16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

MTF-BASED METHOD OF BLIND ESTIMATION OF REVERBERATION TIME IN ROOM ACOUSTICS

Masashi Unoki and Sota Hiramatsu

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 JAPAN
phone: + (81) 761-51-1237, fax: + (81) 761-51-1149, email: {unoki,s0610073}@jaist.ac.jp
web: http://www.jaist.ac.jp/~unoki/

ABSTRACT

This paper proposes a method of blindly estimating the reverberation time based on the concept of the modulation transfer function (MTF). It is used to estimate the reverberation time from the reverberant signal without measuring room impulse response. We incorporated a process for estimating a parameter related to the reverberation time into the method of MTF-based speech dereverberation we previously proposed. We investigated whether the estimation process we then presented worked as a blind method of estimation and found problems with it. We therefore propose a new method of blindly estimating the reverberation time to resolve these problems, where the reverberation time is correctly estimated by inverse-MTF filtering in the modulation frequency domain. We evaluated the new method with the previous approach using both artificial MTF-based signals and speech signals to demonstrate how accurately it could estimate the reverberation time in artificial reverberant environments. The results revealed that it could accurately estimate reverberation times from observed reverberant signals. We also evaluated the new technique in real reverberant environments. The results suggested that it could accurately estimate the reverberation time from observed reverberant speech.

1. INTRODUCTION

Reverberation time is one of the most significant parameters for characterizing room acoustics [1]. Reverberation affects both speech intelligibility and sound localization. Therefore, reverberation time is used as a useful parameter for various speech signal processes such as F_0 estimates from reverberant speech, speech dereverberation, and robust speech recognition in reverberant environments [2, 3, 4, 5, 6].

The reverberation time specifies the duration for which a sound persists after it has been switched off. The persistence of sound is due to the multiple reflections of sound from various surfaces in the room. Thus, the reverberation time is defined as the T_{60} time, which is the time taken for the sound to decay to 60 dB below its value at cessation [1, 7]. This decay curve for the sound energy is precisely calculated using the impulse response in room acoustics [8]. Therefore, stable and accurate methods for measuring the room impulse response (RIR) by bursting balloons, firing gunshots, or the time stretched pulse (TSP) are required to accurately determine the reverberation time [1, 9].

These methods can be used to accurately determine the reverberation time for room acoustics. In practice, they may have problems with use under realistic conditions, such as ambient noise-floor and time-variant conditions due to variations in temperature, humidity, shapes-of-rooms, or moving objects. Prediction and methods of estimating the decay function have been proposed to resolve noise-floor issue. However, it is very difficult to instantaneously measure the RIR and then to simultaneously apply the estimated reverberation time to applications of speech dereverberation or speech recognition in the same situations in reverberant environments. The reverberation time can not only be determined without measuring the RIR under realistic conditions but it can also work on the applications even if the characteristics of the room acoustics are varied.

We therefore incorporated a process for estimating a parameter related to the reverberation time into the MTF-based methods of speech dereverberation we previously proposed [4, 5]. We investigated whether the estimation process worked as a method of blind estimation and found problems with it. Here, we proposed a new method of blind estimation based on the MTF concept to resolve these problems.

2. MTF-BASED POWER ENVELOPE INVERSE FILTERING

2.1 MTF concept

The MTF concept was proposed by Houtgast and Steeneken to account for the relationship between the transfer function in an enclosure in terms of input and output signal envelopes and the characteristics of the enclosure such as reverberation [10]. This concept was introduced as a measure in room acoustics for assessing the effect of the enclosure on speech intelligibility [10, 11, 12]. The complex MTF is defined as

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

where $h(t)$ is the RIR and ω is the radian frequency. A well-known stochastic approximation of the RIR, i.e., artificial reverberant impulse response [13], is defined as

$$h(t) = e_h(t)\mathbf{n}(t) = a\exp(-6.9t/T_R)\mathbf{n}(t), \qquad (2)$$

where $e_h(t)$ is the exponential decay temporal envelope, a is a constant amplitude, T_R is the reverberant time defined as the T_{60}, and $\mathbf{n}(t)$ is the white noise carrier as a random variable (uncorrelated-carrier).

