shows one set of results for the extracted power envelopes and the restored power envelopes when the improved method with a sinusoidal or a band-limited noise power envelope with $T_R = 1.0$ s was used. The top two panels show the result for a sinusoidal power envelope and bottom two panels show the result for a band-limited noise power envelopes. Each power envelope $e_y(t)^2$ in Figs. 8(a) and (c) was estimated from $y(t)$ and is shown as a dashed line. The dash-dot and solid lines indicate the original power envelope of the original signal and the power envelope of the reverberant signal calculated using Eq. (9), respectively. Negative areas, as shown in Figs.

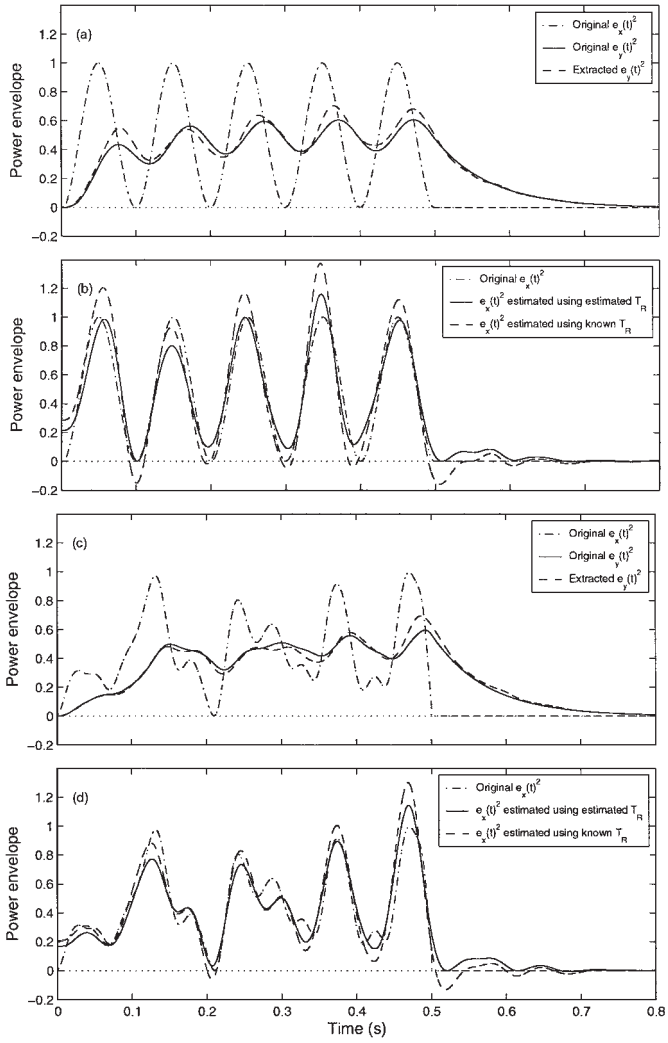

Fig. 8 Sample results: (a) extracted power envelope and (b) restored power envelope for sinusoidal stimulus with $T_R = 1.0$ s; (c) extracted power envelope and (d) restored power envelope for a band-limited noise stimulus with $T_R = 1.0$ s.

8(b) and (d), were detected at the maximum dip of the power envelope (around 0.1 s in Fig. 8(b) and around 0.2 s in Fig. 8(d)). In these cases, the estimated $\hat{T}_R$ was 0.73 s for Fig. 8(b) and 0.78 s for Fig. 8(d). In Fig. 8(b), the correlation and SNR between $e_x(t)^2$ and $\hat{e}_x(t)^2$ with the estimated $\hat{T}_R$ were 0.98 and 15.09 dB, while the correlation and SNR between $e_x(t)^2$ and $e_y(t)^2$ in Fig. 8(a) were 0.15 and 4.44 dB, respectively. The improvements in correlation and SNR were 0.83 and 10.65 dB, respectively. In contrast, the improvements in correlation and SNR for $\hat{e}_x(t)^2$ using the known T_R were 0.82 ($= 0.97 - 0.15$) and 7.59 ($= 12.03 - 4.44$) dB, respectively. Therefore, the effective improvements in correlation and SNR with regard to estimating T_R were 0.01 and 3.06 dB, respectively, in the improved method. The same evaluation for Fig. 8(d) shows that the improvements in correlation and SNR with the improved method were 0.58 ($= 0.98 - 0.40$) and 8.04 dB ($= 14.4 - 6.36$), while the improvements when using the known T_R were 0.57 ($= 0.97 - 0.40$) and 6.44 dB ($= 12.8 - 6.36$), respectively. The effective improvements in correlation and SNR were 0.01 ($= 0.58 - 0.57$) and 1.60 dB ($= 8.04 - 6.44$), respectively.

