euronoise

Blind estimation method of reverberation time based on concept of modulation transfer function

Masashi Unoki and Sota Hiramatsu

JAIST, 1-1 Asahidai, 923-1292 Nomi, Japan
unoki@jaist.ac.jp

Abstract

This paper proposes a method of blindly estimating the reverberation time (RT) based on the concept of the modulation transfer function (MTF). It is used to estimate the RT from the reverberant signal without measuring room impulse response (RIR). We incorporated a process for estimating a parameter related to the RT 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 the RT estimation and found problems with it. We therefore propose a new method of blindly estimating the RT to resolve these problems, where the RT 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 RT in artificial reverberant environments. The results revealed that it could accurately estimate RTs from observed reverberant signals. The results suggested that it could accurately estimate the RT from observed reverberant signals.

1 Introduction

Reverberation time (RT) is one of the most significant parameters for characterizing room acoustics [1]. Reverberation affects both speech intelligibility and sound localization. Therefore, RT 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, 7, 8].

The RT 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 RT 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, 9]. This decay curve for the sound energy is precisely calculated using the room impulse responses (RIRs) [10]. Therefore, stable and accurate methods for measuring the RIRs by bursting balloons, firing gunshots, or the time stretched pulse (TSP) are required to accurately determine the RT [1, 11].

These methods can be used to accurately determine the RT 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 [1]. 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 RIRs and then to simultaneously apply the estimated RT to applications of speech dereverberation or speech recognition in the same situations in reverberant environments. The RT can not only be determined without measuring the RIRs 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 RT into the MTF-based methods of speech dereverberation we previously proposed [5, 6]. We investigated whether the estimation process we then presented 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 [12]. This concept was introduced as a measure in room acoustics for assessing the effect of the enclosure on speech intelligibility [12, 13, 14]. The complex MTF is defined as

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

where $h(t)$ is the RIR and ω is the radian frequency. A well-known stochastic approximation of the impulse response (artificial RIR) for room acoustics [15] is defined as

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

where $e_h(t)$ is the exponential decay temporal envelope, a is a constant amplitude, T_R is the reverberant time defined as the time required for the power of $h(t)$ to decay by 60 dB, 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

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

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 RIR were assumed to correspond

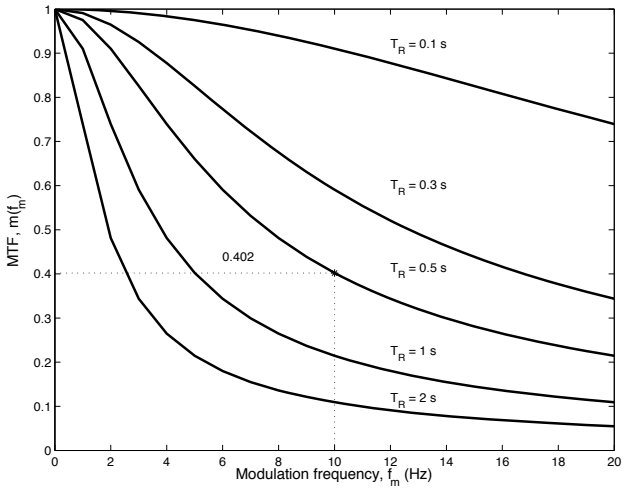

Figure 1: Theoretical curves representing modulation transfer function, $m(f_m)$, under various conditions with $T_R = 0.1$, 0.3, 0.5, 1.0, and 2.0 s.

to $\mathbf{y}(t)$, $\mathbf{x}(t)$, and $\mathbf{h}(t)$ in the MTF-based dereverberation model [5, 6]. These can be modeled as:

$$\begin{aligned} \mathbf{y}(\mathbf{t}) &= \mathbf{x}(\mathbf{t}) * \mathbf{h}(\mathbf{t}), && (4)\\ \mathbf{x}(\mathbf{t}) &= e_x(t)\mathbf{n}_1(t), && (5)\\ \mathbf{h}(\mathbf{t}) &= e_h(t)\mathbf{n}_2(t), && (6)\\ e_h(t) &= a\exp(-6.9t/T_R), && (7)\\ & \langle \mathbf{n}_k(t)\mathbf{n}_k(t-\tau)\rangle = \delta(\tau). && (8) \end{aligned}$$

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 random variables (white noise functions).

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

$$e_y^2(t) = e_x^2(t) * e_h^2(t) \tag{9}$$

due to the independence of $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ [5]. 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

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

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

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

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

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 $x(t)$, reverberant signal $y(t)$, and impulse response $h(t)$. Figures 2(a), (c), and (e) show power envelopes $e_x^2(t)$, $e_y^2(t)$, and $e_h^2(t)$ 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. 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 [5] can be extracted using

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

Here, LPF$[\cdot]$ is low-pass filtering operator, and Hilbert$[\cdot]$ is the Hilbert transform [5]. We set the cut-off frequency of the LPF to 20 Hz to retain most of the important modulation information for speech perception [5, 6].

