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

PAPER

A speech dereverberation method based on the MTF concept in power envelope restoration

Masashi Unoki*, Keigo Sakata†, Masakazu Furukawa‡ 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 21 January 2004*)

Abstract: We previously proposed an improved method for restoring the power envelope from a reverberant signal, based on the modulation transfer function (MTF) concept in order to resolve the problems of Hirobayashi's method. In this paper, to apply our improved method to reverberant speech, we consider three issues related to speech applications: (i) how to apply the improved method to speech dereverberation based on co-modulation characteristics; (ii) whether the MTF concept can also be applied in the sub-band for reverberant signals; and (iii) whether power envelope inverse filtering should be done separately in each channel. We propose an extended filterbank model based on these considerations. We have carried out 15,000 simulations of the power envelope restoration for reverberant speech signals, and our results have shown that the proposed model can adequately restore the power envelopes in all channels from reverberant speech signals. We also found that the estimation of the reverberation time should be done separately in each channel to improve the restoration accuracy of the power envelope.

Keywords: Power envelope, Reverberation time, Inverse filtering, Modulation transfer function (MTF), Co-modulation characteristics

PACS number: 43.72.Ew [DOI: 10.1250/ast.25.243]

1. INTRODUCTION

Speech dereverberation is an important issue concerning various kinds of speech signal processing, such as speech-emphasis for transmission systems and hearing aid systems, as well as 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. There are, for example, the methods of Neely and Allen [1], Miyoshi and Kaneda [2], and Wang and Itakura [3]. These methods use a single microphone or microphone array to dereverberate signals. Miyoshi and Kaneda's method and Wang and Itakura's method can be applied to room acoustics with non-minimum phase characteristics, but Neely and Allen's method can be applied only to the minimum phase characteristics in the room acoustics. 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.

Recently, Nakatani and Miyoshi proposed a blind dereverberation method for a single-channel speech signal based on harmonic structure without measuring the impulse response of the room acoustics [4]. This method, however, requires accurate estimation of the fundamental frequency from the reverberant speech, and they pointed out that it is difficult to meet this requirement. It also seems that this method does not precisely dereverberate the parts corresponding to consonants.

On the other hand, temporal envelope inverse filtering methods have been proposed that not only restore the

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

temporal envelope from reverberant speech, but also improve speech intelligibility that is degraded by reverberation. There are, for example, the methods of Langhans and Strube [5], and Avendano and Hermansky [6]. These method are based on the modulation transfer function (MTF) concept [7,8] and use temporal envelope deconvolution through suitable high-pass filtering on the power spectrum based on the short-term Fourier transform (STFT).

There are also the methods of Mourjopoulos and Hammond [9] and Hirobayashi *et al.* [10]. These methods are also based on the MTF concept and restore the temporal envelope from reverberant speech based on a single- or multi-channel filterbank rather than on the STFT. These methods differ, though, in their signal definition with regard to the temporal envelope (amplitude or power) and the carrier (sine-wave or white noise) based on the amplitude modulation (AM) representation. With regard to the filterbank model, Hirobayashi *et al.* stated that the optimal bandwidth was a constant bandwidth of 100 Hz rather than a constant-Q bandwidth, such as the octave bandwidths used for female speech input [11]. With regard to evaluation of both methods, Hirobayashi *et al.* reported that the power envelope restoration method is superior to the envelope deconvolution methods [12].

These last two methods ([9,10]) represent attempts to restore the temporal envelope from reverberant speech while the first two methods ([5,6]) attempt 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 (temporal fluctuation or modulation index) from reverberant signals 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, they will be useful for preprocessing in 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.* [10,11] will be used as a reasonable model in our work.

In our previous work [13,14], as the first step, we reconsidered the power envelope inverse filtering method proposed by Hirobayashi *et al.* [10]. We pointed out that their methods still have three important problems: (1) how to precisely extract the power envelope, (2) how to determine the model parameters, and (3) whether the MTF concept can be applied to more realistic signals where the carrier is a sinusoidal, but not white noise, signal. Moreover, since Hirobayashi *et al.* applied their basic method to speech signals without carefully considering speech characteristics, we point out that there are other issues regarding speech applications: (i) how to apply the basic method to speech restoration based on co-modulation characteristics; (ii) whether the MTF concept can be applied in a sub-band for reverberant signals; and (iii) whether power envelope inverse filtering should be done separately in each channel. We solved the first group of problems (1)–(3), and then we improved this basic method [13,14], as described in Sect. 2.2 and Sect. 2.3. In this paper, we consider the second group of problems (i)–(iii) and propose a speech dereverberation method based on the MTF concept in the power envelope restoration.

This paper is organized as follows. In Chap. 2, the underlying concept of a basic power envelope restoration method based on the MTF is described. This method restores the power envelope of the reverberant signal as the power envelope is co-modulated over the whole frequency range. However, since the power envelopes of speech signals may not be co-modulated over the whole frequency range, our improved model should be extended to a filterbank (sub-band) model. Section 3 describes application considerations regarding speech signals, as related to the above extension: how does a reasonable bandwidth relate to the co-modulation characteristics; the applicability of the MTF concept in a sub-band; and the usefulness of separately estimating the reverberant time in each channel using another method. Section 4 describes the extended method for speech dereverberation and discusses our results. Section 5 gives our conclusions.

2. RESTORATION METHOD

2.1. 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 [7] 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

$$\boldsymbol{y}(t) = \boldsymbol{x}(t) * \boldsymbol{h}(t), \tag{1}$$

$$\boldsymbol{x}(t) = e_x(t)\boldsymbol{n}_1(t), \tag{2}$$

$$\boldsymbol{h}(t) = e_h(t)\boldsymbol{n}_2(t), \tag{3}$$

$$e_h(t) = a\exp(-6.9t/T_{\mathrm{R}}), \tag{4}$$

$$\langle \boldsymbol{n}_k(t)\boldsymbol{n}_k(t-\tau)\rangle = \delta(\tau), \tag{5}$$

where "$*$" denotes the convolution operation, $e_x(t)$ and $e_h(t)$ are the envelopes of $\boldsymbol{x}(t)$ and $\boldsymbol{h}(t)$, $\boldsymbol{n}_1(t)$ and $\boldsymbol{n}_2(t)$ are the mutually independent respective white noise (random variable) functions, and $\langle\cdot\rangle$ is the ensemble average operation [15]. The parameters of the impulse response, a

and T_R, are a constant amplitude term and the reverberation time, respectively [10].

In the model, the power envelope of the reverberant signal, $e_y(t)^2$, can be determined as [10,13,14]

$$\langle \boldsymbol{y}(t)^2 \rangle = \int_{-\infty}^{\infty} e_x(\tau)^2 e_h(t-\tau)^2 d\tau = e_y(t)^2. \quad (6)$$

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 the 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 the sampling frequency f_s is 20 kHz. Then, the transfer functions of power envelopes $E_x(z)$, $E_h(z)$, and $E_y(z)$ are assumed to be the z-transforms of $e_x[n]^2$, $e_h[n]^2$, and $e_y[n]^2$, respectively. Thus, the transmission function of the power envelope of the original signal, $E_x(z)$, can be determined from

$$E_x(z) = \frac{E_y(z)}{E_h(z)} = \frac{E_y(z)}{a^2}\left\{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right)z^{-1}\right\}. \quad (7)$$

Finally, the power envelope $e_x[n]^2$ can be obtained from the inverse z-transform of $E_x(z)$ [10,13,14]. In this paper, for convenience, we use symbols of continuous definition.

