Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **6**, November **2017**

Method of Blindly Estimating Speech Transmission Index in Noisy Reverberant Environments

Masashi Unoki, Akikazu Miyazaki, Shota Morita, and Masato Akagi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
1-1 Asashidai, Nomi, Ishikawa 923–1292, Japan
{unoki, miyazaki.aki, s-morita, akagi}@jaist.ac.jp

Received March 2017; revised May 2017

Abstract. The speech transmission index (STI) is an objective measurement that is used to assess the quality of speech transmission as well as listening difficulty in room acoustics. Blindly estimating STI in real environments is, therefore, an important challenge. The authors previously developed a simplified method for blindly estimating STI on the basis of the concept of the modulation transfer function (MTF). The proposed scheme could be used to estimate STIs from observed reverberant signals in which the room impulse response (RIR) was approximated by Schroeder's model, without measuring the RIRs. There were, however, four remaining issues: whether the method (1) could suitably approximate RIR, (2) was robust against different types of observed signals, (3) was robust against background noise, and (4) could feasibly estimate STI in real environments. This paper extends our previously proposed scheme to resolve these problems by proposing generalized RIR models, by considering the relationship between MTF and modulation spectrum, and by simultaneously estimating their inverse MTFs in noisy reverberant environments. Simulations were carried out to determine whether the proposed method could correctly estimate STIs from the observed speech signals in noisy reverberant environments even if the RIR could not be approximated as Schroeder's model. The results revealed that the proposed approach could be used to effectively estimate STIs from noisy reverberant speech signals even if people were in the room and background noise existed.

1. **Introduction.** The quality of speech transmission must be evaluated to design room acoustics and to diagnose degradation in the sound field, although many subjective experiments need to be conducted to evaluate it and the costs involved are very expensive. Therefore, prediction, objective indices, and measurements of speech transmission in room acoustics are needed to inexpensively assess the quality and intelligibility of speech. Thus, the articulation index (AI), the degree of contribution of early reflections (or early decay time (EDT)), the Deutlichkeit (early to total sound energy ratio: D_{50}), Clarity (early to late arriving sound energy ratio: C_{50}), and other acoustic parameters (e.g., reverberation time (RT): T_{30} and T_{60}) have been used to assess the quality of speech transmissions [1, 2].

The speech transmission index (STI) is a well-known measurement of speech transmission quality in room acoustics [2, 3]. The correspondence between STI and the assessed quality of speech transmission in room acoustics is summarized in Table 1 (see Fig. 4 in Sato *et al.* [4]). The correlation between listening difficulty ratings and STI is the strongest of all tested objective measures [4, 5]. Therefore, STI can be regarded as one of the most significant measurements for assessing the quality level of speech transmission in room acoustics. Methods of calculating STI have been standardized by IEC 60268-16 [3], which is based on the concept of the modulation transfer function (MTF) [6, 7]. This

Table 1. Relationship between speech quality and STI [4].

Quality	Bad	Poor	Fair	Good	Excellent
STI	0.0 ~ 0.3	0.3 ~ 0.45	0.45 ~ 0.6	0.6 ~ 0.74	0.75 ~ 1.0

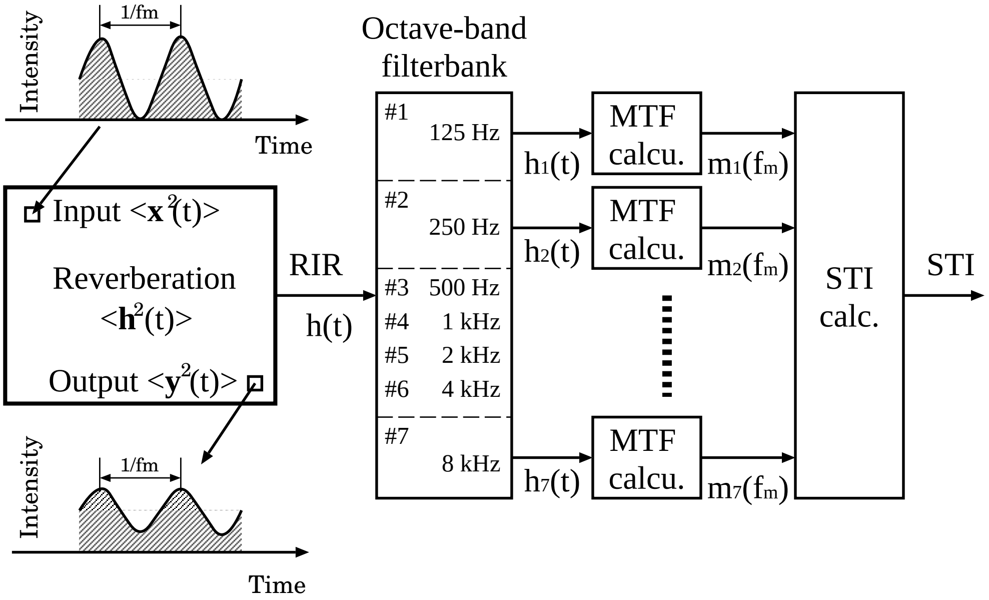

Figure 1. Scheme for STI calculations based on MTF [14].

concept has been an attempt 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 those involving reverberation [6, 7], as shown in Fig. 1.

All objective indices including STI are derived from the characteristics of room impulse responses (RIRs) in assumptions where RIRs have been measured in actual environments that have only low-level background noise and no people. This means that RIRs must be accurately measured to calculate these indices. However, speech transmission generally needs to be assessed in real situations and/or applications such as speech communication and secure announcements in common spaces (e.g., stations, airports, and concourses). Since these measurements must be done in actual environments, these characteristics are quite difficult to obtain by using typical methods of measuring RIRs in sound environments from which people cannot be excluded. In addition, these indices cannot be directly calculated to simultaneously assess the quality of speech transmission in noisy reverberant environments.

There have been a few approaches that can be used to estimate acoustic parameters or objective indices such as the RT, EDT, and C_{50}, from received music and/or speech signals [8, 9, 10, 11]. These approaches have used deep machine learning techniques to estimate these parameters and indices. Although they can accurately estimate these parameters and indices, we need to have massive datasets in real environments to train all of them. It is also very difficult to obtain a corpus of data that include measured RIRs in common spaces from which people cannot be excluded.

We, on the other hand, carried out a preliminary study on the feasibility of blindly estimating the STI in room acoustics on the basis of MTF concept, without measuring RIRs [12]. We previously developed a simplified method of blindly estimating STIs from reverberant signals [13]. This method was used to correctly estimate STI from reverberant amplitude modulation (AM) signals in which RIR was approximated as Schroeder's model of the RIR [15, 16]. The previous results revealed that this method could effectively be used to estimate STIs in artificial reverberant environments. However, four issues remained: whether the method (1) could estimate STIs even if the RIR could not be approximated as Schroeder's model; (2) could not only correctly estimate STIs from reverberant AM but also reverberant speech signals, (3) could estimate STIs from observed signals in noisy reverberant environments; and (4) could estimate STIs from observed signals in real environments where people cannot be excluded.