Figures 9, 10, and 11 show, respectively, the improvement in the restoration accuracy for the power envelopes of sinusoidal, harmonics, and band-limited noise stimuli (at point C in Fig. 3), with $T_R = 0.0$, 0.1, 0.3, 0.5, 1.0, and 2.0 s. These results were obtained by plotting the differences between the restored power envelope with the improved method ($\hat{e}_x(t)^2$) and with no processing ($e_y(t)^2$), as shown by the solid lines. Each point and the error bar show the mean and the standard deviation of the results. The improvements in Figs. 9–11 indicate positive values in all cases and demonstrated that the improved method could effectively restored the power envelope of the signal from the reverberant signals.

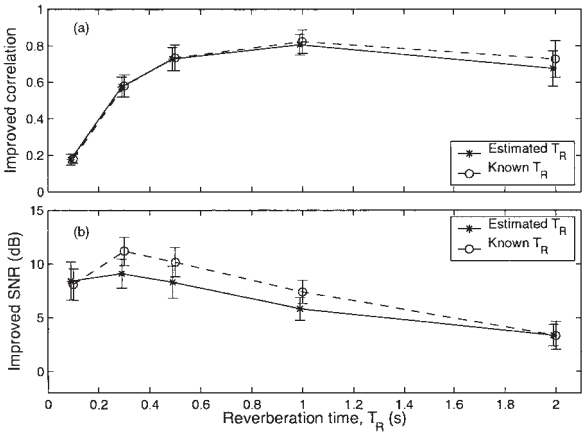

Fig. 9 Comparison with the envelope restoration accuracy for sinusoidal power envelope: (a) improved correlation and (b) improved SNR.

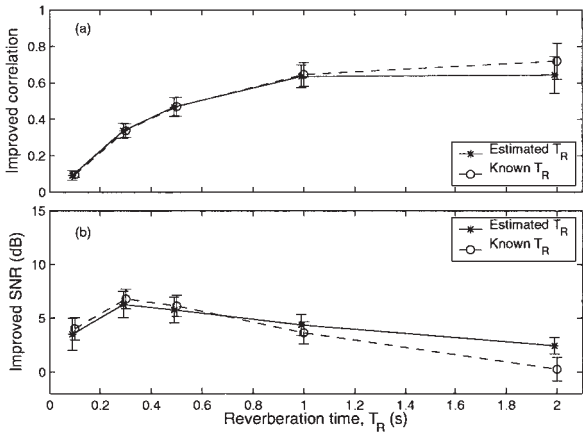

Fig. 10 Comparison with the envelope restoration accuracy for a harmonic power envelope: (a) improved correlation and (b) improved SNR.

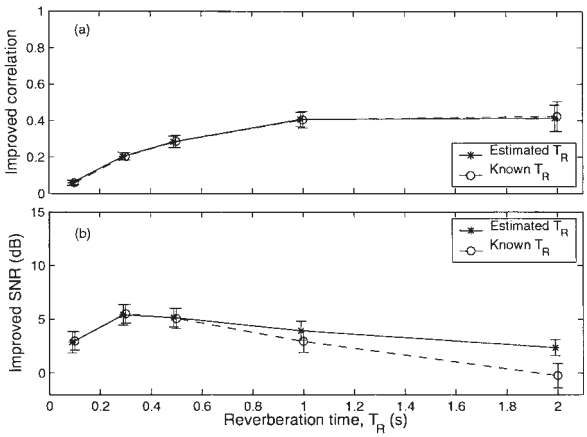

Fig. 11 Comparison with the envelope restoration accuracy for a band-limited noise power envelope: (a) improved correlation and (b) improved SNR.