The corresponding MTF, $m(f_m)$, can be obtained as

$$m(f_m) = |M(f_m)| = \left[1 + \left(2\pi f_m \frac{T_R}{13.8}\right)^2\right]^{-1/2}. \qquad (3)$$

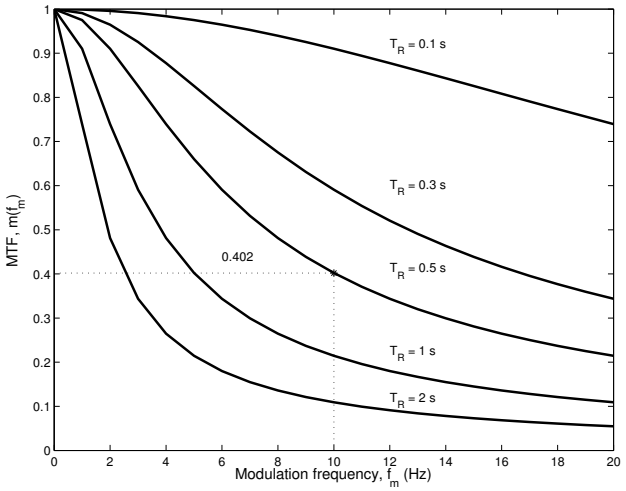

Figure 1: Theoretical curves representing modulation transfer function, $m(f_m)$, under various conditions with $T_R = 0.1$, 0.3, 0.5, 1.0, and 2.0 s.

For a radian modulation frequency $\omega = 2\pi f_m$ of the temporal envelope, Eq. (3) can be regarded as the modulation index, i.e., the degree of relative fluctuation in the normalized amplitude with respect to the modulation frequency f_m. On the basis of this characteristic, T_R can be predicted from a specific modulation frequency by using the MTF. Figure 1 shows the MTF, $m(f_m)$, as a function of T_R. The MTF has characteristics of low-pass filtering as a function of f_m.

2.2 Restoration of power envelope based on MTF

The observed reverberant signal, the original signal, and the stochastic idealized impulse response in room acoustics were assumed to correspond to $\mathbf{y}(t)$, $\mathbf{x}(t)$, and $\mathbf{h}(t)$ in the MTF-based dereverberation model [4, 5]. These can be modeled based on the MTF concept as:

$$\mathbf{y}(\mathbf{t}) = \mathbf{x}(\mathbf{t}) * \mathbf{h}(\mathbf{t}), \quad (4)$$

$$\mathbf{x}(\mathbf{t}) = e_x(t)\mathbf{n}_1(t), \quad (5)$$

$$\mathbf{h}(\mathbf{t}) = e_h(t)\mathbf{n}_2(t), \quad (6)$$

$$e_h(t) = a\exp(-6.9t/T_R), \quad (7)$$

$$\langle \mathbf{n}_k(t)\mathbf{n}_k(t-\tau)\rangle = \delta(\tau). \quad (8)$$

Here, the asterisk "$*$" denotes the operation of convolution and $e_x(t)$ and $e_h(t)$ are the envelopes of $\mathbf{x}(t)$ and $\mathbf{h}(t)$. The $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ indicate respective mutually independent white noise functions.

In this model, $e_y^2(t)$ can be determined as

$$e_y^2(t) = e_x^2(t) * e_h^2(t) \quad (9)$$

due to the independence of $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ [4]. To cope with these signals in a computer simulation, these variables are transformed from a continuous signal to a discrete signal, such as $e_x^2[n]$, $e_h^2[n]$, $e_y^2[n]$, $x[n]$, $h[n]$, and $y[n]$ based on the sampling theorem. Here, n is the number of samples and f_s is the sampling frequency. In this paper, f_s is set to 20 kHz.

The transfer function of the power envelope of the impulse response, $\mathbf{Z}[e_h^2[n]]$, on the modulation frequency domain can be obtained as

$$\mathbf{Z}[e_h^2[n]] = \frac{a^2}{1-\exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}}, \quad (10)$$

where $\mathbf{Z}[\cdot]$ is the z-transformation. Thus, modulation spectrum $\mathbf{Z}[e_x^2[n]]$ can be obtained as

$$\mathbf{Z}[e_x^2[n]] = \frac{\mathbf{Z}[e_y^2[n]]}{\mathbf{Z}[e_h^2[n]]} = \frac{\mathbf{Z}[e_y^2[n]]}{a^2}\left\{1-\exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}\right\}. \quad (11)$$

Since $1/\mathbf{Z}[e_h^2[n]]$ is the inverse filtering of the power envelope of the RIR, this is referred to as inverse MTF. This can be obtained as a 1st order Infinite Impulse Response (IIR) filter.