This paper presents a method for blindly estimating STIs from observed noisy reverberant speech signals. The proposed method involves estimating inverse MTF from the observed signals by the same approach we previously used [12, 13]. The main advantage of our approach is that it enables us to estimate STIs in room acoustics from which people cannot be excluded, without having to measure RIRs or the signal-to-noise ratio (SNR).

2. **Calculation of Speech Transmission Index.** The RIR in IEC 60268-16 [3], is assumed to be a stochastic optimized RIR (Schroeder's RIR [15, 16]):

$$\mathbf{h}(t) = e_h(t)\mathbf{c}_h(t) = a\exp(-6.9t/T_R)\mathbf{c}_h(t), \tag{1}$$

where $\mathbf{c}_h(t)$ is a white noise carrier acting as a random variable and a is a gain factor of RIR. Since the MTF is defined as

$$\mathbf{m}(f_m) = \frac{\int_0^\infty \mathbf{h}^2(t)\exp(-j2\pi f_m t)dt}{\int_0^\infty \mathbf{h}^2(t)dt}, \tag{2}$$

the MTF of the Schroeder's RIR model can be represented as

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

where a is normalized as one. Here, T_R is RT. The MTF, $m(f_m, T_R)$, has characteristics of low-pass filtering as a function of the modulation frequency, f_m, and RT, T_R.

The process of calculating STI can be summarized into five steps (see IEC 60268-16 [3] for details), as outlined in Fig. 1.

(i) Calculating MTFs in seven octave-bands: $m_k(F_i)$, are measured in seven octave-bands (the center frequencies (CFs) range from 125 Hz to 8 kHz and $k = 1, 2, 3, \cdots, 7$). This has fourteen modulation frequencies (the F_i ranges from 0.63 to 12.5 Hz and $i = 1, 2, 3, \cdots, 14$).

$$m_k(F_i) = 1/\sqrt{1 + (2\pi F_i T_R/13.8)^2}. \tag{4}$$

(ii) Calculating SNRs from MTFs: $N(k,i)$ is calculated from $m_k(F_i)$. The $m_k(F_i)$ and $N(k,i)$ are represented as:

$$N(k,i) = 10\log_{10} m_k(F_i)/(1 - m_k(F_i)). \tag{5}$$

(iii) Calculating transmission indices (TIs): TIs, $T(k,i)$, are calculated by normalizing the SNRs, $N(k,i)$, as:

$$T(k,i) = \begin{cases} 1, & (15 < N(k,i)) \\ \frac{N(k,i)+15}{30}, & (-15 \le N(k,i) \le 15) \\ 0, & (N(k,i) < -15) \end{cases} \tag{6}$$

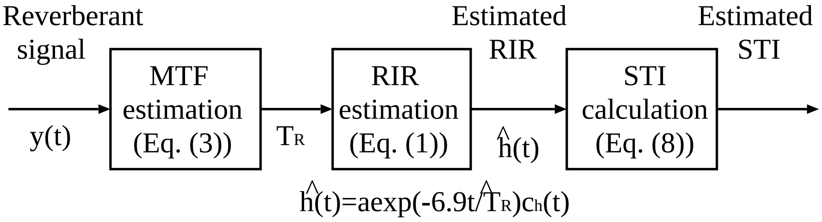

FIGURE 2. Block diagram for previous method of estimating STIs.

(iv) Calculating modulation transmission indices (MTIs): MTIs, $M(k)$, are calculated by averaging $T(k,i)$ as:

$$M(k) = \frac{1}{14}\sum_{i=1}^{14} T(k,i). \tag{7}$$

(v) Calculating STI: Finally, STI is calculated as:

$$\mathrm{STI} = \sum_{k=1}^{7} W(k)M(k). \tag{8}$$

Here, the contribution rates, $W(k)$, are determined to be $W(1) = 0.129$, $W(2) = 0.143$, $W(3) = W(4) = 0.114$, $W(5) = 0.186$, $W(6) = 0.171$, and $W(7) = 0.143$.

3. Previous Method Using Schroeder's RIR Model.

3.1. **Blind estimation of MTF/STI.** In the previous methods, there is assumed to be no background noise. Our previous method used three useful characteristics to estimate MTF: (i) the MTF at 0 Hz was 0 dB, i.e., a modulation index of 1.0, (ii) the original modulation spectrum at the dominant modulation frequency, f_m, was the same as that at 0 Hz, and (iii) the entire modulation spectrum of the reverberant signal was reduced as RT increased in accordance with the MTF. These useful characteristics enabled us to model a strategy to blindly estimate the RT, T_R, from the observed signal, $y(t)$. This meant that a specific T_R could be determined to compensate for the reduced modulation spectrum at a dominant f_m on the basis of the MTF being 0 dB ($m(f_m)$ was restored to 1.0 for all f_ms). Thus, T_R can be determined as

$$\hat{T}_R = \arg\min_{T_R} \left(\left|\log|E_y(f_d)| - \log|E_y(0)| - \log \hat{m}(f_d, T_R)\right|\right), \tag{9}$$

where $\log|E_y(f_d)| - \log|E_y(0)|$ is the reduced modulation spectrum at specific f_d and $\hat{m}(f_d, T_R)$ is the derived MTF at specific f_d as a function of T_R. This equation means T_R is determined as the value at which $m(f_d)$ can be restored to 1.0.

Figure 2 shows a block diagram of the previous method of estimating STI from $y(t)$. This block diagram was developed to adapt speech signals in our preliminary studies [12] in which we found that although the AM-noise signal was suitable for estimating MTFs in the octave-band filterbank, speech signals did not have the same characteristics of whiteness as AM in the bands. The previous method is composed of three blocks: MTF estimation, RIR estimation, and STI calculation.

First, an RT, $\hat{T}_R$, and an MTF, $\hat{m}(f_m, \hat{T}_R)$, are estimated from $y(t)$ by using Eqs. (1) and (3). Then, an RIR, $\hat{h}(t)$, is estimated on the basis of Schroeder's RIR model with $\hat{T}_R$. The $\hat{h}(t)$ is decomposed into seven sub-band components by using the octave-band filterbank. Next, the MTF in each octave-band is calculated from the corresponding observed sub-band signal. Finally, the process described in Section 2 is used to estimate STI from the estimated MTFs.