We compared these results with the result obtained when using the known value of the reverberation time T_R instead of $\hat{T}_R$ in Eq. (14) (denoted by the dashed line) in the improved method (Figs. 9–11). We found no large differences in the improvements obtained through the improved method, although differences were found when using a sinusoidal power envelope with T_R of 0.3 to 1.0 s and when using harmonics and band-limited noise power envelopes with T_R of 0.5 to 2.0 s. These differences were caused by the differences in the estimation results of T_R in Fig. 7. From these results, when the reverberation time T_R is relatively short (less than about 0.5 s), it seems that the estimated reverberation time $\hat{T}_R$ should tend to match the original value. However, when T_R is above about 0.5 s, the opposite tendency seems to hold. In addition, the improved SNR when using a known T_R of 2.0 s reached zero or a negative dB value, which indicates that there was no improvement, in Figs. 10 and 11.

These results suggest that $\hat{T}_R$ should be adequately determined to accurately restore the power envelope and achieve improvements, rather than accurately estimating $\hat{T}_R$ to obtain the same value as T_R. Consequently Eq. (14) constraints the over-modulation of the power envelope or does not permit a negative power envelope in this model so that the power envelope is adequately restored. Therefore, Eq. (14) can be regarded as a reasonable constraint for restoration of the power envelope inverse filtering method. Note that the $\hat{T}_R$ estimated using Eq. (14) can completely match the original one if the power envelope extracted from a reverberant signal, using Eq. (12) or Eq. (13), can completely match the power envelope in Eq. (9).

3.3.4. Consideration of signal carriers

These results show that the improved method can restore the power envelope of the original signal from the reverberant signal through blind processing. Finally, based on these results, we also considered whether the MTF concept can be applied to realistic signals in which the carrier is harmonic but not white noise. To apply this concept, we should ensure that carriers are not correlated with each other; however, speech carriers may not remain uncorrelated. Let us thus consider the modeling in terms of this difference. Figure 12 shows the results of power envelope restoration using the improved method for the same power envelopes, except with carriers, as in Figs. 9–11. The signal carriers in Eq. (5) were 100 types of harmonics (99-order) with a fundamental frequency of 100 Hz and random phases, while the carriers of $h(t)$ in Eq. (6) were the same 100 white noise elements. We found that the improved method restored the power envelope from the reverberant signal in this case as well as is shown by the solid lines in Figs. 9–11, although there was a large deviation. This suggested that the improved method can also be applied to power envelope restoration for realistic signals.

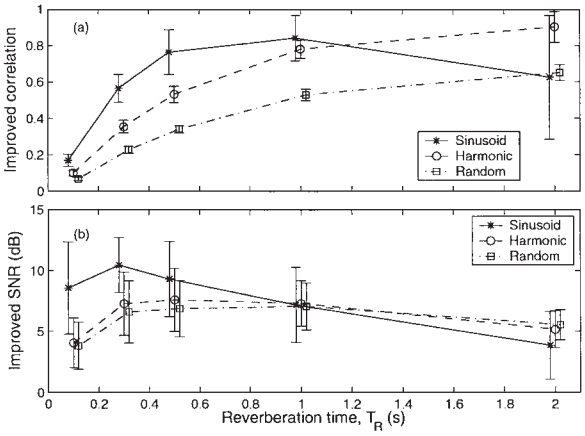

Fig. 12 Improvement of the restoration accuracy: (a) improved correlation and (b) improved SNR (three types of power envelope with harmonic carriers).