In our previous method, T_R could be blindly determined as

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

where T is the signal duration. The RT is constrained as $T_{R,\min} < T_R < T_{R,\max}$. These are the lower and upper bounds of T_R (in this paper, $0 < T_R < 3$ s was used). This equation means that when the biggest dip of the restored power envelope $\hat{e}_x^2(t)$ 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, 7), and 8. Therefore, we could not use our previous method as a method of blindly estimating the RT. 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 RT 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

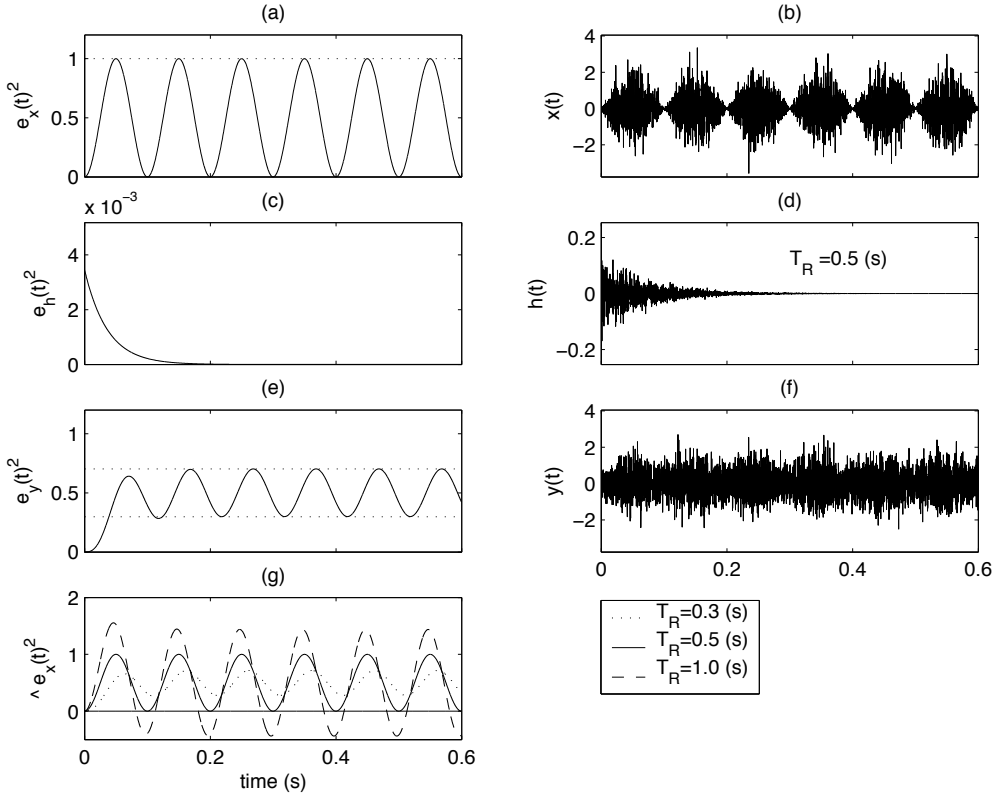

Figure 2: Examples of relationships between power envelopes of system based on MTF concept: (a) power envelope $e_x^2(t)$ of (b) original signal $x(t)$, (c) power envelope $e_h^2(t)$ of (d) impulse response $h(t)$, (e) power envelope $e_y^2(t)$ derived from $e_x^2(t) * e_h^2(t)$, (f) reverberant signal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

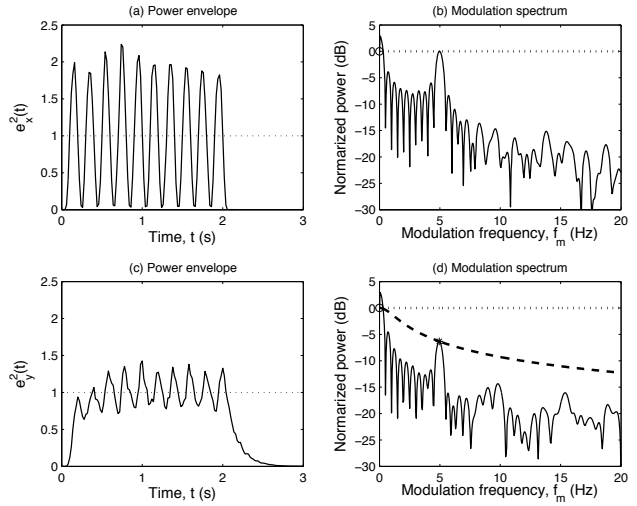

Figure 3: Extracted power envelopes ((a) and (c)) and modulation spectra ((b) and (d)) of reverberant sinusoids.

modulation spectra at 0 Hz (DC, (b) and (d)) are the same so that the MTF at 0 Hz is 0 dB. The original modulation spectrum at 5 Hz is the same as that at 0 Hz. As shown in Figs. 3(b) and (d), we found that the entire modulation spectrum of the reverberant signal is reduced as the RT increases, according to the MTF, as shown in Fig. 1.

These useful characteristics enabled us to model a strategy for blindly the estimating RT from the observed reverberant signal. This meant that a specific RT could be determined by compensating for the reduced modulation spectrum at a dominant frequency based on the MTF being 0 dB (the modulation index 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 RT.

3.2 Proposed method of estimation

Based on the model concept, we propose a method of blindly estimating the RT 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

$$\log E_y(0) = \log E_x(0), \tag{14}$$

$$\log E_x(f_{dm}) = \log E_x(0) \tag{15}$$

in the proposed method, where $E_x(f_m)$ is the modulation spectrum of $e_x^2(t)$ and $E_y(f_m)$ is that of $e_y^2(t)$. The f_{dm} is the dominant modulation frequency (e.g., $f_{dm} = 5$ Hz in Fig. 3). 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 RT, $\hat{T}_R$, can be obtained from the reduced spectrum and the MTF:

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

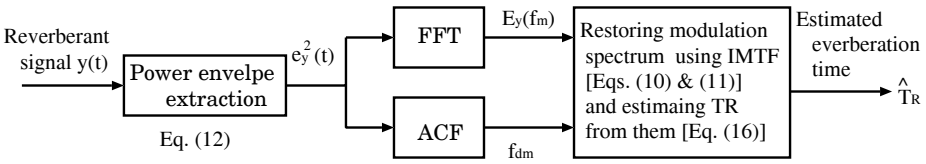

Figure 4: Block diagram for estimating reverberation time.

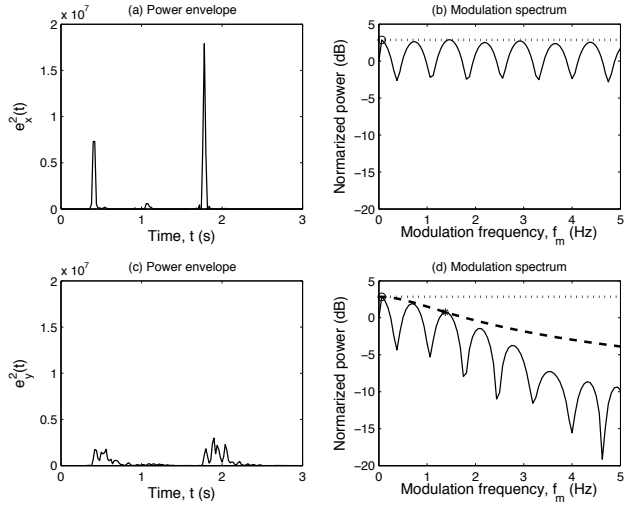

Figure 5: Extracted power envelopes ((a) and (c)) and modulation spectra ((b) and (d)) of reverberant speech.

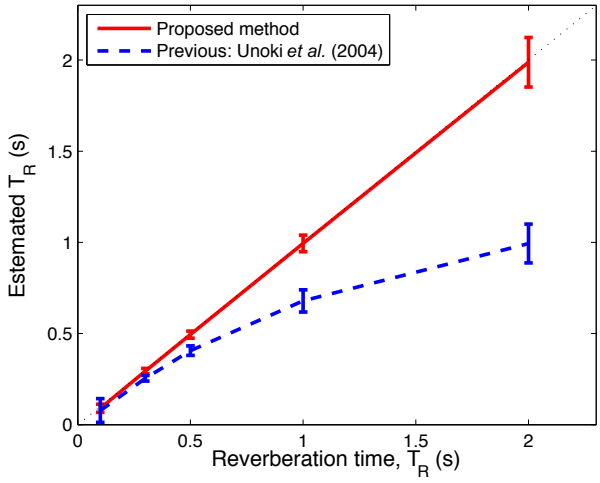

Figure 6: Estimated reverberation time from reverberant sinusoids.

where $\hat{m}(f_m, T_R)$ is the derived MTF at specific f_m as a function of T_R.

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 $e_y^2(t)$ 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 dotted lines in Figs. 3(a)-(d) indicate the MTF at the $\hat{T}_R$, derived with the proposed method. Figures 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

In this section, we discuss our evaluations of the proposed method using reverberant (artificial) sinusoidal and speech signals to confirm whether it worked on blind estimates based on our basic concept. We used the 100 artificial RIRs ($h(t)$s in Eq. (6)), five RTs (T_R = 0.1, 0.3, 0.5, 1.0, and 2.0 s) for the artificial signal (sinusoid of 5 Hz), $x(t)$, 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 [16]. 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 (for artificial signals) and 4,000 ($= 100 \times 5 \times 8$, for speech signals) convolutions of $x(t)$ with $h(t)$.

4.1 Test for reverberant sinusoidal signal

Figure 6 plots the estimated RTs, $\hat{T}_R$s, from reverberant (artificial) sinusoidal signals. The points represent the means for $\hat{T}_R s$ and the error bars represent their standard deviations. The dotted lines indicate the original RT and the dashed lines indicate the RT estimated by the previous method we proposed [5, 6]. In this case, the $\hat{T}_R$ is underestimated by the previous method as the original T_R increases. $\hat{T}_R$s are almost completely matched to the original at all T_Rs in Fig. 6.

4.2 Test for reverberant speech signal

Figure 7 plots the estimated RTs, $\hat{T}_R$s, from reverberant speech signals in which the adequate eight-channels for estimating RTs were chosen in advance for each speech signal and then the estimated RT was determined as the averaged of the RTs in these channels. Figure 8 plots the estimated RTs, $\hat{T}_R$s, from the same reverberant speech signals, by using an automatic channel selection method. In this paper, a channel, which two-specific peaks exist in the power envelope, is chosen as an adequate channel for estimating the RTs.

The figure format is the same as Fig. 6. 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

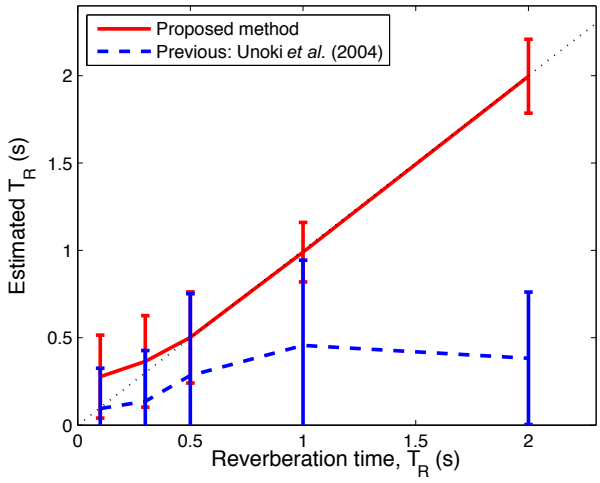

Figure 7: Estimated reverberation time from reverberant speech (fixed channel).

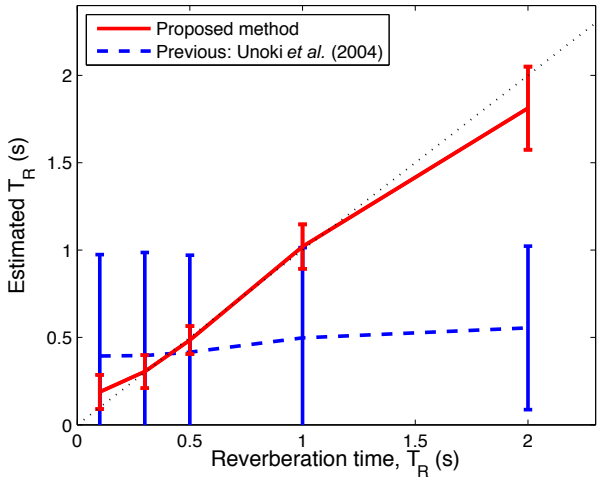

Figure 8: Estimated reverberation time from reverberant speech (with automatic channel selection).

original at all T_Rs in Fig. 7. $\hat{T}_R$s are somewhat mismatched to the original at longer T_R in Fig. 8. In Figs. 7 and 8, 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.

5 Conclusion

This paper proposed a method of blindly estimating the RT 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 RTs from observed reverberant signals.

In the future, we intend to deal with real recorded speech in various reverberant environments by adapting and modifying our MTF-based RT estimation method.

Acknowledgments

This work was partially supported by a Grant-in-Aid for Scientific Research (No. 18680017) from the Ministry of Education, Calture, Sports, Science, and Technology of 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, T. Hosorogiya, and Y. Ishimoto, "Comparative evaluations of robust and accurate F0 estimates in reverberant environments," Proc. *ICASSP2008*, 4569–4572, 2008.

[4] 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.

[5] 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.

[6] 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.

[7] X. Lu, M. Unoki, and M. Akagi, "A robust feature extraction based on the MTF concept for speech recognition in reverberant environment," Proc. *Interspeech2005*, 2546–2549, 2005.

[8] X. Lu, M. Unoki, and M. Akagi, "A comparative evaluation of modulation-transfer-function based blind restoration of the sub-band power envelopes of speech as a front-end processor for automatic speech recognition systems," *Acoust. Sci. & Tech.*, 2008 (in Press).

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

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

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

[12] 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.

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

[14] 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.

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

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