3.2. **Remaining issues.** The previous method could estimate the MTF/STI without having to measure RIR, where there is no background noise. However, there were four issues remaining from our preliminary studies [12] as to whether the method could (1) estimate STIs even if the RIR could not be approximated as Schroeder's model, (2) estimate STIs from not only reverberant AM but also reverberant speech signals, (3) estimate STIs from observed signals in noisy reverberant environments, and (4) estimate STIs from observed signals in real environments where people could not be excluded.

The STI and $\hat{T}_R$ were frequently estimated incorrect by the previous method, in which the measured RIRs were approximated as Schroeder's RIR model. Issue (1) was caused by mismatches between the temporal envelope of the measured RIRs and its approximation ($\exp(-6.9t/T_R)$). There were a number of corresponding RIRs in which the approximated temporal envelope mismatched that of the measured RIRs, since the corresponding RIRs had onset-transition in the temporal envelope, as can be seen from Fig. 3(a). Since AM signals were used to evaluate the concept of the previous method, issues (2) – (4) have not yet been resolved. To resolve them, general sounds such as speech signals should be used to reconsider these issues.

4. **Proposed Method.**

4.1. **Generalized RIR model.** The previous method assumed that room acoustics could be regarded as reverberant environments without noise and had a diffuse sound field [14]. In addition, Schroeder's RIR model was modified as a generalized RIR model to account for the temporal envelope of the real RIR as [14]:

$$\mathbf{h}(t) = at^{(b-1)}\exp(-6.9t/T_R)\mathbf{c}_h(t), \tag{10}$$

where a is a gain factor of RIR and b is the order of the RIR. This is the same as Schroeder's RIR at $b = 1$. The generalized RIR has greater flexibility than Schroeder's RIR. The MTF of the generalized RIR model is:

$$m(f_m, T_R, b) = \left[1 + \left(2\pi f_m \frac{T_R}{13.8}\right)^2\right]^{-(2b-1)/2}. \tag{11}$$

The difference between the MTFs of Schroeder's RIR and generalized RIR is an exponent of $-(2b-1)/2$.

The temporal envelope and the MTF of RIR models were fitted to those of the measured RIRs to check whether the generalized RIR could correctly approximate the measured RIR. Figure 3 provides results for an example of fitting these characteristics. The root-mean-squared errors (RMSEs) of the temporal power envelopes between the measured RIR and the two models of Schroeder's and the generalized RIRs and the RMSEs of their modulation indices are plotted in these panels. Figure 3(a) indicates that the generalized RIR model could more correctly approximate the temporal envelope of the measured RIR than Schroeder's RIR model. Figure 3(b) also indicates that the MTF of generalized RIR could more correctly represent the MTF of measured RIR than Schroeder's RIR model. This is one of the confirmed results, and the same advantage of the generalized RIR could also be observed in the other RIRs.

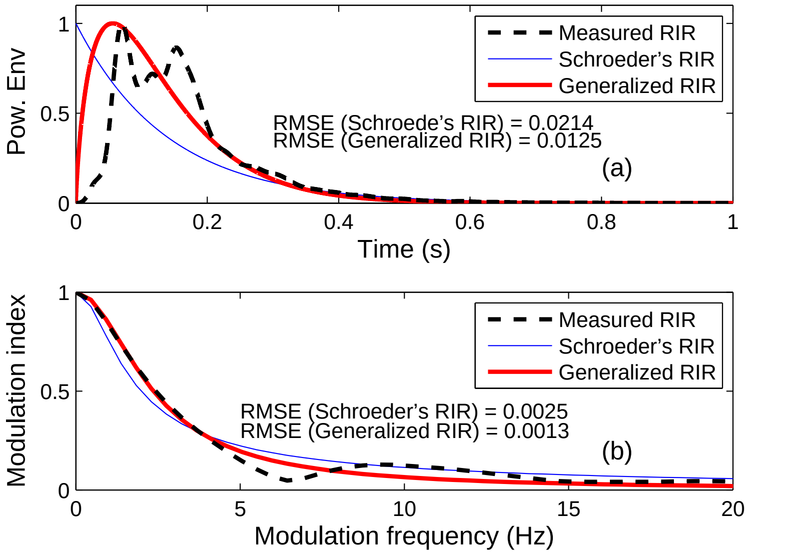

FIGURE 3. Results for fits of RIRs measured with two RIR models: (a) power envelope of RIR and (b) modulation index (MTF) of RIR.

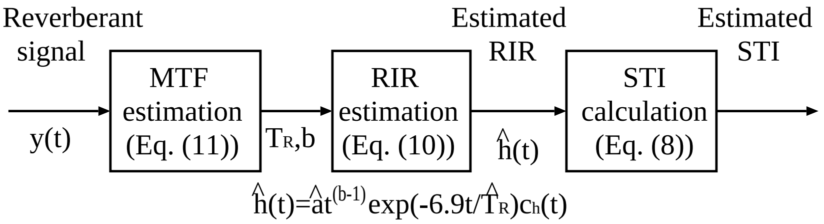
Reverberant signal
y(t)
MTF estimation (Eq. (11))
T_R,b
RIR estimation (Eq. (10))
Estimated RIR
ĥ(t)
STI calculation (Eq. (8))
Estimated STI
ĥ(t)=ât^(b-1)exp(-6.9t/T̂_R)c_h(t)

FIGURE 4. Block diagram for extending previous STI estimation in Fig. 2.

4.2. **Extension to use generalized RIR model.** Figure 4 is a block diagram of the method we have extended for blindly estimating STIs in Fig. 2. This diagram is similar to that for the previous method as shown in Fig. 2, and its main modifications are in the first and second blocks in Fig. 4. Here, the measured RIR is approximated by using Eq. (10) so that the MTF of the measured RIR is approximated by using Eq. (11) [14].

The extended method had three useful characteristics to estimate MTF: (i) MTF at 0 Hz was 0 dB, (ii) the original modulation spectrum at the dominant modulation frequency of f_m was the same as that at 0 Hz, (iii) and the entire modulation spectrum of the reverberant signal was reduced as RT increased in accordance with MTF [14]. These useful characteristics enabled us to model a strategy to blindly estimate the T_R and b of inverse MTF $m^{-1}(f_m)$ that restores the original modulation spectrum from the entire modulation spectrum. The optimal T_R and b were specifically obtained by using the minimum root mean square (RMS). These are defined as:

$$\{\hat{T}_R, \hat{b}\} = \arg\min_{T_R, b} \mathrm{RMS}(T_R, b), \tag{12}$$

$$\mathrm{RMS}(T_R, b) = \sqrt{\frac{1}{L}\sum_{\ell=1}^{L}\left[|E_y(f_{m\ell})| - m(f_{m\ell}, T_R, b)\right]^2}, \tag{13}$$

where $E_y(f_{m\ell})$ is the modulation spectrum of output at specific $f_{m\ell}$ and $m(f_{m\ell}, T_R, n)$ is the derived MTF of the generalized RIR at specific $f_{m\ell}$ as a function of T_R and b. Here,

L is two. Then, an RIR $\mathbf{h}(t)$ is estimated on the basis of the generalized RIR model with T_R and b. Finally, the process described in Section 2 is used to calculate the STI from the estimated MTF.