4. SUMMARY

In this paper, we have improved upon the basic method of Hirobayashi *et al.* in three ways: (i) to precisely extract the power envelope from the observed signal; (ii) to adequately determine the parameters (a and T_R) of the impulse response for the power envelope inverse filtering; and (iii) to consider whether the MTF concept can be applied to more realistic signals. We have carried out many simulations in which the improved method was applied to power envelope restoration for 1,500 types of reverberant signals where the carriers were white noise or harmonics. Our results demonstrate that the improved method can be used to accurately restore the power envelope from a reverberant signal with a white noise carrier as well as with a harmonic carrier, as a blind-restoration method. These results show that the improved method can be applied to power envelope restoration for more realistic signals in which the carrier is harmonic but not white noise. Therefore, this suggests that the improved model can also be applied to power envelope restoration for speech signals.

We still need to consider speech characteristics with regard to the temporal envelope before applying the improved method to reverberant speech, and should extend it for speech applications based on this consideration. Also, while the temporal deconvolution methods mentioned in the Introduction can also restore the envelope information from the reverberant signal, there is no significant improvement in speech intelligibility. Most of existing methods use non-processed phase information or carriers (fine-structure) affected by reverberation to synthesize the restored signal. Therefore, speech intelligibility cannot be restored without causing artifacts in the fine-structure. We stress the need to consider the carrier restoration as well as the temporal envelope restoration when attempting to both dereverberate the signal from a reverberant signal and improve speech intelligibility.

Thus, in our future work, our next step will be to (1) consider speech characteristics such as co-modulation in the temporal envelope mentioned above, (2) extend this model into a filterbank model for speech applications, and (3) reconsider what constitutes a reasonable trade-off between co-modulation bandwidths and the minimum bandwidth to be held for the MTF in a sub-band. We will then (4) reconsider how to restore the carrier and how to resynthesize the dereverberated signal from the restored power envelope using results (1)–(3) and the restored carrier. As a final step, we will be able to test whether our approach can suppress the reduction in speech intelligibility caused by reverberation.

ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid for Science

Research from the Ministry of Education (No. 14780267) and by special coordination funds for promoting science and technology (supporting young researchers with fixed-term appointments).

REFERENCES

[1] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, **66**, 165–169 (1979).

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech Signal Process.*, **ASSP-36**, 145–152 (1988).

[3] H. Wang and F. Itakura, "Realization of acoustic inverse filtering through multi-microphone sub-band processing," *IEICE Trans. Fundam.*, **E75-A**, 1474–1483 (1992).

[4] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acustica*, **28**, 66–73 (1973).

[5] T. Houtgast, H. J. M. Steeneken and R. Plomp, "Predicting speech intelligibility in room acoustics," *Acustica*, **46**, 60–72 (1980).

[6] 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. Am.*, **77**, 1069–1077 (1985).

[7] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," *Proc. ICASSP 82*, pp. 156–159 (1982).

[8] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," *Proc. ICSLP 96*, pp. 889–892 (1996).

[9] J. Mourjopoulos and J. K. Hammond, "Modelling and enhancement of reverberant speech using an envelope convolution method," *Proc. ICASSP 83*, pp. 1144–1147 (1983).

[10] S. Hirobayashi, H. Nomura, T. Koike and M. Tohyama, "Speech waveform recovery from a reverberant speech signal using inverse filtering of the power envelope transfer function," *IEICE Trans. A*, **J81-A**, 1323–1330 (1998).

[11] M. R. Schroeder, "Modulation transfer functions: definition and measurement," *Acustica*, **49**, 179–182 (1981).

[12] A. Papouris, *Probability, Random Variables and Stochastic Processes*, 3rd Ed. (MacGraw-Hill, Inc., New York, 1991).

[13] F. J. Taylor, *Principles of Signals and Systems* (MacGraw-Hill, Inc., New York, 1994).

[14] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing* (Prentice-Hall, Inc., London, 1975).

[15] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, **105**, 2783–2791 (1999).

[16] N. Kanedera, T. Arai, H. Hermansky and M. Pavel, "On the importance of various modulation frequencies for speech recognition," *Proc. EuroSpeech 97*, 1079–1082 (1997).

[17] N. Kanedera, T. Arai and T. Funada, "Robust automatic speech recognition emphasizing important modulation spectrum," *IEICE Trans. D-II*, **J84-D-II**, 1261–1269 (2001).

APPENDIX: PROOF OF EQ. (9)

Let $\boldsymbol{v}(t)$ and $f(v,t)$ to be a random variable and the density function of $\boldsymbol{v}(t)$, respectively. The ensemble average $\langle \boldsymbol{v}(t) \rangle$ is defined as [12]

$$\langle \boldsymbol{v}(t) \rangle = \int_{-\infty}^{\infty} v f(v,t) dv. \tag{A·1}$$

$\langle \boldsymbol{y}(t)^2 \rangle$ in Eq. (9) equals the ensemble average of the convolution of $x(t)$ with $h(t)$, as follows.

$$\langle \boldsymbol{y}(t)^2 \rangle = \left\langle \left\{ \int_{-\infty}^{\infty} \boldsymbol{x}(\tau) \boldsymbol{h}(t-\tau) d\tau \right\}^2 \right\rangle. \tag{A·2}$$

Here, using Eqs. (5) and (6), we can obtain

$$\begin{aligned}
&\langle \boldsymbol{y}(t)^2 \rangle \\
&= \left\langle \int_{-\infty}^{\infty} e_x(\tau_1) \boldsymbol{n}_1(\tau_1) e_h(t-\tau_1) \boldsymbol{n}_2(t-\tau_1) d\tau_1 \right. \\
&\quad \left. \times \int_{-\infty}^{\infty} e_x(\tau_2) \boldsymbol{n}_1(\tau_2) e_h(t-\tau_2) \boldsymbol{n}_2(t-\tau_2) d\tau_2 \right\rangle \\
&= \int_{-\infty}^{\infty} e_x(\tau_1) e_h(t-\tau_1) \int_{-\infty}^{\infty} e_x(\tau_2) e_h(t-\tau_2) \\
&\quad \times \langle \boldsymbol{n}_1(\tau_1) \boldsymbol{n}_1(\tau_2) \rangle \\
&\quad \times \langle \boldsymbol{n}_2(t-\tau_1) \boldsymbol{n}_2(t-\tau_2) \rangle d\tau_1 d\tau_2,
\end{aligned} \tag{A·3}$$

where $\boldsymbol{n}_k(t)$, $k = 1, 2$, are the mutually independent respective white noise (Gaussian) random variables with mean of 0 and variance of 1; hence,

$$\langle \boldsymbol{n}_k(t) \boldsymbol{n}_k(t-\tau) \rangle = \delta(\tau). \tag{A·4}$$

Using $\tau = \tau_1 = \tau_2$, we can derive the following.

$$\begin{aligned}
&\langle \boldsymbol{y}(t)^2 \rangle \\
&= \int_{-\infty}^{\infty} e_x(\tau_1) e_h(t-\tau_1) \int_{-\infty}^{\infty} e_x(\tau_2) e_h(t-\tau_2) \\
&\quad \times \delta(\tau_2 - \tau_1)^2 d\tau_1 d\tau_2 \\
&= \int_{-\infty}^{\infty} e_x(\tau)^2 e_h(t-\tau)^2 d\tau \\
&= e_x(t)^2 * e_h(t)^2.
\end{aligned} \tag{A·5}$$

On the other hand, if we calculate the left side of Eq. (A·2) directly, we can obtain

$$\begin{aligned}
\langle y(t)^2 \rangle &= \langle e_y(t)^2 \boldsymbol{n}(t)^2 \rangle \\
&= e_y(t)^2 \langle \boldsymbol{n}(t)^2 \rangle \\
&= e_y(t)^2,
\end{aligned} \tag{A·6}$$

where $\langle \boldsymbol{n}(t)^2 \rangle = \delta(0)$. Hence, from Eqs. (A·5) and (A·6), we can obtain

$$e_y(t)^2 = e_x(t)^2 * e_h(t)^2. \tag{A·7}$$