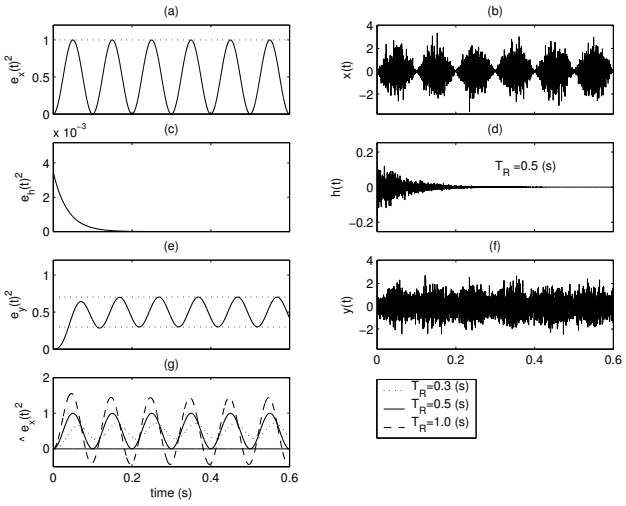

Figure 2: Examples of relationships between power envelopes of system based on MTF concept: (a) power envelope $e_x^2(t)$ of (b) original signal $x(t)$, (c) power envelope $e_h^2(t)$ of (d) impulse response $h(t)$, (e) power envelope $e_y^2(t)$ derived from $e_x^2(t) * e_h^2(t)$, (f) reverberant signal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

Figure 2 shows these modulation relations on the time domain when the original power envelope is sinusoidal (10 Hz). Figures 2(b), (d), and (f) show original signal $x(t)$, impulse response $h(t)$, and reverberant signal $y(t)$. Figures 2(a), (c), and (e) show power envelopes $e_x^2(t)$, $e_h^2(t)$, and $e_y^2(t)$ of all signals. Figure 2(e) shows result of convolution of Figs. 2(a) and (c) at $T_R = 0.5$ s as derived in Eq. (9). Figure 2(g) shows the power envelope restored from Fig. 2(e) by inverse filtering in Eq. (11). When $T_R = 0.5$ s as a parameter of the inverse filter, the restored power envelope is the same as that in Fig. 2(a). In Fig. 1, this restoration was done by inverse filtering at $m(f_m) = 0.402$, where $f_m = 10$ Hz and $T_R = 0.5$ s, to obtain $m(f_m) = 1$. When $T_R = 1.0$ s, the restored power envelope is over modulated.

2.3 T_R estimates and problems

The power envelope, $e_y^2(t)$, in inverse filtering [4] can be extracted using

$$\hat{e}_y^2(t) := \mathrm{LPF}\left[|y(t) + j\mathrm{Hilbert}[y(t)]|^2\right]. \quad (12)$$

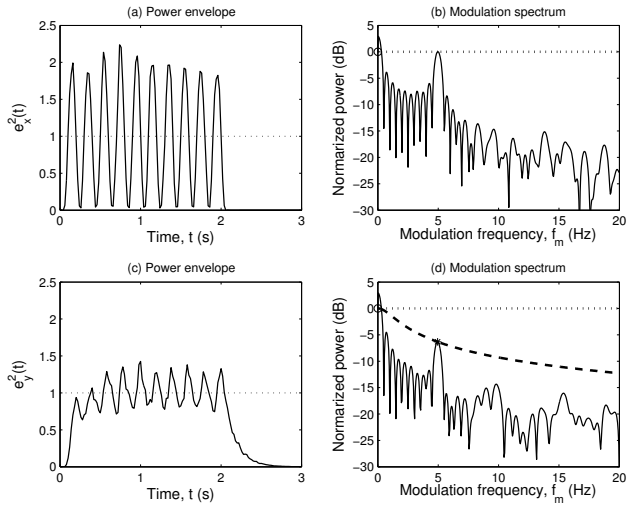

Figure 3: Extracted power envelopes ((a) and (c)) and modulation spectra ((b) and (d)) of reverberant sinusoids.

Here, LPF$[\cdot]$ is low-pass filtering and Hilbert$[\cdot]$ is the Hilbert transform [4]. The LPF cut-off frequency is 20 Hz. In our previous method, T_R could be blindly determined as

$$\hat{T}_R = \max\left(\arg\min_{T_R}\int_0^T |\min(\hat{e}^2_{x,T_R}(t),0)|dt\right), \qquad (13)$$

where T is signal duration and $\hat{e}^2_{x,T_R}(t)$ is the set of candidates of the restored power envelopes as a function of T_R. This equation means that when the biggest dip of the restored power envelope $\hat{e}^2_x(t)$ is 0 in the restoration, $\hat{T}_R$ can be determined. This is because the power envelope does not have a negative value.

In our previous method, $\hat{T}_R$ was an appropriate value for restoring the power envelope; however, we found that $\hat{T}_R$ was less than the value of T_R in the system as T_R increased (this will be described in Sec. 4 in more detail, see dashed lines in Figs. 6 and 7). Therefore, we could not use our previous method as a method of blindly estimating the reverberation time. This problem was caused because when the power envelope was extracted from the reverberant signal by using Eq. (12), the high frequency components were not completely removed from the power envelope after realistic low-pass filtering and they were emphasized by the inverse MTF filter. The dips in the restored power envelope were therefore the sharpest due to these emphasized components. Since Eq. (13) can be used to determine the lowest zero points in the restored power envelope (modulation index of 1), the deepest dips caused the reverberation time to be underestimated.

3. PROPOSED METHOD

3.1 Model concept

Figure 3 shows the power envelopes ((a) and (c)) extracted using Eq. (12) and the modulation spectra ((b) and (d)) of an artificial signal, which has a sinusoidal power envelope (modulation frequency of 5 Hz). Figs. 3(a) and (b) show the non-reverberated originals and Figs. 3(c) and (d) show them at $T_R = 2.0$ s. Both modulation spectra at 0 Hz (DC, (b) and (d)) are the same so that the MTF at 0 Hz is 0 dB (denoted by "○"; **characteristic 1**). The original modulation spectrum at 5 Hz is the same as that at 0 Hz (**characteristic 2**). As shown in Figs. 3(b) and (d), we found that the entire modulation spectrum of the reverberant signal is reduced as the reverberation time increases, according to the MTF (**characteristic 3**), as shown in Fig. 1. These characteristics were also observed in various f_ms (≤ 20 Hz).

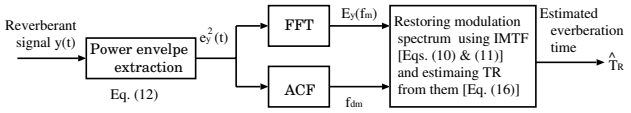

Figure 4: Block diagram for estimating reverberation time.

These useful characteristics enabled us to model a strategy for blindly estimating the reverberation time from the observed reverberant signal. This meant that a specific reverberation time could be determined by compensating for the reduced modulation spectrum at a dominant modulation frequency, f_{dm}, (e.g., $f_{dm} = 5$ Hz in Fig. 3) based on the MTF being 0 dB ($m(f_m)$ was restored to 1). While our previous method dealt with restoring the power envelope of the signal to a modulation index of 1 from the reduced index in the time domain, as this new method dealt with the modulation spectrum related to the modulation index of the power envelope in the modulation frequency domain, it should be possible to stably and accurately estimate the reverberation time.

3.2 Proposed method of estimation

Based on the model concept in Subsection 3.1, we propose a method of blindly estimating the reverberation time in the modulation frequency domain. This is since it can be used to manipulate the dominant modulation frequency component of the power envelope in this domain. We assumed that

$$\log|E_x(f_{dm})| = \log|E_x(0)|, \qquad (14)$$
$$\log|E_y(0)| = \log|E_x(0)| \qquad (15)$$

in the proposed method, based on the two useful characteristics, where $E_x(f_m)$ is the modulation spectrum of $e^2_x(t)$ and $E_y(f_m)$ is that of $e^2_y(t)$. Eqs. (14) and (15) were represented as **characteristics 1 and 2**. Although these were our initial assumptions in the proposed method, we also found that they were useful characteristics in practice. Based on these, the estimated reverberation time, $\hat{T}_R$, can be obtained from the reduced spectrum and the MTF, based on **characteristic 3**:

$$\hat{T}_R = \arg\min_{T_R}\left(\left|\log|E_y(f_{dm})| - \log|E_y(0)| - \log\hat{m}(f_{dm},T_R)\right|\right), \qquad (16)$$

where $\log|E_y(f_{dm})| - \log|E_y(0)|$ is the reduced modulation spectrum at specific f_{dm} and $\hat{m}(f_{dm},T_R)$ is the derived MTF at specific f_{dm} as a function of T_R. This equation means to determine T_R in which $m(f_{dm})$ could be restored to 1.

Figure 4 is block diagram for blindly estimating T_R using Eq. (16). Here, FFT is the fast Fourier transform and ACF is the auto-correlation function. ACF was used for $e^2_y(t)$ in the time domain to determine the dominant frequency, f_{dm}, in the modulation spectrum, $E_y(f_m)$. In restoring the power envelope, the inverse MTF filter in Eqs. (10) and (11) was used to derive $\hat{m}(f_{dm},T_R)$.

For example, the dashed line in Fig. 3(d) indicates the MTF at the $\hat{T}_R$, derived with the proposed method. Figures

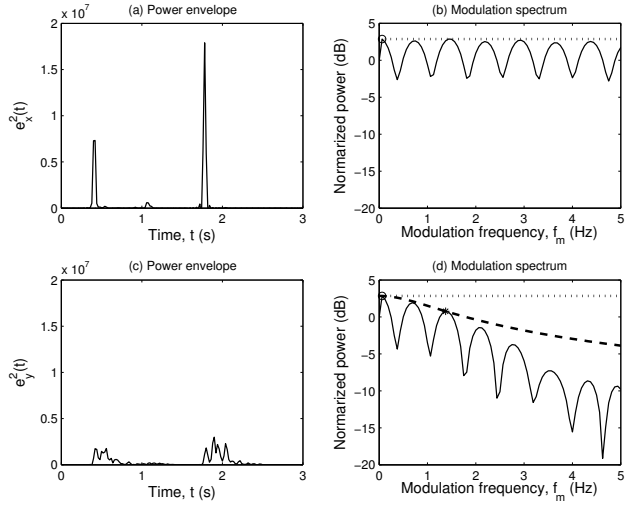

Figure 5: Extracted power envelopes ((a) and (c)) and modulation spectra ((b) and (d)) of reverberant speech.

5(a) and (c) show the power envelopes and (b) and (d) show the modulation spectra of a band limited speech signal. The format for Fig. 5 is the same as that for Fig. 3. In the power envelope in Fig. 3(a), its modulation spectrum at the dominant frequency (f_{dm} Hz) is the same as that at near 0 Hz (f_L Hz). The power envelopes as shown in Fig. 5 can often be found in band-limited speech signals.

4. EVALUATION

4.1 Test: artificial environments

In this section, we discuss our evaluations of the proposed method using reverberant speech signals to confirm whether it worked on blind estimates based on our basic concept. We used the 100 artificial impulse responses ($h(t)$s in Eq. (6)), five reverberation times ($T_R = 0.1, 0.3, 0.5, 1.0$, and 2.0 s) for the artificial signal, $x(t)$, whose power envelope is in Fig. 3(a) and eight speech signals ($x(t)$s) in the evaluation, which were Japanese sentences uttered by a female speaker [14]. All speech signals were decomposed using constant bandwidth filterbank (100-Hz bandwidth and 100-channels). The power envelope had to have restrictions to enable our model concept to be applied to speech signals. All channels we used in the evaluations were chosen beforehand. All reverberant signals, $y(t)$, were obtained through 500 ($= 100 \times 5$ for artificial signals) and $4,000$ ($= 100 \times 5 \times 8$, for speech signals) convolutions of $x(t)$ with $h(t)$.

Figures 6 and 7 plot the estimated reverberation times, $\hat{T}_R$s, from reverberant artificial signals (Fig. 3) and speech signals (Fig. 5). The points represent the means for $\hat{T}_R$s and the error bars represent their standard deviations. The dotted lines indicate the original reverberation time and the dashed lines indicate the reverberation time estimated by the previous method we proposed [4, 5]. In both cases, the $\hat{T}_R$ is underestimated by the previous method as the original T_R increases. $\hat{T}_R$s are matched to the original at all T_Rs in Fig. 6 and from $T_R = 0.3$ to 2.0 s in Fig. 7. In Fig. 7, the standard deviation for $\hat{T}_R$ using the proposed method tends to be reduced when T_R estimates of some channels for reverberant speech signals are used.

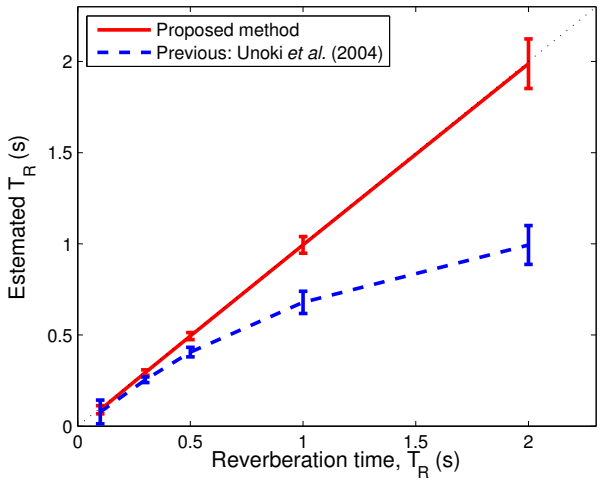

Figure 6: Estimated reverberation time from reverberant sinusoids.

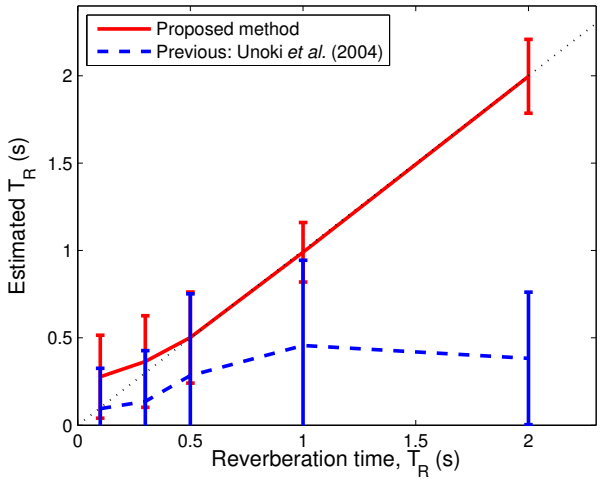

Figure 7: Estimated reverberation time from reverberant speech.

4.2 Applications: real reverberant environments

We tested the proposed method using reverberant speech signals to confirm whether the proposed method worked on blind estimates based on our basic concept in real reverberant environments. We used 17 real impulse responses (IRs) in this test, produced in SMILE datasets [15]. These IRs were measured in a living room in a wooden house ($T_R = 0.36$ s), a movie theater (0.38 s), a meeting room (0.62 s), a church (0.71 s), nine multi-purpose halls (0.80, 1.04, 1.09, 1.35, 1.42, 1.47, 1.54, 1.93, and 2.16 s), a theater hall (0.85 s), and three classic concert halls (1.69, 1.77, and 1.96 s). We used the same speech data [14] as in the above evaluations. Figure 8 shows the estimated T_R from 17 reverberant speech signals. In relatively short reverberation cases (≤ 1.0 s), the blindly-estimated T_R results were accurate and stable. However, in relatively long reverberation cases, the blind estimated T_R results were underestimated. Labels A, B, and C indicate the results for IRs in the multipurpose halls (0.71 s), multipurpose halls (1.42 s), and classic concert halls (1.96 s).

Figure 9 shows three IRs and their respective power envelopes, labeled as A, B, and C in Fig. 8. As a result,

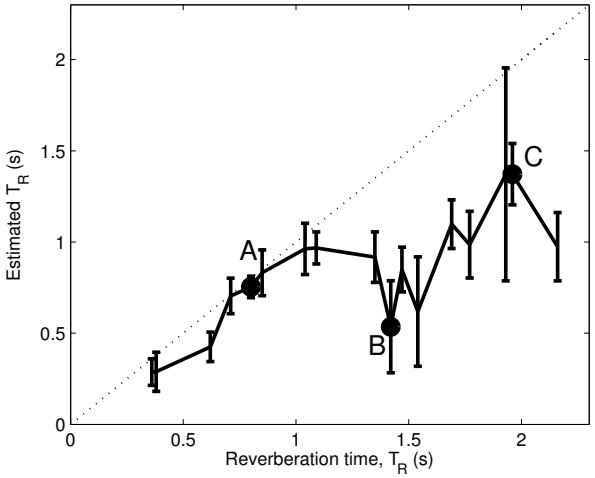

Figure 8: Estimated reverberation time in real reverberant environments. A, B, and C correspond to results for IRs in multipurpose halls (0.80 s), multipurpose halls (1.42 s), and classic concert halls (1.96 s).

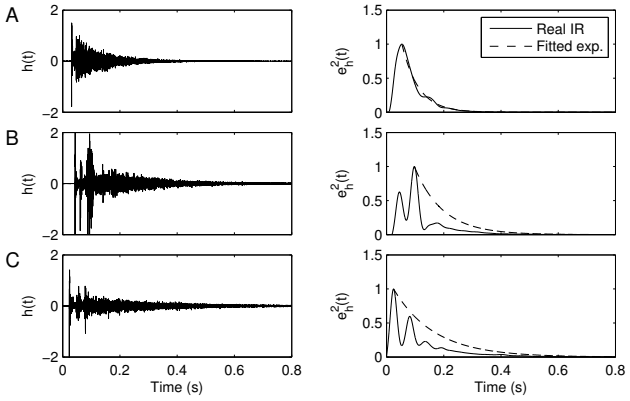

Figure 9: Power envelopes and impulse responses $h(t)$s.

we found that the degree of approximation of the power envelopes of IRs may have affected the accuracy of estimating T_R, especially in case B. Although this suggests that the proposed method can be used to blindly estimate T_R in real environments as a first approximation, but this model should be improved especially in the case of B.

5. CONCLUSION

This paper proposed a method of blindly estimating the reverberation time from observed speech signals based on the MTF concept. We identified problems with the method of estimating T_R we previously presented in MTF-based speech dereverberation. This was because inverse MTF filtering amplifies higher frequency components in the power envelope. We proposed a blind method of estimating T_R in the modulation frequency domain. We evaluated the new method with the previous approach using $4{,}000$ reverberant speech signals. The results revealed that it could correctly estimate the reverberation times from observed reverberant signals. Additional results from real applications demonstrated that the proposed method was effective for blindly estimating the reverberation time in real reverberant environments; however, these also suggested that room acoustics should be improved more with non-exponential decay IRs.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (No. 18680017) from the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

REFERENCES

[1] H. Kuttruff, *Room Acoustics*, 3rd ed. (Elsevier Science Publishers Ltd., Lindin), 1991.

[2] M. Unoki and T. Hosorogiya, "Estimation of fundamental frequency of reverberant speech by utilizing complex cepstrum analysis," *J. Signal Processing*, **12**(1), 31–44, 2008.

[3] M. Unoki, M. Toi, and M. Akagi, "Development of the MTF-based speech dereverberation method using adaptive time-frequency division," Proc. Forum Acusticum 2005, 51–56, Budapest, Hungary, 2005.

[4] M. Unoki, M. Fukai, K. Sakata, and M. Akagi, "An improvement method based on the MTF concept for restoring the power envelope from a reverberant signal," *Acoust. Sci. & Tech.*, **25**(4), 232–242, 2004.

[5] M. Unoki, K. Sakata, M. Furukawa, and M. Akagi, "A speech dereverberation method based on the MTF concept in power envelope restoration," *Acoust. Sci. & Tech.*,**25**(4), 243–254, 2004.

[6] X. Lu, M. Unoki, and M. Akagi, "Comparative evaluation of MTF-based feature extraction for speech recognition in reverberant environments," *SPECOM2007* ! **1**, 124–133, Moscow, Russia, 2007.

[7] ISO 3382, *Acoustics–Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters,* 2nd ed. (International Organization for Standardization, Géneve), 1997.

[8] M. R. Schroeder, "New Method of Measuring Reverberation Time," *J. Acoust. Soc. Am.*, **37**(6), 1187–1188, 1965.

[9] J. Ohga, Y. Yamasaki, and Y. Kaneda, *Acoustic Systems and Digital Processing for Them*, IEICE, Tokyo, 1995.

[10] 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.

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

[12] 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**(3), 1069–1077, 1985.

[13] M. R. Schroeder, "Modulation transfer function: definition and measurement," *Acustica*, **49**, 179–182, 1981.

[14] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe, and H. Kuwabara, *Speech Database*, ATR Interpreting Telephony Research Laboratories, Kyoto, 1988

[15] SMILE2004, Sound Material in Living Environment, Architectual Institute of Japan and GIHODO SHUPPAN Co., Ltd., 2004.