As mentioned in [13,14], in the basic method proposed by Hirobayashi *et al.*, there are two main problems: (1) how to precisely extract the power envelope from the observed signal and (2) how to determine the parameters of the reverberant time and the amplitude term (T_R and a) of the impulse response of the room acoustics. We have resolved these problems as follows. Figure 1 shows a block diagram of the improved method [13,14].

2.2. Power Envelope Extraction

In general, there are well-known techniques for amplitude demodulation such as low-pass half-wave rectification (HWR), but these do not work in this model because the carrier is a white-noise rather than a monotone sine-wave. In a previous paper [13,14], we proposed two methods that can be used to extract the power envelope. One is a method using ensemble averaging as follows.

$$\hat{e}_y(t)^2 := \mathbf{LPF}[\langle \hat{\boldsymbol{y}}(t)^2 \rangle] = \mathbf{LPF}[\langle (y(t)\hat{\boldsymbol{n}}(t))^2 \rangle], \quad (8)$$

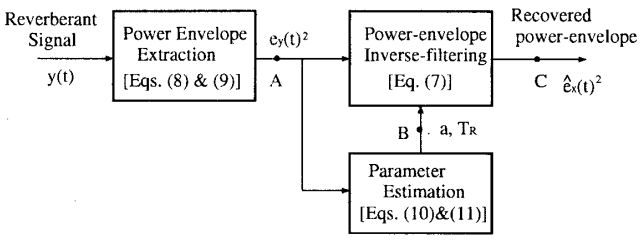

Fig. 1 Block diagram of the power envelope inverse filtering method.

where $\hat{\boldsymbol{y}}(t) = y(t)\hat{\boldsymbol{n}}(t)$, $\hat{\boldsymbol{n}}(t)$ is a set of white noise, $\hat{\boldsymbol{y}}(t)$ is a quasi-set of the observed $y(t)$, and $\mathbf{LPF}[\cdot]$ is a low-pass filtering. This method is an approximation method used instead of Eq. (6) because $y(t)$ is a single observed signal and so Eq. (6) cannot be calculated directly.

The second is a method using the Hilbert transform relations:

$$\hat{e}_y(t)^2 := \mathbf{LPF}[|y(t) + j \cdot \mathbf{Hilbert}(y(t))|^2]. \quad (9)$$

This method is based on calculation of the instantaneous amplitude of the signal. Both methods used low-pass filtering as post-processing to remove the higher frequency components in the power envelope.

In this paper, we use an LPF cut-off frequency of 20 Hz in both equations because an important modulation region for speech perception [16] and speech recognition is from 1 to 16 Hz [17,18].

2.3. Determination of the Impulse Response Parameters

In our previous paper [13,14], we assumed that the modulation index of the original power envelope is 1 and that a gain of the impulse response can be regarded as the normalized total power of the impulse response. We then proposed that the impulse response parameters (T_R and a) can be determined as follows [13,14].

$$\hat{T}_R = \max\left(\underset{T_{R,\min} \le T_R \le T_{R,\max}}{\arg\min} \int_0^T |\min(\hat{e}_{x,T_R}(t)^2, 0)| dt\right), \quad (10)$$

and

$$a = \sqrt{1 \Big/ \int_0^T \exp(-13.8t/T_R) dt}, \quad (11)$$

where T is signal duration and $\hat{e}_{x,T_R}(t)^2$ is the candidates of the power envelope {restored} as a function of T_R. Here, $T_{R,\min}$ and $T_{R,\max}$ are the lower limited region and the upper limited region, respectively, of T_R. Equation (10) means that T_R is determined with the restored power envelope constrained to prevent it being a non-negative power envelope and to prevent over-modulation (that the modulation index not be over 1). Equation (11) means that a is determined as the summarized power envelope of the impulse response that is normalized as 1.

3. APPLICATION TO SPEECH

Previously, we proposed the improved method described in the above and have showed that our proposed model can precisely restore the power envelope from reverberant artificial signals [13,14].

In this paper, we consider the model's application to reverberant speech signals.

3.1. Co-modulation Characteristics

Both the basic method and the improved method restore the power envelope from the reverberant signal as the power envelope is co-modulated across the whole frequency region. In general, however, the power envelope of speech may not have the same power envelopes across the entire range of frequencies. The first point to be considered is whether the power envelopes of speech for all frequencies have a co-modulation characteristic. If they do not, the next consideration should be what is an appropriate bandwidth which can be regarded as similar to the co-modulation characteristics for speech.

We examined the correlation between the power envelopes on channels (calculated using Eq. (9)) in a constant narrow-band (40 Hz) filterbank to verify the co-modulation characteristics. In this consideration, the speech signals were three Japanese sentences (/aikawarazu/, /shinbun/, and /joudan/) uttered by ten speakers (five males: Mau, Mht, Mnm, Mtm, and Mtt and five females: Faf, Ffs, Fkn, Fsu, and Fyn) from the ATR-database [19]. Correlation conditions of 0.50, 0.60, 0.75, 0.80, 0.85, 0.90, 0.92, 0.94, 0.96, and 0.98 were used. Bandwidths obtained from co-modulation characteristics were calculated from the contours showing the region of correlation for each condition, and then they were averaged over all channels (from 0.50 and 0.98) and for all speech signals.

Figure 2, for example, shows the outcome of this analysis for speech using a constant narrow-band filterbank. The speech signal was a Japanese sentence (/aikawarazu/) uttered by a male speaker (Mau). In Fig. 2 (a), the contour shows the region of correlation over 0.98 as a co-modulation characteristic. From this contour, a bandwidth of 90 Hz is regarded as an approximate constant bandwidth for the entire frequency range. Through the same procedure, the contour in Fig. 2 (b) was obtained from the region of correlation over 0.80 as a co-modulation characteristic. In this case, the approximated constant bandwidth is about 250 Hz. From these results, we found that these contours tended to be wide in middle and/or higher frequency regions as the correlation related to the co-modulation characteristics fell. Therefore, the power envelopes of speech have to be analyzed in a sub-band since the basic method restores the same power envelope at all frequencies.

Figure 3 shows the relationship between the correlation and the averaged bandwidth obtained from the above calculation (whose results are shown in Fig. 2) with regard to all correlation conditions (0.50–0.98). This figure was plotted as a function of correlation. Each point shows the average bandwidth on the horizontal axis and the error bar shows the standard deviation of the results. The standard deviation increased as the correlation between the power envelopes in adjacent channels fell. This figure indicates that correlation between the power envelopes on channels tends to fall as the averaged bandwidth widens. This result also suggests that we have to extend our improved method to the filterbank model to separately apply it to speech dereverberation in the sub-band. It is desirable to use the narrowest bandwidth possible on the channel for speech applications.

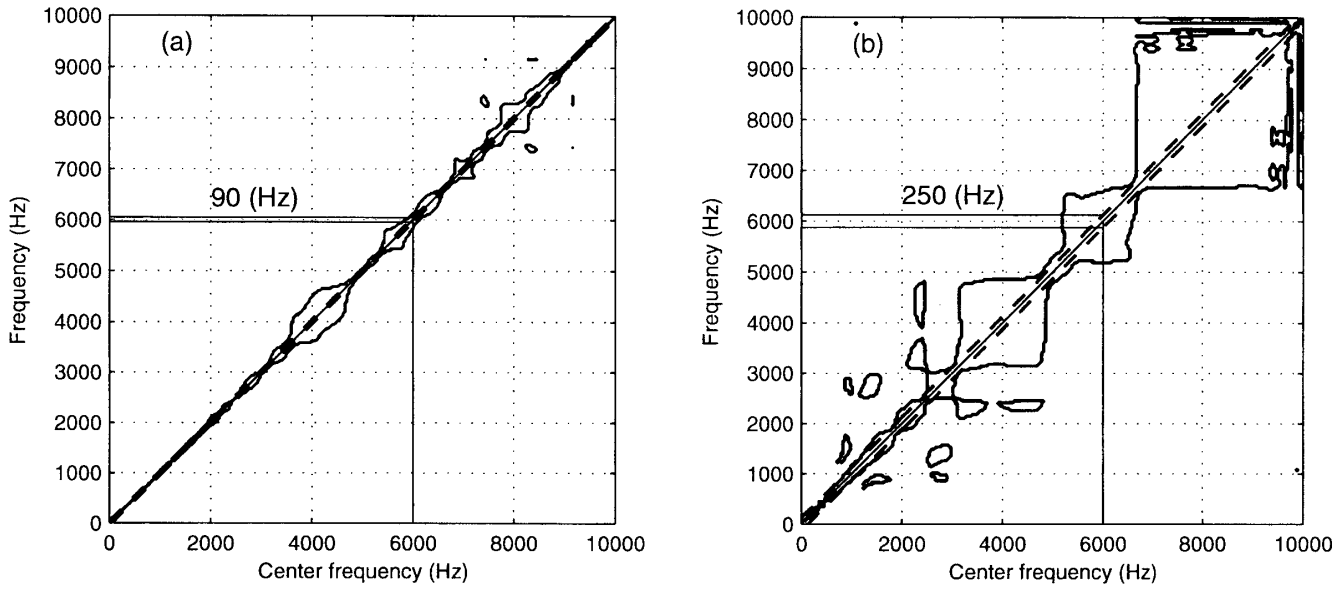

Fig. 2 Correlation between the power envelopes of one channel and the adjacent channels for speech on the filterbank. The dashed line shows the region of an approximated constant bandwidth ((a) when correlation was over 0.98 and (b) when correlation was over 0.80).

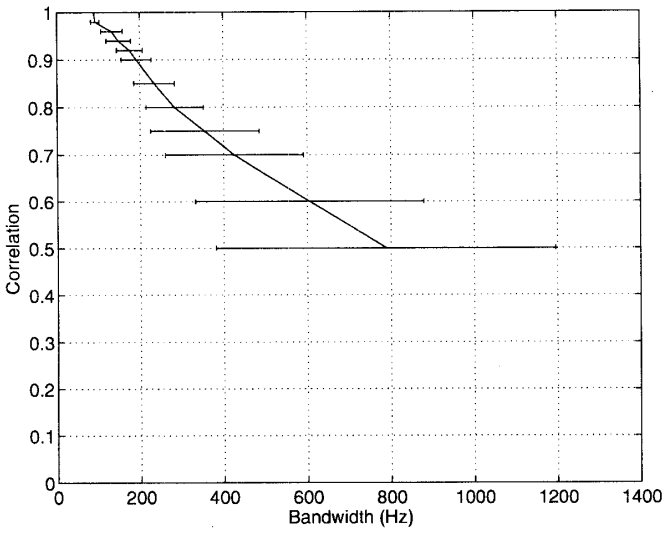

Fig. 3 Relationship between the averaged bandwidth and correlation between the power envelopes on speech signal channels.

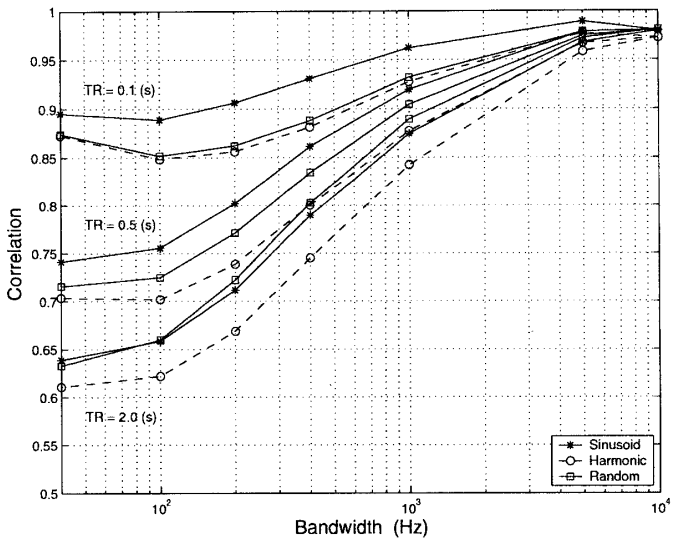

Fig. 4 Relationship between power envelope correlation and bandwidth in the filterbank with regard to the MTF concept (artificial signals).

3.2. Applicability of the MTF Concept

Based on the above consideration, we next considered the applicability of the MTF concept in the sub-band. Thus, the second point to consider was whether the MTF concept can be applied to reverberant speech signals. To apply this concept, we should ensure that carriers are not correlated with each other in the sub-band (the white noise in the model or the harmonics in the application), but speech carriers may remain correlated. We have shown that the model can restore the power envelope from reverberant signals even though the carrier is a harmonic, but not white noise, signal in which the power envelopes are the same (see Sect. 3.3 in [13]; Sect. 3.3.4 in [14]). However, this applies to carriers within the whole frequency range (or in waveforms). As explained, we have to consider the applicability of the MTF concept in a channel according to the filterbank model. We expected that we would not be able to apply the MTF concept through a narrow sub-band because the assumption of mutual independence between the carriers would not consistently hold.

In this paper, with regard to signal definition based on the MTF concept (Eqs. (1)–(5)), we have examined the consistency between $e_y(t)^2$ calculated from $e_x(t)^2$ using Eq. (6) according to the MTF concept and $e_y(t)^2$ extracted from $y(t)$ using Eq. (8), to test this expectation. The values of $x(t)$ consisted of the white noise multiplied by three types of power envelope:

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

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

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

Here, $F = 10\,\mathrm{Hz}$, $F_0 = 1\,\mathrm{Hz}$, $K = 20$, θ_k is a random phase, and the cut-off frequency of **LPF**[·] is 20 Hz. The impulse responses, $h(t)$, consisted of five types of envelope, with $T_R = 0.1$, 0.3, 0.5, 1.0, and 2.0 s, 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 $h(t)$. The channel bandwidth was 40, 100, 200, 400, 1,000, 5,000, or 10,000 Hz. Reverberation time T_R was 0.1, 0.3, 0.5, 1.0, and 2.0 s.

Figure 4 shows the correlation results with the MTF concept in the sub-band for artificial signals related in the model definition (Eqs. (1)–(5)). In this figure, results are shown only for T_R of 0.1, 0.5, and 2.0 s. The correlation when the MTF concept was applied tended to fall as the channel bandwidth became narrower and/or T_R increased. Results under the other conditions were similar to those in Fig. 4. This figure indicates that it is desirable to widen the bandwidth of the channel to satisfy the MTF concept for artificial signals.

We also examined the consistency between the power envelopes under the MTF concept in the sub-band for speech signals using the same input (three sentences and ten speakers) as described in Sect. 3.1. The original power envelope $e_x(t)^2$ in each channel was set to be the calculated envelope from the original speech because we did not know original power envelope of the speech signal.

Figure 5 shows the correlation results with the MTF concept in the sub-band for speech signals. In this figure, results are shown for only T_R of 0.1, 0.5, and 2.0 s. Contrary to our expectation, the correlation did not drastically fall as the channel bandwidth became narrower or T_R increased. In contrast with Fig. 4, the correlation fell a little with wider bandwidths. There are two points of difference for evaluation when comparing these figures:

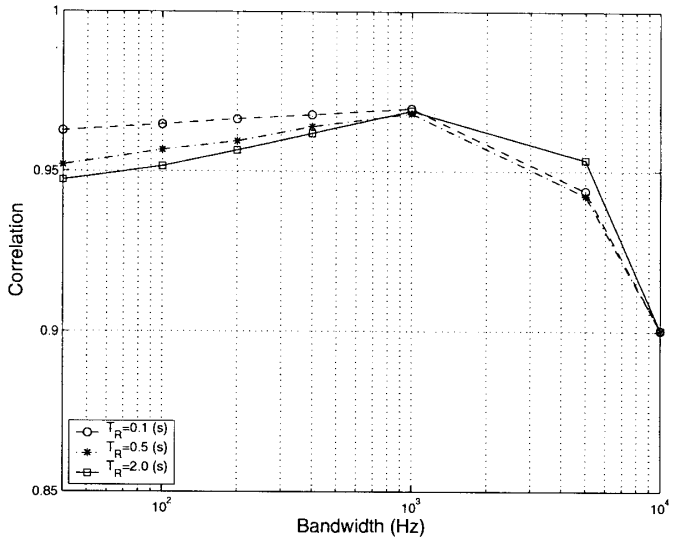

Fig. 5 Relationship between power envelope correlation and bandwidth in the filterbank with regard to the MTF concept (speech signals).

one is the difference between the setting of the original power envelopes for artificial and speech signals and the other is the difference between the carriers in the channels (band-limited white noise and band-limited harmonic signals). With regard to the first point, whether the accuracy of the power envelope extraction is included in the total evaluation may affect the above results, but the effect seems to be small. However, the second difference may be significant because the mutual independence between the carriers of $x(t)$ and $h(t)$ for a speech signal in a sub-band (a band-limited harmonic signal in $x(t)$ and band-limited white noise in $h(t)$) is shown in Fig. 5, while that for a sub-band for artificial signals (band-limited white noise signals in both $x(t)$ and $h(t)$) is shown in Fig. 4.

Consequently, the correlation between the carriers for speech signals in the sub-band was lower than that for artificial signals, which indicates that the MTF concept (Eq. (6)) can be applied in each sub-band for speech signals. Moreover, the correlation at narrower bandwidths was still high compared with the Fig. 4 results. Results under the other conditions were almost the same. These results suggest that we can use a narrower channel bandwidth for speech applications, in contrast what is shown in Fig. 4.

We also think that there is a reasonable trade-off between the divided co-modulation bandwidths and the bandwidth to be held for the MTF concept expressed by Eq. (6) when we compare these results with Figs. 3 and 5. Based on the above results, a reasonable trade-off point between the right-down slope in Fig. 3 and the left-down slope in Fig. 5 is near 100 Hz. Thus, in this paper, we regarded a bandwidth of 100 Hz to be reasonable for speech applications.

3.3. Another Estimation Method for T_R

Next, we reconsider a method for estimating the reverberation time T_R in the channel. In general, to reasonably restore the power envelope of a reverberant signal based on the MTF concept, we can use the proposed method (Eq. (10)) to estimate T_R [13,14]. For example, Fig. 6 (a) shows a sinusoidal power envelope of 10 Hz (solid line) and a reverberant power envelope with $T_R = 0.5$ s (dashed line). In this figure, we can precisely estimate $\hat{T}_R$ using Eq. (10) and then restore $\hat{e}_x(t)^2$ from $e_y(t)^2$ using Eq. (7) with $\hat{T}_R$ of 0.2 to 0.8 in 0.2 steps as described in Sect. 3.2. Equation (10) has an important effect of constraining negative power envelopes and this also works well in simulations of artificial signals, so this equation is useful for the general case.

However, this processing cannot work in the case shown in Fig. 6 (b), which seems to be a special case. This is because we assumed that the modulation index of the original signal was set to be 1, and the silence interval corresponding to a modulation index of 1 is not long [13,14]; the processing will not work if this assumption is not satisfied. This special case will often occur in the power envelope of a speech signal in a narrow-band channel. This problem also occurs when there is a long silence in the power envelope on a channel even if there is no silence in the power envelope of the waveform. Since this problem is caused by dividing a whole band into sub-bands, we need a better method to estimate T_R in a channel when there is a long silence within the power envelope.

In this section, instead of focusing on the negative area in Eq. (10), we focus on the shifting variations at a related position at the same threshold, depending upon inverse filtering with $\hat{T}_R$, as shown in Fig. 6 (c) (denoted by "$*$"). The dotted lines show the over- and/or under-restored envelopes when various values of $\hat{T}_R$ were used. Thus, we propose another T_R estimation method that can be used for sub-band signals even if there is a long silence, as follows.

$$\hat{T}_R = \underset{T_{R,\min} \le T_R \le T_{R,\max}}{\arg\min} \frac{dT_P(T_R)}{dT_R}, \tag{12}$$

$$T_P(T_R) = \min\left(\underset{t_{\min} \le t \le t_{\max}}{\arg\min} |\hat{e}_{x,T_R}(t)^2 - \theta| \right), \tag{13}$$

where θ is a threshold for detecting a point (described by "$*$" in Fig. 6 (c)) from the maximum of $e_y(t)^2$. Here, we set θ to a value of 0.01 multiplying the maximum envelope $e_y(t)^2$ (value of -20 dB down) within $t_{\min}$ to $t_{\max}$, where $t_{\min}$ and $t_{\max}$ are, respectively, the lower and upper limited regions for determining a point. The purpose of this method is to detect the point where there is the smallest shift with the varying T_R corresponding to an idealized T_R. Figure 6 (d) shows the variation of the shifting point with any T_R. In this figure, T_R at $T_P = 0.2$ could be used to identify the

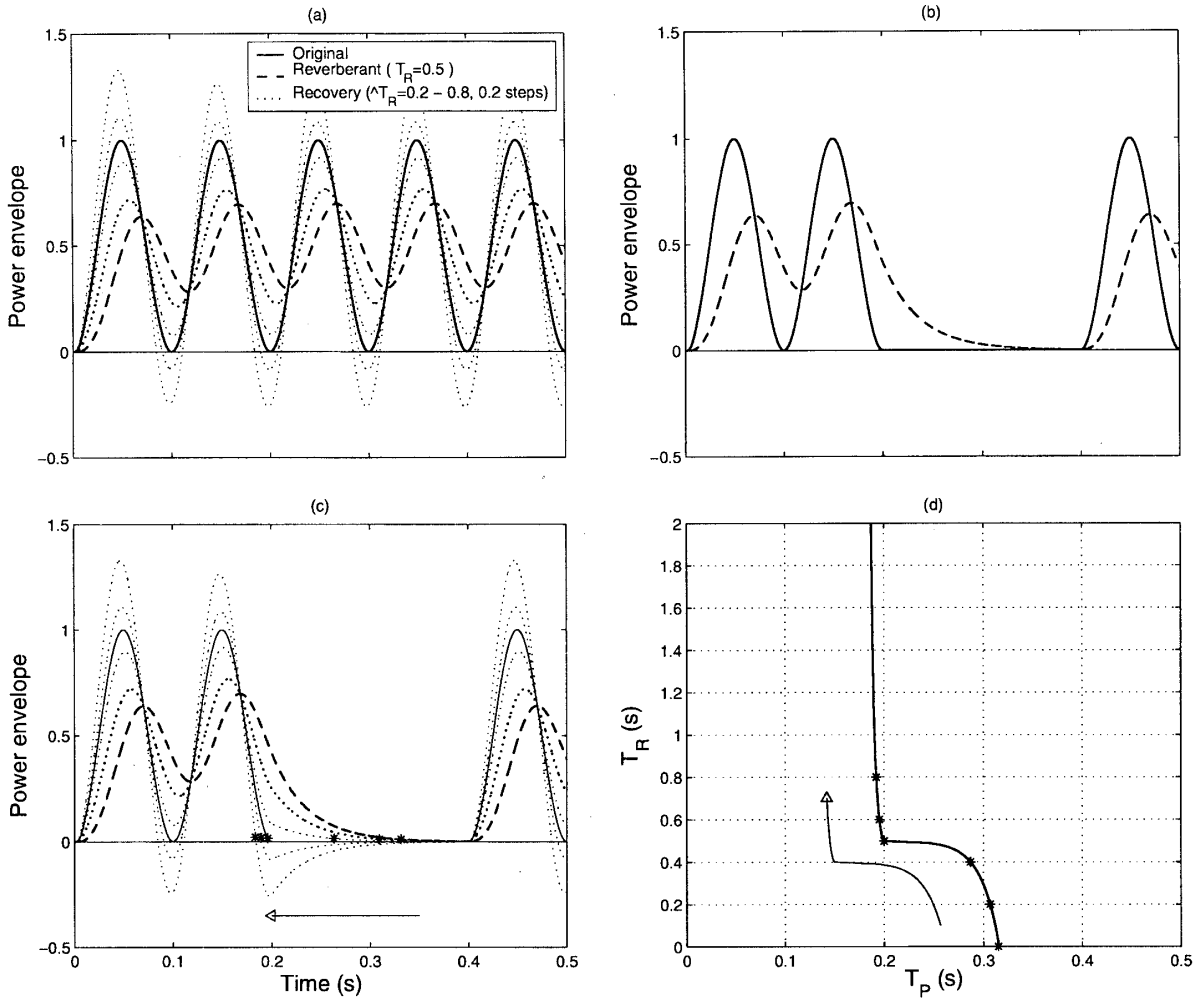

Fig. 6 An example of power envelopes based on the MTF and improved estimation of T_R: (a) envelopes with no silence, (b) an envelope with a long silence, (c) candidates of the restored envelopes, and (d) $T_P(t)$.

correct T_R (= 0.5), where we detected a rapid variation at $T_P(T_R)$. Therefore, we can precisely estimate T_R using Eq. (12). Each point denoted by "∗" varied from right to left with increasing T_R in Fig. 6 (c) and varied from right-bottom to left-top with increasing T_R in Fig. 6 (d).

Figure 7 shows $\hat{T}_R$ estimated using Eq. (12) from the extracted power envelope (which is the same evaluation as shown in Fig. 4 of [13] or Fig. 7 of [14]). In these stimuli, the power envelope was composed of the same two envelopes (sinusoidal, harmonics, or band-limited noise) as described in Sect. 3.3 and one long silence (1 s) as shown in Fig. 6 (b). Each point and error bar shows the mean and standard deviation for $\hat{T}_R$. The dotted line in Fig. 7 shows the idealized $\hat{T}_R$. We found that $\hat{T}_R$ matched the idealized value from 0 to about 0.5, but there were discrepancies with the idealized value above about 0.5. This is also related to a reasonable constraint for an adequate restored power envelope, with the same meaning as in our previous study [13,14].

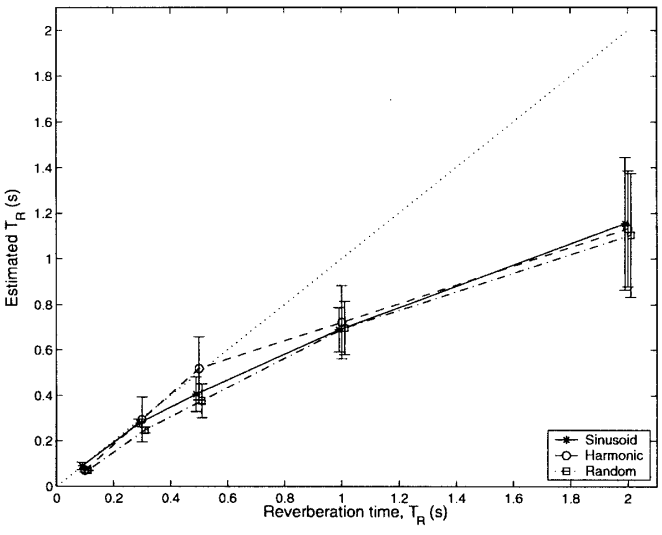

Fig. 7 Estimated reverberation time. The dotted line shows the idealized reverberation time.

Figure 8 shows the improvement in restoration accuracy for the restoration of the power envelope of signals with a long silence (1 s). In these comparisons, correlation and SNR as the evaluation measure are used to show the improvement in restoration accuracy achieved through our model, as follows.

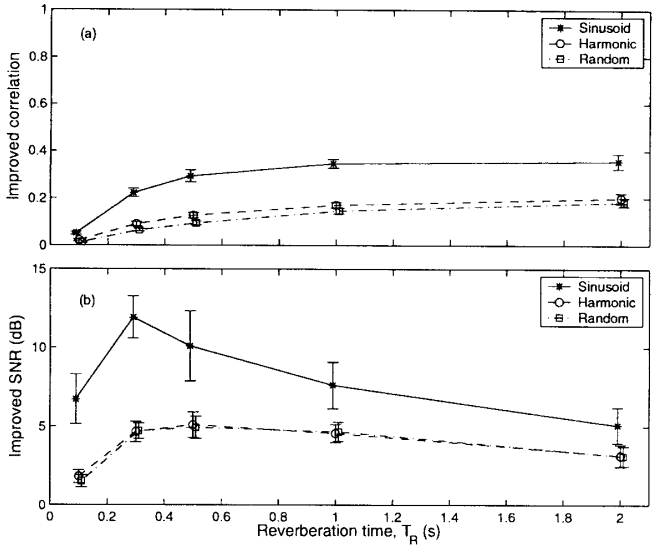

Fig. 8 Improvement in the restoration accuracy: (a) improved correlation and (b) improved SNR (within the silence).

$$\mathrm{Corr}(e_x^2, \hat{e}_x^2) = \frac{\int_0^T \left(e_x(t)^2 - \overline{e_x(t)^2}\right)\left(\hat{e}_x(t)^2 - \overline{\hat{e}_x(t)^2}\right)dt}{\sqrt{\left\{\int_0^T \left(e_x(t)^2 - \overline{e_x(t)^2}\right)^2 dt\right\}\left\{\int_0^T \left(\hat{e}_x(t)^2 - \overline{\hat{e}_x(t)^2}\right)^2 dt\right\}}}, \tag{14}$$

$$\mathrm{SNR}(e_x^2, \hat{e}_x^2) = 20\log_{10}\frac{\int_0^T e_x(t)^2 dt}{\int_0^T \left(e_x(t)^2 - \hat{e}_x(t)^2\right)dt}, \quad \text{(dB)} \tag{15}$$

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. The improvement in correlation is calculated from $\mathrm{Corr}(e_x^2, \hat{e}_x^2) - \mathrm{Corr}(e_x^2, e_y^2)$ and the improvement in SNR is calculated from $\mathrm{SNR}(e_x^2, \hat{e}_x^2) - \mathrm{SNR}(e_x^2, e_y^2)$. 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 [13,14]. Therefore, if the power envelope was restored from a reverberant signal, both measures 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 completely improved.

The improvements in Fig. 8 are all positive values, showing that our alternative method (Eq. (12)) can be used to adequately estimate T_R from the reverberant envelope.

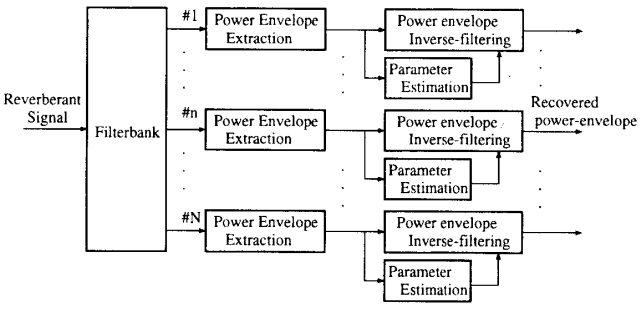

Fig. 9 Power envelope inverse filtering in the constant bandwidth filterbank model. The number of filterbank channels is denoted as N.

Note that this method also works in the general case (e.g. as shown in Fig. 6 (a)).

4. FILTERBANK MODEL

Based on the above considerations, we extended our improved method [13,14] into a filterbank model for speech. Figure 9 shows the architecture of the extended filterbank model. This model was designed as a constant-band filterbank using a FIR-type bandpass filter, the bandwidth of each channel was set to 100 Hz ($N = 100$) and an extraction method using the Hilbert transform relations (Eq. (9)) was used to reduce the computational cost. The blocks where T_R is estimated were processed separately.

We carried out the following simulations to evaluate the proposed model. The speech signals were the same Japanese sentences (three sentences and ten speakers) as described in Sect. 3.1. There were 100 types of impulse response $h(t)$, and five reverberation times: $T_R = 0.1$, 0.3, 0.5, 1.0, and 2.0 s. All stimuli, $y(t)$, were composed through 15,000 ($= 3 \times 10 \times 5 \times 100$) convolutions of $x(t)$ with $h(t)$.

Figure 10 shows the improved correlation and the improved SNR of each channel for the restoration from the speech signals with T_R estimated using Eq. (12). Each $\hat{T}_R$ was separately estimated for each channel. In this figure, the bar height and the error bar show, respectively, the mean and the standard deviation for each result. The improvements in correlation and SNR increased as T_R increased, except for T_R of 0.1.

In contrast, Figure 11 shows the improved correlation and the improved SNR of each channel for the power envelope restoration from the reverberant speech signals with no estimation of T_R when just using the known T_R (original value) in all cases and using constant values in all channels. Again, as in Fig. 10, the improvements in correlation and SNR increased as T_R increased, except for T_R of 0.1. The main difference between these results is the improved SNR at lower frequencies (left side) and higher frequencies (right side). The standard deviations of the

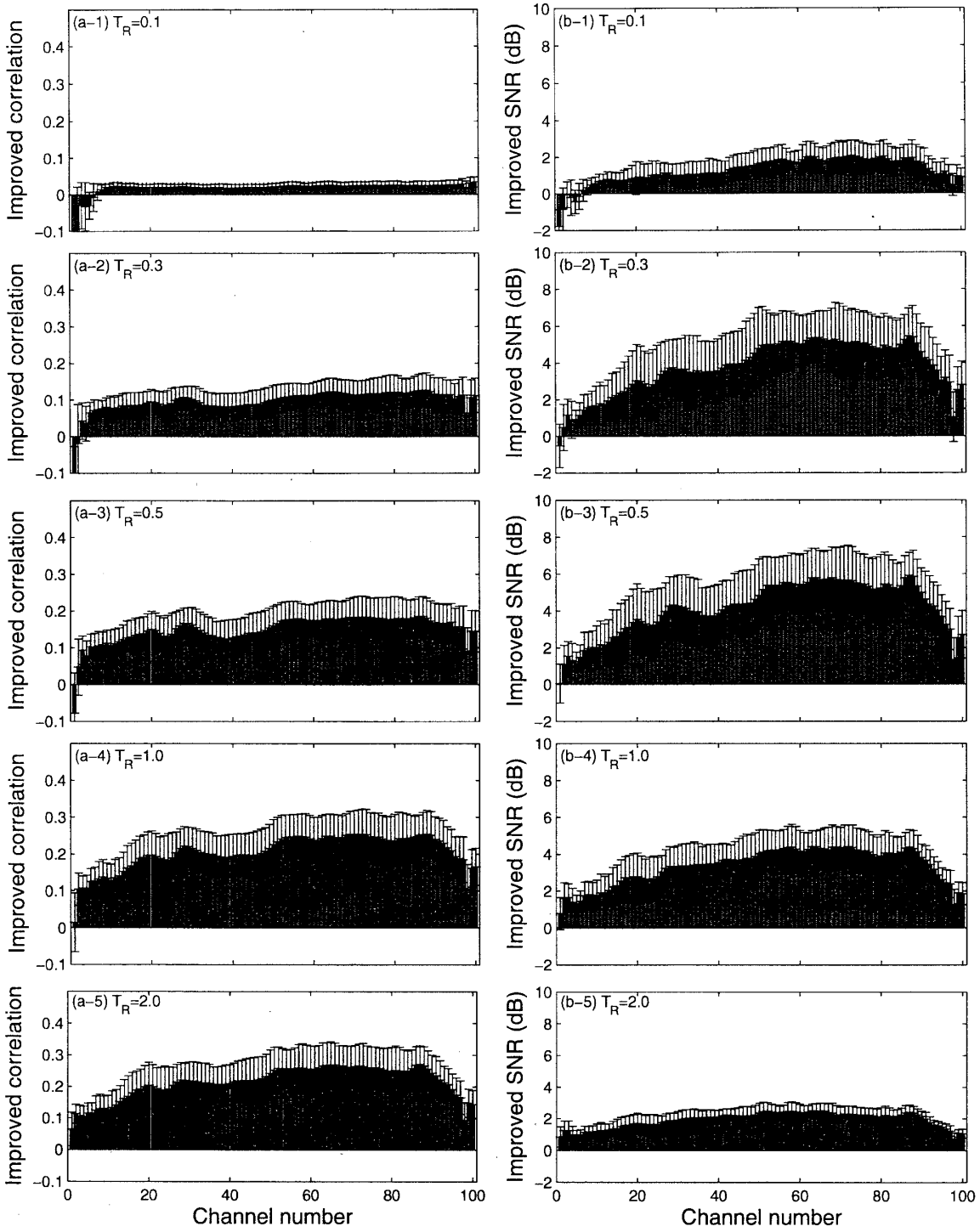

Fig. 10 Improvement in the restoration accuracy for the power envelope of speech on the filterbank: (a) improved correlation and (b) improved SNR (using Eq. (12)).

improvements in Fig. 11 are larger than those in Fig. 10. These results show that Eq. (12) is a reasonable means of estimation for dereverberation and that the estimation should be done separately in each channel.

Figure 12 shows an example of a restoration result obtained using the proposed model for a Japanese sentence (/aikawarazu/) uttered by a male speaker (Mau) (in panel (a)) and the reverberation time of $T_R = 1.0$ s (in panel (b)). The power envelopes of only a quarter of all channels are plotted in this figure (#1, #5, #9, and so on). We can see many matches between the power envelopes of the original and the restored envelopes in panel (d), but there are fewer matches in panel (c). These results demonstrate that the proposed model can be used to adequately restore the power envelope from reverberant speech.

In this case, the improved SNRs for a specific channel were calculated as $\mathrm{SNR}(e_x^2, \hat{e}_x^2) - \mathrm{SNR}(e_x^2, e_y^2)$ over all durations in Figs. 12 (c) and (d). The averaged improved

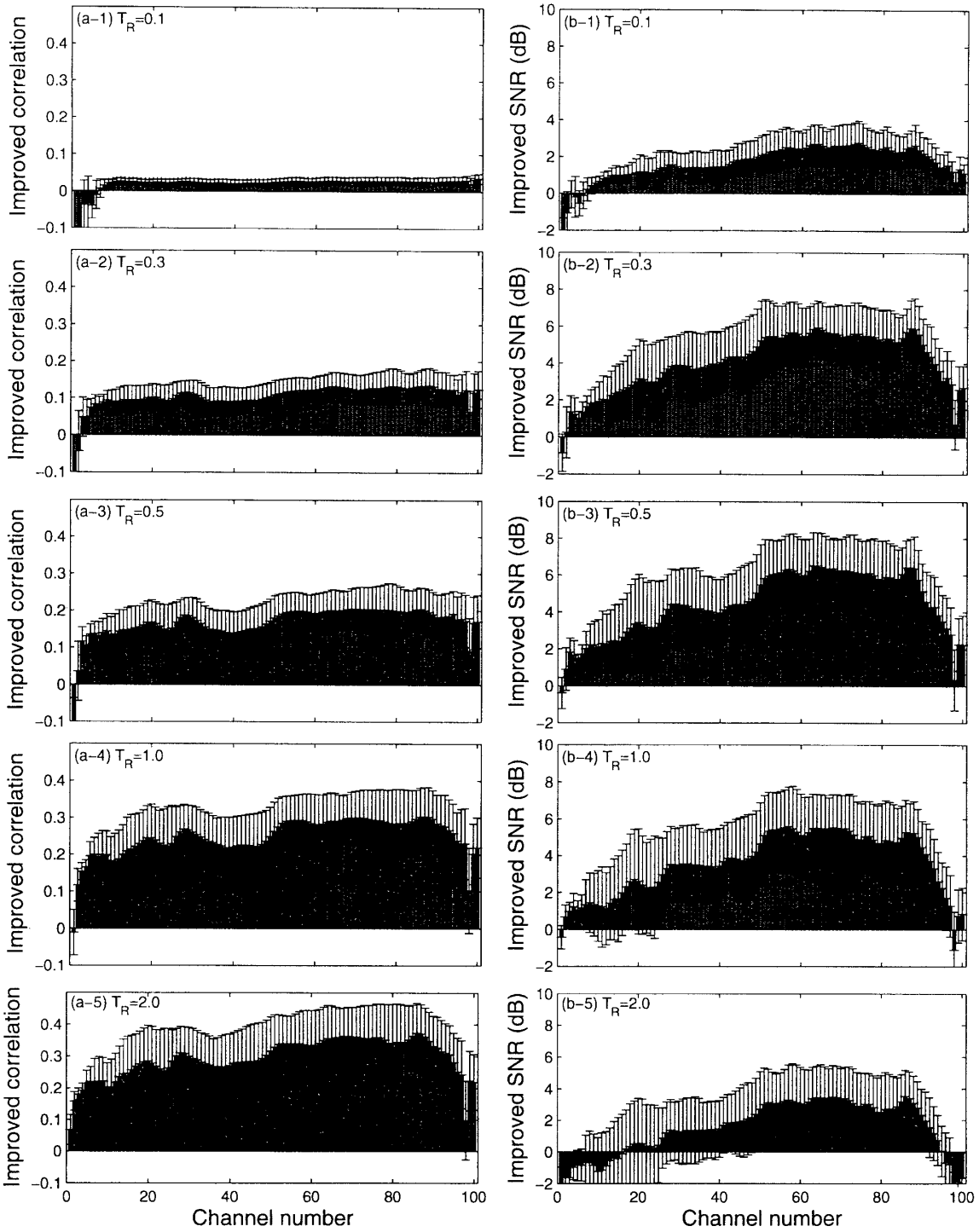

Fig. 11 Improvement in the restoration accuracy for the power envelope of speech on the filterbank: (a) improved correlation and (b) improved SNR (using known T_R (ideal $\hat{T}_R$) and with all values constant in the channels).

SNR with channels was 3.16 dB, and this indicates improvement in the average temporal fluctuation. In contrast, when we measure the root mean squared (rms) difference between the original and the restored power envelopes at a specific t, such as the rms of $10\log_{10}(e_x(t)^2/\hat{e}_x(t)^2)$ and $10\log_{10}(e_x(t)^2/e_y(t)^2)$, the averaged rms with every 50 ms duration by shifting 25 ms for the restored and the reverberant power envelopes were 6.56 dB and 9.17 dB, respectively, in Figs. 12 (c) and (d). Therefore, the reduced rms (improvement) was 2.61 dB. Since both measures indicate improvements in the restored power envelope over duration or over channel, we can interpret this rms as being due to the spectrum distortion (SD) and can estimate the reduction of the averaged SD from the average of the improved SNR.

5. SUMMARY

In this paper, we have proposed a speech dereverberation method based on the MTF concept in power envelope restoration without measuring the impulse response of

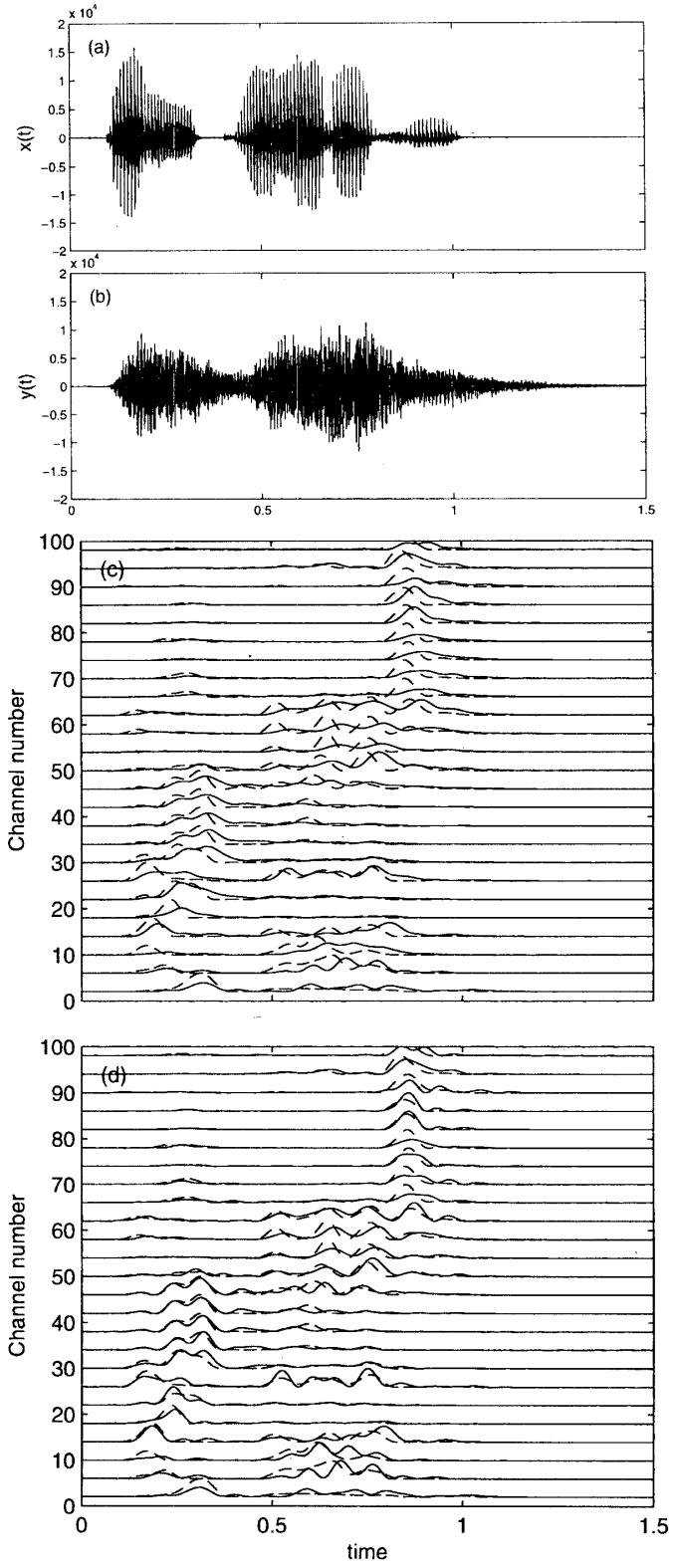

Fig. 12 Simulation results for the reverberant speech: (a) a Japanese sentence (/aikawarazu/) uttered by a male speaker from the ATR database, (b) reverberant speech with $T_R = 1.0$ s, (c) no processing (solid lines), and (d) restoration using the proposed model (solid lines). Dotted lines show the power envelopes of the original speech. $N = 100$.

room acoustics. This method, based on a filterbank, was extended from the improved basic method [13,14] by considering the issues regarding speech applications: (i) how does a reasonable bandwidth depend on the co-modulation characteristics; (ii) how applicable is the MTF concept to reverberant speech and the related bandwidth; and (iii) the usefulness of separately estimating T_R in each channel through another method. We have carried out many simulations in which the proposed model was applied to the power envelope restoration for 15,000 types of reverberant speech signals. We found that the proposed model can be used to adequately restore the power envelopes in a sub-band from reverberant speech. We also showed that another means of estimating T_R (Eq. (12)) is reasonable for the power envelope restoration and that the estimation of T_R should be done in each channel separately.

While the temporal deconvolution methods mentioned in the Introduction can restore the envelope information (the temporal envelope or the modulation index) from the reverberant signal, there is no significant improvement in speech intelligibility. Most existing methods use non-processed phase information or carriers (fine-structure), which are 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.

In our future work, as the next step toward development of blind speech dereverberation, we will (1) reconsider an adaptive power envelope restoration method for time-division and frequency-division processing using a reconstructed filterbank depending on each speech signal, (2) investigate how to restore the carrier or remove the effect of reverberation from a reverberant carrier based on the same filterbank model, and (3) then attempt to solve the problem of dereverberating a signal from a reverberant signal, in waveform, by restoring not only the envelope, but also the carrier from a reverberant signal. We will finally 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. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," *Proc. ICASSP 2003*, Vol. I, pp. 92–95 (2003).
[5] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," *Proc. ICASSP 82*, pp. 156–159 (1982).
[6] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," *Proc. ICSLP 96*, pp. 889–892 (1996).
[7] T. Houtgast, H. J. M. Steeneken, and R. Plomp, "Predicting speech intelligibility in room acoustics," *Acustica*, **46**, 60–72 (1980).
[8] 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).
[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] S. Hirobayashi and T. Yamabuchi, "Validation of blind dereverberation using power envelope inverse filtering and filter banks," *IEICE Trans. A*, **J83-A**, 1029–1033 (2000).
[12] S. Hirobayashi, H. Terashima and T. Yamabuchi, "Evaluation of envelope estimation method of acoustic signal in reverberant field," *J. Jpn. Soc. Simulation Technol.*, **22**, 208–215 (2003).
[13] M. Unoki, M. Furukawa, K. Sakata and M. Akagi, "A method based on the MTF concept for dereverberating the power envelope from the reverberant signal," *Proc. ICASSP 2003*, Vol. I, pp. 840–843 (2003).
[14] M. Unoki, M. Furukawa, K. Sakata and M. Akagi, "An improved method based on the MTF concept for restoring the power envelope from a reverberant signal," *Acoust. Sci. & Tech.* **25**, 232–242 (2004).
[15] A. Papouris, *Probability, Random Variables, and Stochastic Processes*, 3rd Ed. (MacGraw-Hill, Inc., New York, 1991).
[16] 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).
[17] N. Kanedera, T. Arai, H. Hermansky and M. Pavel, "On the importance of various modulation frequencies for speech recognition," *Proc. EuroSpeech 97*, pp. 1079–1082 (1997).
[18] 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).
[19] K. Takeda, Y. Sagisaka, K. Katagiri, M. Abe and H. Kuwabara, "Speech Database User's Manual," *ATR Tech. Rep.*, TR-I-0028 (1988).