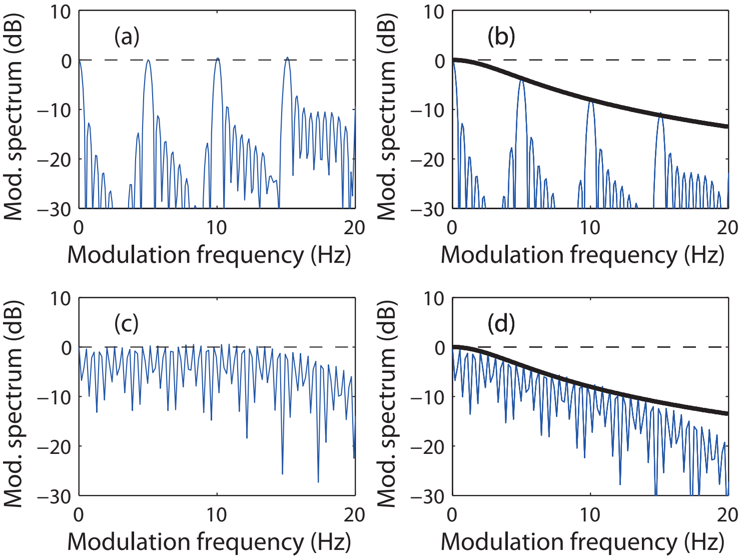

FIGURE 5. Estimated MTFs from reverberant speech signals. Modulation spectra of (a) clean and (b) reverberant AM signal in which power envelope has periodicity. Modulation spectra of (c) clean and (d) reverberant power envelope of speech signal.

Figure 5 (top) plots the relationship between the modulation spectra of the input (original) and output (reverberant) signals that include harmonicity on the modulation spectrum (or periodicity in the power envelope). The solid curve is the MTF, $m(f_m, T_R, b)$, in Eq. (11). The modulation spectrum of input has peaks of 0 dB at the corresponding modulation frequencies, and the corresponding peaks are reduced in accordance with $m(f_m, T_R, b)$. Therefore, $\hat{T}_R$ and $\hat{b}$ are estimated from $y(t)$ by using Eq. (12) when these peaks in Fig. 5(b) are restored to 0 dB. Figure 5 (bottom) plots the same relationship for speech signals so that the proposed method can also determine these two parameters, $\hat{T}_R$ and $\hat{b}$.

4.3. **Extension to gain robustness against background noise.** The previous method studied a method of blindly estimating STI in reverberant environments [14]. Therefore, the previous method could estimate STI without having to measure RIR in reverberant environments. However, there is a critical problem in that the accuracy of the estimated STI was drastically reduced in noisy reverberant environments as there was no modeling effect of background noise.

The proposed method expands the previous method to noisy reverberant environments to resolve these problems. We have already developed a method for restoring an MTF-based power envelope in noisy reverberant environments [17]. The main concept in deriving the inverse MTF with this method can be used to estimate the STI in noisy reverberant environments.

Assume that $\mathbf{x}(t)$, $\mathbf{y}(t)$, $\mathbf{h}(t)$, and $\mathbf{n}(t)$ correspond to the original signal, noisy reverberant signal, RIR, and background noise. The signal is also assumed to be composed of temporal envelope $e(t)$ and carrier $\mathbf{c}(t)$ as random variables of white Gaussian noise. The $e_y^2(t)$ can be represented as $e_y^2(t) = e_x^2(t) * e_h^2(t) + e_n^2(t)$, where the asterisk $(*)$ indicates

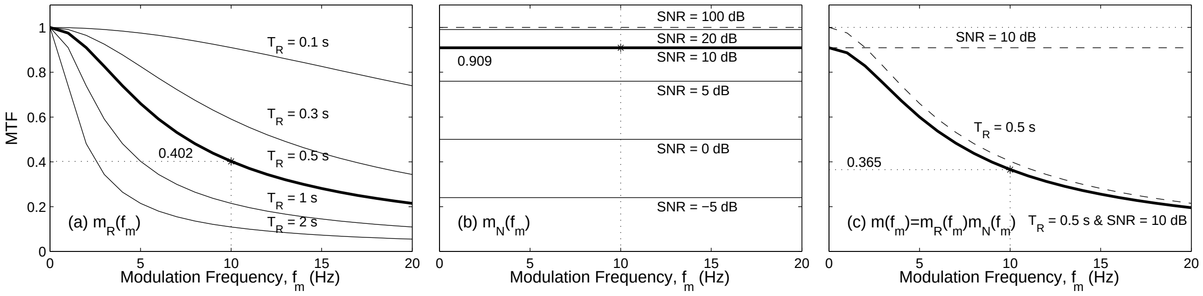

FIGURE 6. Theoretical representations of MTFs, $m(f_m)$, in (a) reverberant environment, (b) noisy environment, and (c) both noisy and reverberant environments. Bold solid lines indicate MTF with $T_R = 0.5$ s and SNR $= 10$ dB.

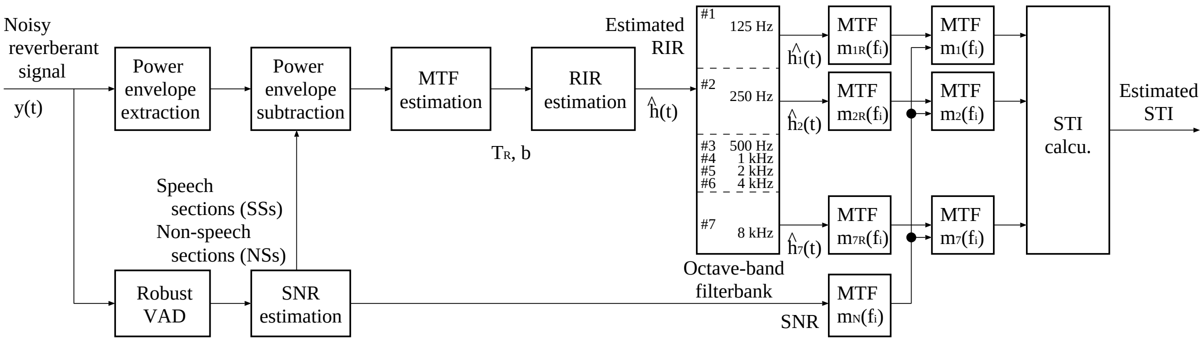

FIGURE 7. Block diagram of proposed method.

convolution by assuming linear systems and mutual independence between carriers. The MTF in a noisy reverberant environment can be represented as [17]:

$$m(f_m, T_R, b, \mathrm{SNR}) = m_R(f_m, T_R, b) \cdot m_N(f_m, \mathrm{SNR}). \tag{14}$$

Here, the MTF in a reverberant environment, $m_R(f_m, T_R, b)$, is defined in Eq. (11) and means the low-pass characteristics as a function of T_R (as shown in Fig. 6(a)). In the case of a T_R of 0.5 s, $m(f_m)$ at $f_m = 10$ Hz is 0.402. The MTF in a noisy environment is defined as $m_N(f_m, \mathrm{SNR}) = 1/(1 + 10^{-\frac{\mathrm{SNR}}{10}})$. This MTF is independent of f_m and reduced as a function of SNR (Fig. 6(b)). In the case of SNR of 10 dB, $m(f_m)$ is 0.909. Therefore, the MTF in a noisy reverberant environment, $m(f_m)$, is defined as:

$$m(f_m, T_R, b, \mathrm{SNR}) = \left[1 + \left(2\pi f_m \frac{T_R}{13.8}\right)^2\right]^{-\frac{(2b-1)}{2}} \left(\frac{1}{1 + 10^{-\frac{\mathrm{SNR}}{10}}}\right). \tag{15}$$

The MTF in noisy reverberant environments depends on f_m and means the low-pass characteristics resulting from reverberation as a function of T_R and the constant attenuation resulting from noise as a function of SNR (Fig. 6(c)). In the case of a T_R of 0.5 s and SNR $= 10$ dB, $m(f_m)$ at $f_m = 10$ Hz is 0.365 ($= 0.402 \times 0.909$). When the previous method was used in noisy reverberant environments, errors in estimation were caused by the effect of MTF in noisy environments (Eq. (15)).

Figure 7 shows a block diagram of the proposed method. The power envelopes of observed signals $e_y^2(t)$ are calculated from observed noisy reverberant signals $y(t)$ as:

$$\hat{e}_y^2(t) = \mathbf{LPF}\left[|y(t) + j \cdot \mathbf{Hilbert}(y(t))|^2\right], \tag{16}$$

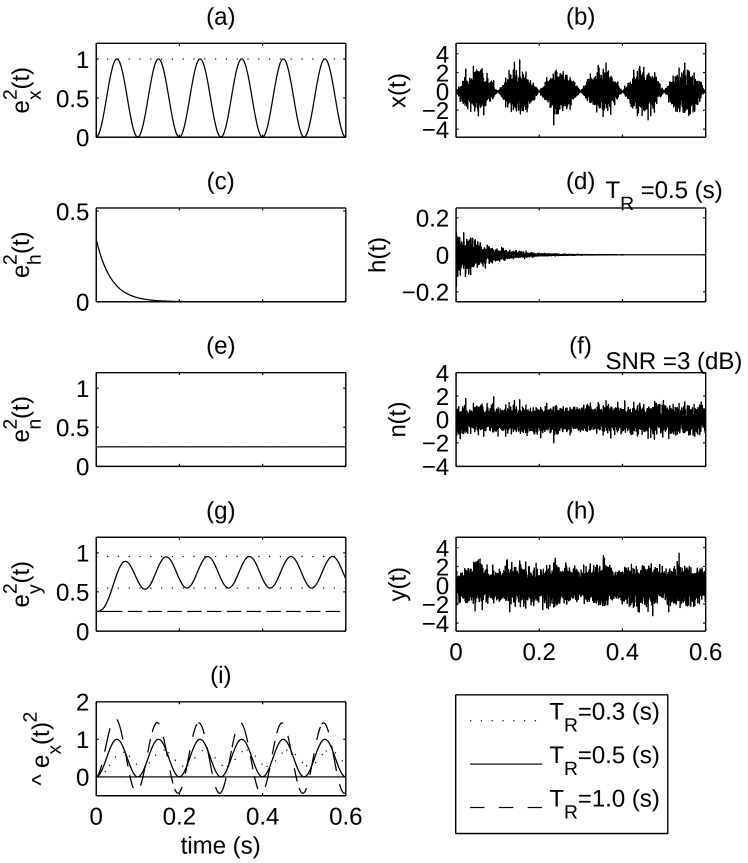

Figure 8. Example of relationship between power envelopes of system based on MTF concept: (a) power envelope $e_x^2(t)$ of (b) original signal $x(t)$, (c) power envelope $e_h^2(t)$ of (d) simulated room impulse response $h(t)$ ($T_R = 0.5$ s), (e) power envelope $e_n^2(t)$ of (f) noise signal $n(t)$, (g) power envelope $e_y^2(t)$ derived from $e_x^2(t) * e_h^2(t) + e_n^2(t)$, (h) noisy reverberant signal $y(t)$ derived from $x(t) * h(t) + n(t)$, and (i) restored power envelope $\hat{e}_x^2(t)$.

where **Hilbert**$(\cdot)$ is the Hilbert transform and **LPF**$[\cdot]$ is a low-pass filter with a cutoff frequency of 20 Hz. Speech sections and noise sections of the observed signals were estimated by using the robust voice activity detection (VAD) in noisy reverberant environments [18, 19]. The VAD algorithm consisted of three blocks. The first block is an estimate of the SNR that was used to mitigate against the effect of additive noise on the speech power envelope. The second block is a speech power envelope dereverberation based on the MTF concept. The last block is threshold processing on the dereverberated speech power envelope for a speech/non-speech decision.

The SNR was estimated from the mean power ratio of speech sections to noise sections. Speech sections were extracted by using a robust VAD algorithm [18, 19]. Since speech sections were affected due to the effect of additive noise, the estimated SNR could be obtained by removing this effect from speech sections. Next, the MTF in noisy environments $m_N(f_m)$ was calculated by using the estimated SNR of the noisy reverberant signal. The proposed method can generally calculate the STI in the same way as the previous method. However, MTFs in noisy reverberant environments multiply MTFs in seven octave-bands $m_{kR}(f_m), k = 1, 2, \cdots, 7$ by $m_N(f_m)$. Finally, the process described in Section 2 is used to calculate STI from the estimated MTFs.

Let us provide an example of how power envelope processing is related to the MTF concept. A sinusoidal power envelope as the original $e_x^2(t)$ $(= 0.5(1 + \sin(2\pi f_m t)))$ and $x(t)$ calculated from $e_x^2(t)$ and white noise carrier $c_x(t)$ are shown in Figs. 8(a) and (b); f_m was 10 Hz and $m(f_m)$ was 1. Figures 8(c) and (d) show $e_h^2(t)$ with $T_R = 0.5$ s and $h(t)$. Figures 8(e) and (f) show $e_n^2(t)$ and an $n(t)$ with an SNR of 3 dB, and Figures. 8(g) and (h) show $e_y^2(t)$ $(= e_x^2(t) * e_h^2(t) + e_n^2(t))$ and the observed noisy reverberant signal, $y(t)$ $(=x(t)*h(t)+n(t))$. The panels on the left ((a), (c), (e), and (g)) plot the power envelopes and those on the right ((b), (d), (f), and (h)) show the corresponding signals. This figure indicates $m(f_m)$ decreased from 1.0 (in Fig. 8(a)) to 0.404×0.5. The maximum deviation in the envelope between the dotted lines in Fig. 8(g) is relative to that in Fig. 8(a) and the reduction in Fig. 8(g). The solid line in Fig. 8(g) indicates restored power envelope $\hat{e}_x^2(t)$ obtained from noisy reverberant power envelope $e_y^2(t)$ (Fig. 8(g)) with $T_R = 0.5$ s and SNR $= 3$ dB. These are the estimated MTF and SNR in Fig. 7. We can see that power envelope processing could precisely restore the power envelope from a noisy reverberant signal in terms of its shape and magnitude.

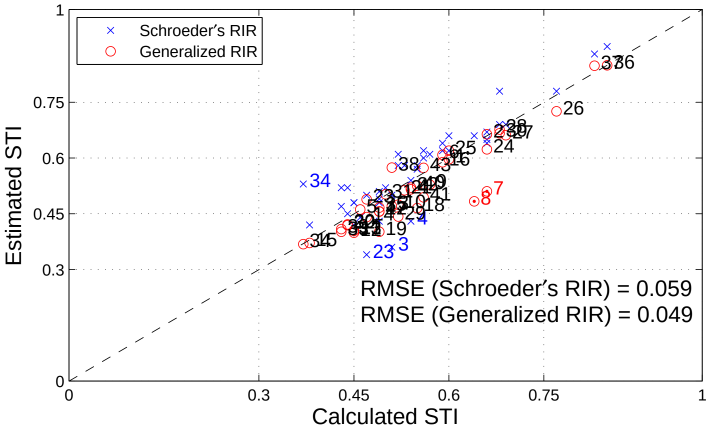

Figure 9. Estimated STIs from reverberant AM signals.

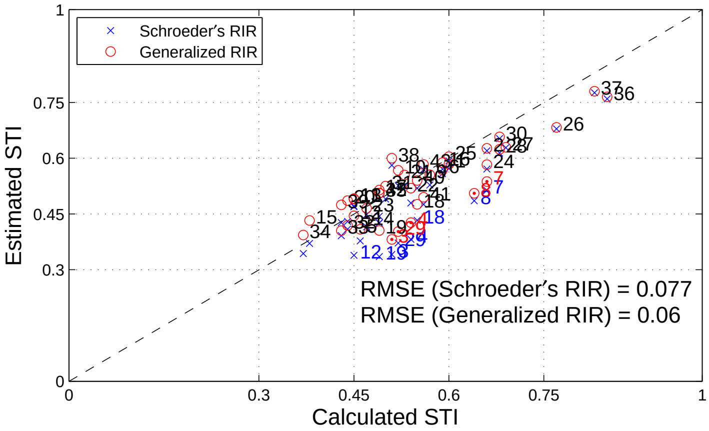

Figure 10. Estimated STIs from reverberant speech signals.

5. Evaluations.

5.1. **Evaluation for issue (1).** We carried out simulated evaluations using reverberant signals to determine whether they worked on blind estimates on the basis of our concept as well as to consider issue (1): whether the proposed method can estimate STIs even if the

RIR cannot be approximated as Schroeder's RIR model. We used reverberant signals that were generated by convolving the AM-signal with RIRs. This was because AM-noise can be regarded as simulated signals and the AM-noise signal was designed to have periodic information in the power envelope. The period in the power envelope was set to 0.2 s so that the fundamental modulation frequency was 5 Hz. We used 43 realistic RIRs in these simulations, which were produced in the SMILE2004 datasets [21] summarized in Table 2 (Room ID Nos. 1–43).

Figure 9 plots the STIs estimated from reverberant AM signals. The horizontal axis indicates STIs directly calculated from RIRs and the vertical axis indicates estimated STIs. The symbols "·" and "∘" correspond to the estimated STIs using the previous and proposed methods. The numbers in Fig. 9 correspond to the results for 43 realistic RIRs. The red numbers indicate over- or under-estimates of STIs by 0.1 by the proposed method, and the blue numbers indicate those of STIs by the previous method. The dashed line in the figure indicates the optimal estimated values for STIs. The root-mean-squared error, RMSE is 0.049 with the proposed method and 0.059 with the previous method. This means all STIs should be on this line if the method can accurately estimate them.

5.2. **Evaluation for issue (2).** We then carried out subsequent simulations using the reverberant speech signals to consider issue (2): whether the proposed method can estimate STIs from not only reverberant AM but also reverberant speech signals. The speech signals were ten long Japanese sentences uttered by ten speakers (five males and five females) from the ATR database [20]. We used the reverberant speech signals generated by convolving speech signals with 43 realistic RIRs from the SMILE datasets.

Figure 10 plots the estimated STIs from reverberant speech signals. The figure format is the same as that for Fig. 9. This figure indicates that most estimated STIs are accurate because most plots are on the optimal line. Here, RMSE is 0.060 with the proposed method and is 0.077 with the previous method. The results for realistic RIRs indicate that the proposed approach could effectively estimate STIs from the observed reverberant speech signals (long sentences) even if the RIR could not be approximated as Schroeder's RIR model.

5.3. **Evaluation for issue (3).** We carried out simulated evaluations using noisy reverberant signals to consider issue (3): whether the proposed method can correctly estimate STI in noisy reverberant environments. The speech signals were ten long Japanese sentences uttered by ten speakers (five males and five females) from the ATR database [20]. We used 43 realistic RIRs in these simulations, which were produced in the SMILE2004 datasets [21], as shown in Table 2 (Room ID Nos. 1–43), and four types of noise (NOISEX-92: [22], white, pink, babble, and factory noise) under two SNR conditions (SNR= 20 and 5 dB). We used noisy reverberant speech signals that were generated by convolving these signals with 43 realistic RIRs and then adding white noise.

The estimated STIs from the noisy reverberant speech signal are plotted in Fig. 11. The horizontal axis indicates STIs directly calculated from RIRs and the vertical axis indicates estimated STIs. The symbols "×" and "∘" correspond to the STIs estimated by the previous and proposed methods. The red and blue symbols indicate the estimated STIs at SNR= 20 dB and SNR= 5 dB. The RMSEs, between the calculated and estimated STIs were used to evaluate the previous and proposed methods.

RMSEs were 0.253 at SNR= 20 dB and 0.336 at SNR= 5 dB with the proposed method and 8.96 at SNR= 20 dB and 5.92 at SNR= 5 dB with the previous method when observed speech signals were used under the white noise and reverberation conditions given in Fig. 11(a). This means all STIs should be on the dashed line if the method can accurately estimate them. These results have almost the same trend as those under pink noise and

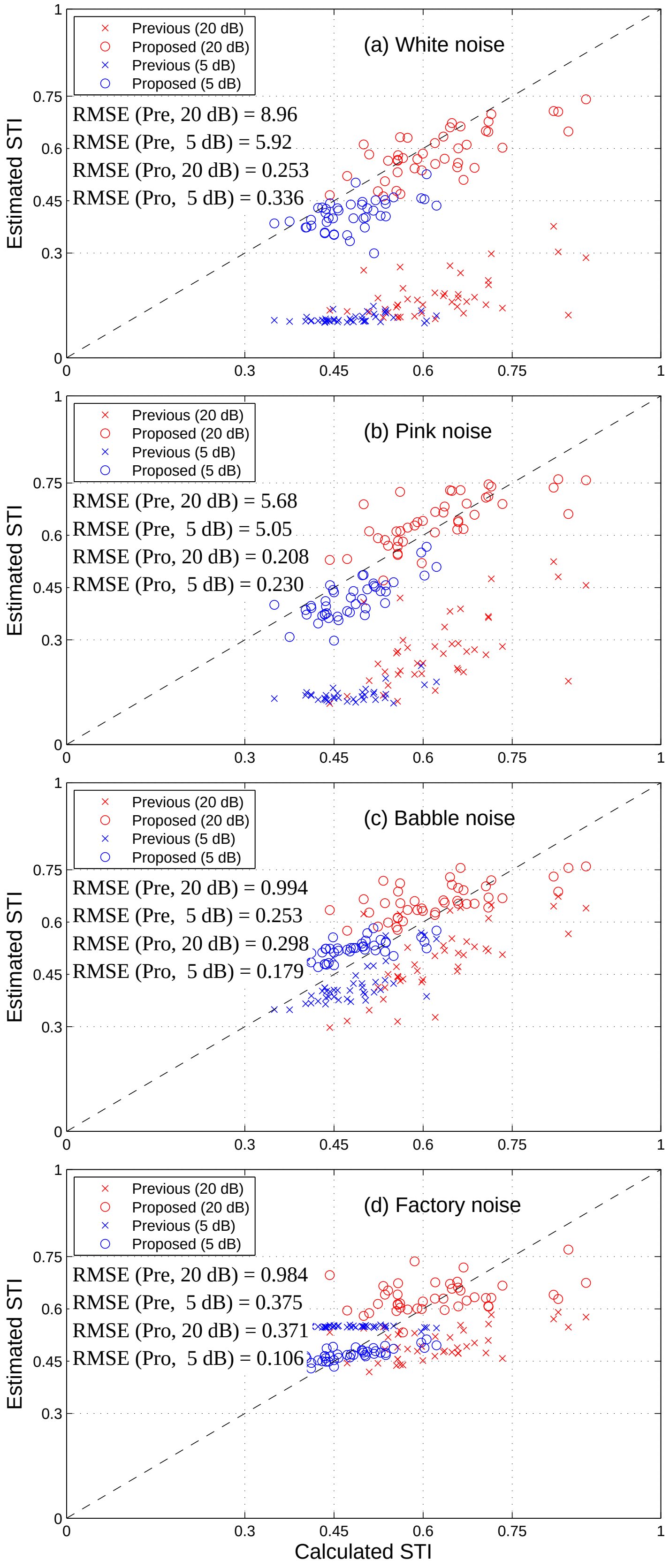

FIGURE 11. Estimated STIs from observed speech signals under background noise and reverberation conditions where noise types are: (a) white noise, (b) pink noise, (c) babble noise, and (d) factory noise.

reverberation conditions in Fig. 11(b). On the other hand, these results do not have the same trend as those in Figs. 11(c) and 11(d) when observed speech signals were used under babble noise or factory noise and under reverberation conditions. The RMSEs for noisy reverberant speech signals under the last two conditions were less than those for white or pink noise and reverberation. In the MTF concept, we assumed that background noise is stationary. Therefore, the MTF in noisy environments can be represented as Eq. (15). Since babble and factory noise are not stationary noise, this mismatching provides a different trend in our observation. In these simulations, we aimed to investigate the feasibility of the proposed method under various noise types. As the results, it was found that the proposed method could be used in all cases to effectively estimate STIs from observed noisy reverberant signals.

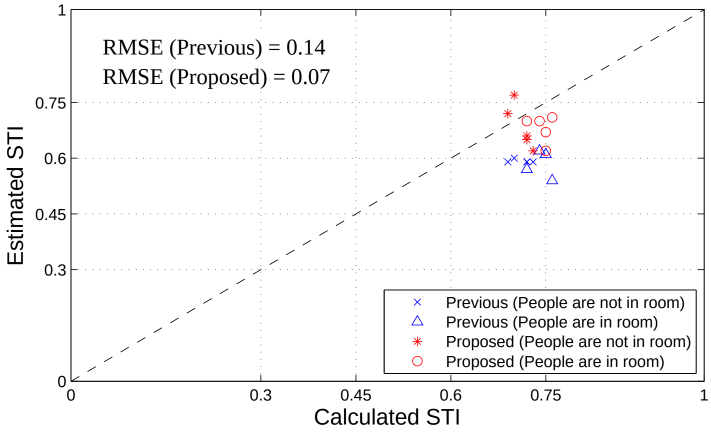

FIGURE 12. Estimated STIs from observed speech signals in real environments.

5.4. **Evaluation for issue (4).** We then carried out subsequent experiments using RIR measuring systems to consider issue (4): whether the proposed method can estimate STIs from observed signals in real environments where people cannot be excluded. The speech signals were the same as those used in the second simulations (ten long Japanese sentences uttered by ten speakers). The RIRs we tested were measured in rooms at our university by using an RIR measuring system [23] (B&K Omni-power Omnidirectional Sound Source: Type 4292-L, B&K Power Amplifier: Type 2734, B&K Hand-held analyzer: Type 2250, and B&K DIRAC Room acoustics software: Type 7841, ver. 5.0). Here, we measured the RIRs under two conditions: (i) no people were in the rooms and (ii) sixteen people with ear protectors were in the rooms. The original source of the speech signals was output from the omni-speakers, and then reverberant speech signals were observed with a hand-held analyzer to estimate STIs without having to measure RIRs.

Figure 12 plots the estimated STIs from reverberant speech signals. The figure format is the same as that for Figs. 9, 10, and 12. The symbols "×" and "△" indicate the STIs estimated by the previous method where people were not and were in rooms. The symbols "*" and "∘" indicate the STIs estimated by the proposed method where people were not and were in rooms.

Figure 12 reconfirms that real STIs were affected when people were in the room. This figure also indicates that most STIs estimated by the proposed method were accurate whereas those by the previous method were under-estimated in all cases. This is because the corresponding T_Rs estimated by the previous method were not suitable values and most tended to be extremely under- and over-estimated due to background noise (effect of flooring noise). In contrast, the proposed method could adequately estimate T_R so that the STI could also be adequately estimated in realistic conditions. It is, therefore, important for the MTF in Eq. (11) to be close to the measured MTF when estimating STIs.

5.5. **Discussion.** According to the above evaluations, our approach could resolve the four remaining issues. Important findings are summarized as follows.

1. The generalized RIR model could be used to account for important characteristics of RIR, that is, the shapes of the power envelope and the corresponding MTF, so that STIs could be correctly estimated from the observed signal by the proposed scheme.
2. The common features on the modulation spectra of AM signals and speech signals could be characterized as the modulation peaks related to periodicity in the power envelope and resulting tilt of modulation spectra due to reverberation. Therefore, these common features could be used to estimate STI correctly under various types of signal (AM and speech).
3. The MTF in noisy reverberant environments could be modeled as the product of the MTF in reverberant environments with the MTF in noisy reverberant environments separately, such like Eq. (15). The MTF in reverberant environments could be estimated by our current approach, that is, by estimating T_R. The MTF in noisy reverberant environments could be estimated by estimating SNR via a noise-robust VAD technique. Therefore, the STI could be correctly estimated under noisy reverberant conditions by the proposed method.
4. By resolving the first three issues, it was found that the proposed method could estimate STIs under real conditions.

These positive results could not have been obtained if the four issues had been reconsidered sequentially and then resolved step by step.

6. **Conclusions.** This paper presented a specified method of blindly estimating speech transmission indices (STIs) from observed speech signals under noise and reverberation conditions, on the basis of the modulation transfer function (MTF) concept, to resolve the four issues remaining from our previous paper. We carried out simulations using speech signals in realistic environments (under noisy and reverberant conditions) and experiments using speech signals where people were and were not in rooms. The results obtained from the simulations revealed that the proposed method could accurately estimate STIs from noisy reverberant speech signals. The results from the experiments revealed that the proposed approach could effectively estimate these STIs in realistic situations where people could not be excluded. This means that the proposed method can now obtain optimal estimates of MTFs/STIs with background noise.

Acknowledgment. This work was supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE; 131205001) of the Ministry of Internal Affairs and Communications (MIC), Japan, by a Grant-in-Aid for challenging Exploratory Research (No. 16K12458) and Innovative Areas (No. 16H01669) from MEXT, Japan, and by the Secom Science and Technology Foundation. and by the Secom Science and Technology Foundation. The authors thank our collaborators, Mr. Kyohei Sasaki, Mr. Tomohiro Ikeda, and Dr. Ryota Miyauchi to discuss our results.

REFERENCES

[1] ISO 3382, *Acoustics–Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters,* 2nd ed. Géneve, 1997.

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

[3] IEC 60268-16:2003. *Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index.*

[4] H. Sato, M. Morimoto, H. Sato, and M. Wada, Relationship between listening difficulty and acoustical objective measures in reverberation fields, *J. Acoust. Soc. Am.*, vol. 123, no. 4, pp. 2087–2093, 2008.

[5] H. Sato, M. Morimoto, H. Sato, and M. Wada, Relationship between listening difficulty and objective measures in reverberant and noisy fields for young adults and elderly persons, *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4596–4605, 2012.

[6] T. Houtgast and H. J. M. Steeneken, The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility, *Acustica.*, vol. 28, pp. 66–73, 1973.

[7] T. Houtgast, H. J. M. Steeneken, and R. Plomp, Predicting speech intelligibility in rooms from the Modulation Transfer Function. I. General Room Acoustics, *Acustica*, vol. 46, pp. 60–72, 1980.

[8] F. F. Li, and T. J. Cox, Speech transmission index from running speech: A neural network approach, *J. Acoust. Soc. Am.*, vol. 113, pp. 1999-2008, 2003.

[9] P. Kendrick, T. J. Cox, Y. Zhang, J. A. Chambers, and F. F. Li, Room acoustic Parameter extraction from music signals, *Proc. ICASSP2006*, **V**, pp. 801–804, 2008.

[10] P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, and J. A. Chambers, Monaural room acoustic parameters from music and speech, *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 278–287, 2008.

[11] P. P. Parada, D. Shama, and P. A. Naylor, Non-intrusive estimation of the level of reverberation in speech, *Proc. ICASSP2014*, pp. 4718–4722, 2014.

[12] M. Unoki, T. Ikeda, and M. Akagi, Blind Estimation Method of Speech Transmission Index in Room Acoustics,*Proc. Forum Acousticum 2011*, CDROM, 2011.

[13] M. Unoki, T. Ikeda, K. Sasaki, R. Miyauchi, M. Akagi, and N. S. Kim, Blind method of estimating speech transmission index in room acoustics based on concept of modulation transfer function, *Proc. ChinaSIP2013*, pp. 308–312, 2013.

[14] M. Unoki, K. Sasaki, R. Miyauchi, M. Akagi, and N. S. Kim, Blind method of estimating speech transmission index from reverberant speech signals, *Proc. EUSIPCO2013*, 1569746153, pp. 1–5, 2013.

[15] M. R. Schroeder, New method of measuring reverberation time, *J. Acoust. Soc. Am*, vol. 37, pp. 409-412, 1965.

[16] M. R. Schroeder, Modulation transfer functions: definition and measurement, *Acustica*, vol. 49, pp. 179–182, 1981.

[17] M. Unoki, Y. Yamasaki, and M. Akagi, MTF-based power envelope restoration in noisy reverberant environments, *Proc. EUSIPCO2009*, pp. 228–232, 2009.

[18] S. Morita, X. Lu, and M. Unoki, Signal to noise ration estimation based on an optimal design of subband voice activity detection, *Proc. ISCSLP2014*, pp. 560–564, 2014.

[19] S. Morita, M. Unoki, X. Lu, and M. Akagi, Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments, *Proc. ISCSLP2014*, pp. 108–112, 2014.

[20] T. Takeda, Y. Sagisaka, K. Katagiri, M. Abe, and H. Kuwabara, Speech Database User's Manual, *ATR Technical Report*, TR-I-0028, 1988.

[21] Architectural Institute of Japan, *Sound library of architecture and environment*, Gihodo Shuppan Co., Ltd., Tokyo, 2004.

[22] A. Varga and H. J. M. Steeneken, ssessment for automatic speech recognition: II NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[23] Room acoustics measurements - DIRAC.
http://www.bksv.com/Products/analysis-software/acoustics/building-acoustics/

TABLE 2. