A METHOD BASED ON THE MTF CONCEPT FOR DEREVERBERATING THE POWER ENVELOPE FROM THE 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
{unoki, m-furuka, k-sakata, akagi}@jaist.ac.jp

ABSTRACT

This paper proposes a method for dereverberating the power envelope from the reverberant signal. This method is based on the modulation transfer function (MTF) and does not require that the impulse response of an environment be measured. It improves upon the basic model proposed by Hirobayashi et al. regarding 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; and (iii) a lack of consideration as to whether the MTF concept can be applied to a more realistic signal. We have shown that the proposed method can accurately dereverberate the power envelope from the reverberant signal.

1. INTRODUCTION

Recovery of the original signal from a reverberant signal is an important issue concerning speech signal processing such as speech-emphasis for transmission system (speaker to microphone) and preprocessing for robust speech recognition systems.

Inverse filtering methods have been proposed to dereverberate the original signal from the 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]. This method can be applied to room acoustics with non-minimum phase characteristics. However, these methods have to measure the impulse response of the room 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 measured each time these methods are used. This is a drawback with regard to these methods.

On the other hand, Hirobayashi et al. proposed the power envelope inverse filtering method [3]. This method, based on the modulation transfer function (MTF) [4], can be used to recover the power envelope of the original signal from the reverberant signal without measuring the impulse response of the room. However, this method still has some problems concerning applications, as described in Sec. 2.1.

In this paper, we improve a method, also based on the MTF concept, for dereverberating the power envelope from the reverberant signal. Our improved method resolves the above problems and is a step towards evolution of a general method for speech from this basic model.

2. POWER ENVELOPE INVERSE FILTERING

2.1. The MTF concept and the model concept based on the MTF

The concept of the MTF was introduced as a measure in room acoustics for assessing the effect of an enclosure on speech intelligibility [4]. The complex MTF, $\mathbf{m}(\omega)$, is defined as

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

where $h(t)$ is the impulse response of the room [6]. 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, $h(t) = \exp(-6.9t/T_R)n(t)$; the MTF, $m(\omega)$, can be obtained as

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

where $n(t)$ is white noise. T_R is the reverberant time; i.e., the time needed for the power of $h(t)$ to decay to 60 dB [4, 6].

In the model of Hirobayashi et al. [3], the observed signal, the original signal, and the stochastic-idealized impulse response in the room acoustics [4] are assumed to be $y(t)$, $x(t)$, and $h(t)$, respectively, and these are modeled based on the MTF as follows:

$$y(t) = x(t) * h(t), \qquad (3)$$
$$x(t) = e_x(t)n_1(t), \qquad (4)$$
$$h(t) = e_h(t)n_2(t) = a\exp(-6.9t/T_R)n_2(t), \qquad (5)$$

where "*" denotes the convolution operation, $e_x(t)$ and $e_h(t)$ are the envelopes of $x(t)$ and $h(t)$, and $n_1(t)$ and $n_2(t)$ are the mutually independent respective white noise functions, $< n_k(t), n_k(t-\tau) >= \delta(\tau)$. The parameters of the impulse response, a and T_R, are a constant amplitude term and the reverberation time, respectively [3]. In this model, the power envelope of the reverberant signal, $e_y(t)^2$, can be determined as

$$\langle y(t)^2 \rangle = \left\langle \left\{ \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \right\}^2 \right\rangle = \int_{-\infty}^{\infty} e_x(\tau)^2 e_h(t-\tau)^2 d\tau = e_y(t)^2, \qquad (6)$$

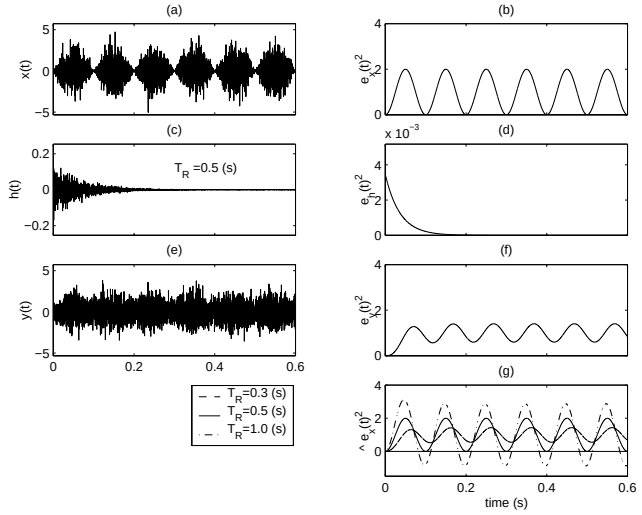

Fig. 1. Example of the relationship between the power envelopes of a system based on the MTF concept.

where $< \cdot >$ is a set-averagin operation [3].

Based on this result, $e_x(t)^2$ can be recovered by deconvoluting $e_y(t)^2$ with $e_h(t)^2$. Here, the transmission functions of power envelopes $E_x(z)$, $E_h(z)$, and $E_y(z)$ are assumed to be the z-transforms of $e_x(t)^2$, $e_h(t)^2$, and $e_y(t)^2$, respectively. Thus, the transmission function of the power envelope of the original signal, $E_x(z)$, can be determined from

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

where f_s is the sampling frequency. Finally, the power envelope $e_x(t)^2$ can be obtained from the inverse z-transform of $E_x(z)$ [3].

Figure 1 shows an example of how the power inverse filtering method is related to the MTF concept. Figure 1 (a) shows the original signal with a sinusoidal power envelope as shown in Fig. 1 (b) (the modulation frequency F_c was 10 Hz and the modulation index was 1). Figure 1 (c) shows the impulse response of Eq. (5) with $T_R = 0.5$, and Fig. 1 (e) shows the observed signal with a convolution of $x(t)$ with $h(t)$. The right panels ((b), (d), and (f)) show the power envelopes of the signals. In this figure, the modulation index decreased from 1 (Fig. 1 (b)) to about 0.29 (Fig. 1 (f)). It can also be shown that the decreased modulation index is derived from $m(2\pi F_c) = 0.31$ using Eq. (2). This means that the MTF concept is to show the modulation index as a function of the modulation frequency (F_c) and the reverberation time (T_R) [4].

The solid line in Fig. 1 (g) shows the power envelope recovered from the observed power envelope (Fig. 1 (f)) through this method with $T_R = 0.5$. If the method is applied with T_R set to an inappropriate value (for example, $T_R = 0.3$ or $T_R = 1.0$), the recovered power envelopes are not precisely dereverberated as the other lines in Fig. 1 (g) show.

2.2. Problems

With this concept, the basic model can dereverberate the power envelope of an original signal from an observed signal if it can detect the power envelope precisely and the parameters of the room impulse response are known before processing, as shown in Fig. 1 (g). However, we still have to overcome the problems associated with the basic model. Here, we consider (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 terms (T_R and a) of the impulse response of room acoustics.

3. IMPROVED METHOD

3.1. Extraction of the power envelope

Extracting the power envelope from an observed signal based on the MTF concept using well-known techniques (such as half-wave rectification (HWR) of the signal demodulation) is difficult because the carrier is white noise rather than a sinusoidal signal.

In this paper, we propose a method using set-averaging as shown in Eq. (6), to extract the power envelope precisely. In this case, the observed signal $y(t)$ does not have a random variable to calculate the set-averaging, so Eq. (6) cannot be directly applied to extract the power envelope from $y(t)$. We assume that a product of each white noise signal becomes the other white noise signal. Based on this assumption, the observed quasi-set of $\hat{\mathbf{y}}(t)$ can be created by producing $y(t)$ with a set of white noise $\hat{\mathbf{n}}(t)$. We can therefore extract the power envelope from $y(t)$ using

$$\hat{e}_y(y)^2 =< \hat{\mathbf{y}}(t)^2 >= \mathbf{LPF}\left[< y(t)\hat{\mathbf{n}}(t)^2 >\right]. \qquad (8)$$

In this equation, we used low-pass filtering (LPF) as post-processing to remove the high-pass 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 and speech recognition is from 1 to 16 Hz [5].

3.2. Determination of the impulse response parameters

In this method, parameter T_R must be precisely determined from the observed signal for dereverberation. Hirobayashi et al. used the known T_R in their model [3], so that their model is restricted for any application. For example, estimating T_R from the relationship between the modulation frequency F_c and the MTF of $|\mathbf{m}(2\pi F_c)|$ using Eq. (2) may be considered. However, it is difficult to precisely determine an optimal T_R from this relationship in Eq. (2) because the frequency component of the power envelope usually does not take a single value (F_c).

In this paper, we consider over- and/or under-dereverberation of the power envelope with T_R as shown in Fig. 1 (g). A matching-condition of the original and recovered power envelope is to recover a modulation index of 1 from the reverberation if the modulation index of the original signal is assumed to be 1. This condition can be satisfied by detecting a timing-point where the maximum dip will be 0 or the negative area of the recovered envelope will be 0. In this paper, we assume that the modulation index of the original signal is set to 1, so T_R can be estimated using

$$\hat{T}_R = \max\left(\underset{T_{R,\min}\le T_R \le T_{R,\max}}{\arg\min}\left\{\int_0^T \min\left(\hat{e}_{x,T_R}(t)^2, 0\right) dt\right\}\right) \qquad (9)$$

where T is a signal duration and $\hat{e}_{x,T_R}(t)^2$ is the set of candidates of the power envelope dereverberated as a function of T_R. Note that the operation of "$\max(\arg\min\{\cdot\})$" means to determine the maximum argument of T_R from a timing point where the negative area of $\hat{e}_{x,T_R}(t)^2$ is approximately equal to zero. Here, $T_{R,\min}$ and $T_{R,\max}$ is the lower limited region and the upper limited region of

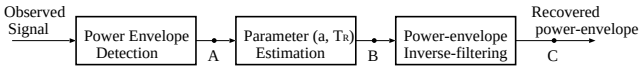

Fig. 2. The power envelope inverse filtering method.

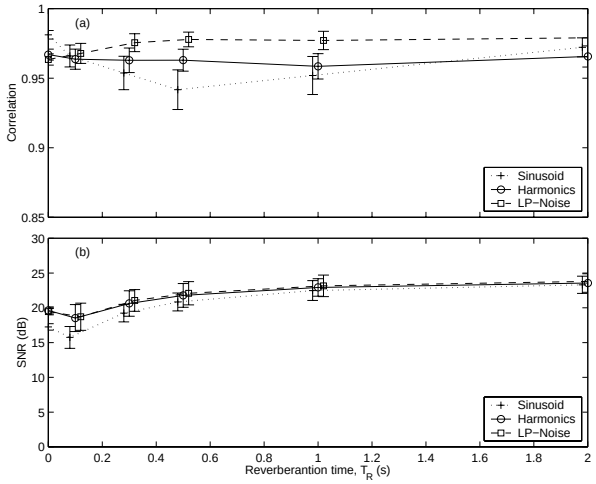

Fig. 3. Extraction accuracy of the power envelope for the three types of stimuli: (a) correlation and (b) SNR.

T_R, respectively. If the original signal has more silences in the signal duration, this assumption is reasonable because more zero-dips will exist in the power envelope.

For example, three candidates of $\hat{e}_{x,T_R}(t)^2$ for $T_R = 0.3, 0.5, 1.0$ were shown in Fig. 1 (g). Here, we assume that $T_{R,\min} = 0.0$ and $T_{R,\max} = 1.0$. One candidate of $\hat{e}_{x,T_R}(t)$, when $T_R = 0.3$, is an under-dereverberation of the power envelope and the other candidate of $\hat{e}_{x,T_R}(t)$, when $T_R = 1.0$, is an over-dereverberation. If we use Eq. (9) to estimate $\hat{T}_R$, we can obtain the optimal $\hat{T}_R$ of 0.5 from three candidates.

Hirobayashi et al. did not describe how to determine parameter a for their model. In our model, however, we find that a is given the same value for both forms of Eq. (5), so determining its value may not be critical problem. Since a is related to the gain of the room acoustics, we assume that the value of a determined from the summarized $e_h(t)^2$ is 1 for applications. In practice, if we try to fit parameter a to various values of $e_h(t)^2$ obtained from many impulse responses, we can set an appropriate value of a for applications in a real environment.

3.3. Evaluation

In this section, we evaluate the improved model. Figure 2 shows a block diagram of the power envelope inverse filtering. The values of $x(t)$ consisted of white noise multiplied by the three types of power envelope:

1. Sinusoidal: $e_x(t)^2 = 1 - \cos(2\pi F t)$;
2. Harmonics: $e_x(t)^2 = 1 + \dfrac{1}{K}\sum_{k=1}^{K} \sin(2\pi k t + \theta_k)$;
3. Band-limited noise: $e_x(t)^2 = \mathbf{LPF}[n(t)]$.

Here, $F = 15$ Hz, $f_s = 20$ kHz, $K = 20$, and θ_k is a random phase. The impulse responses, $h(t)$, consisted of five types of envelope, with $T_R = 0.1, 0.3, 0.5, 1.0, 2.0$, 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 $h(t)$.

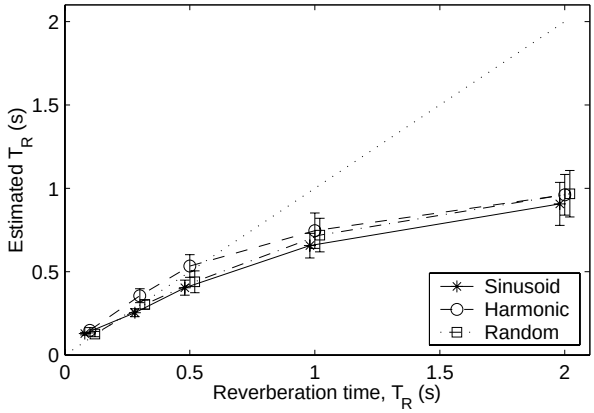

Fig. 4. Estimated reverberation time. The dotted line shows the idealized reverberation time. $T_R = 0.1, 0.3, 0.5, 1.0, 2.0$ s.

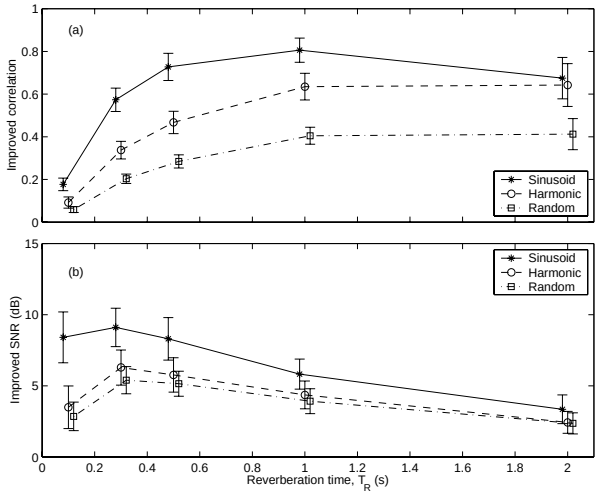

Fig. 5. Improvement of dereverberation accuracy in the proposed model: (a) improved correlation and (b) improved SNR. $T_R = 0.1, 0.3, 0.5, 1.0, 2.0$ s.

As evaluation measures, we used (1) the correlation and (2) the SNR (where S is the original signal and N is the difference between S and the estimated signal) between the original envelope and the extracted/recovered envelope.

Figure 3 shows the extraction accuracy for the power envelopes from all stimuli using the set-averaging method (at point A in Fig. 2). Each point and the error bar show the mean and the standard deviation of the results. We found that the proposed method could precisely extract the power envelope from the observed signal, but the HWR method could not.

Figure 4 shows T_R (at point B in Fig. 2) estimated using the results of Fig. 3. Each point and the error bar show the mean and the standard deviation for $\hat{T}_R$. The dotted line 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.

Figure 5 shows the improvement in accuracy for the dereverberation (at point C in Fig. 2), obtained by plotting the differ-

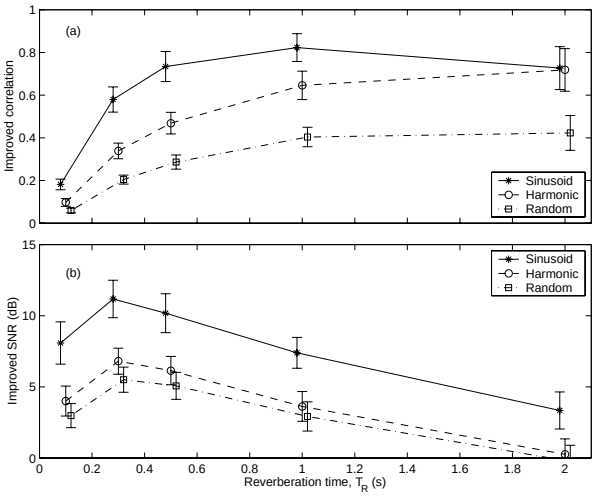

Fig. 6. Improvement of dereverberation accuracy in the proposed model: (a) improved correlation and (b) improved SNR (using ideal T_R, dotted line in Fig. 4). $T_R = 0.1, 0.3, 0.5, 1.0, 2.0$ s.

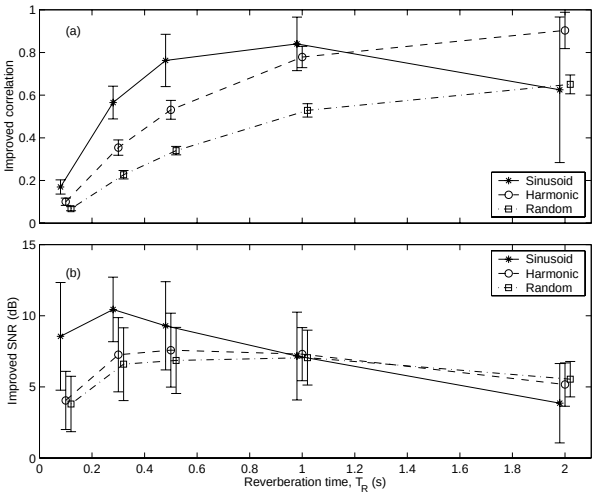

Fig. 7. Improvement of the dereverberation accuracy in the proposed model: (a) improved correlation and (b) improved SNR (Carriers are harmonics). $T_R = 0.1, 0.3, 0.5, 1.0, 2.0$ s.

ences between the recovered power envelope with and without the model. The improvements in Fig. 5 are positive values, demonstrating that the proposed model could effectively dereverberate the power envelope of the signal from the observed signal. We compared these results with the result when T_R was set to a known value (the idealized value). Results from this case are shown in Fig. 6. There were no large differences in the improvements with the proposed model when $\hat{T}_R$ was from 0.0 to about 0.5, although the improvements in the SNR fell by about 1 dB at $T_R = 1.0$ and by about 3 dB at $T_R = 2.0$. Therefore, Eq. (9) can be taken as a reasonable constraint for dereverberation of the power envelope in this model. This means Eq. (9) does not allow over-modulation of the power envelope in this model.

Based on these results, we also considered whether the MTF concept can be applied to realistic signals such as speech. 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 7 shows the results of dereverberation using the proposed model for the same stimuli, except with carriers, as in Fig. 5. The carriers were 100 types of harmonics with F0 of 100 Hz and random phases. We found that the proposed model could dereverberate the power envelope from the observed signal in this case as well as in Fig. 5, although there was a large deviation.

4. CONCLUSION

In this paper, we have developed a method for dereverberating the power envelope from reverberant speech based on the MTF concept without measuring the impulse response of an environment. This method improves upon the method of Hirobayashi et al in two ways: (i) extraction of the power envelope from the observed signal; (ii) determination of the impulse response parameters (a and T_R). We have carried out many simulations in which the proposed model was applied to dereverberation for $1,500$ types of reverberant signals where the carriers were white noise or harmonics. Our results demonstrated that the proposed model can be used to accurately dereverberate the power envelope from a reverberant signal with a white noise carrier as well as with a harmonic carrier.

In our future work, we will (1) extend this model into a filterbank model for speech applications and (2) attempt to solve the problem for dereverberate waveform from the reverberant signal by estimating the carrier from reverberant signal as well as the power envelope.

Acknowledgements

This work was supported by a grant-in-aid for scientific 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).

5. REFERENCES

[1] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response, " J. Acoust. Soc. Am. Vol. 66, No. 1, pp. 165–169, July 1979.

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," IEEE Trans. ASSP, Vol. 36, No. 2, pp. 145–152, Feb. 1988.

[3] 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, Vol. J81-A, No. 10, pp. 1323–1330, Oct. 1998 (in Japanese).

[4] T. Houtgast and H. J. M. Steenken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in audioria," J. Acoust. Soc. Am., Vol. 77, No. 3, pp. 1069–1077, March 1985.

[5] N. Kanedera et al., "On the importance of various modulation frequencies for speech recognition," Proc. EuroSpeech97, pp. 1079-1082, Rhodes, Greece, Sept. 1997.

[6] M. R. Schroeder, "Modulation transfer functions: definition and measurement," Acustica Vol. 49, pp. 179-182, 1981.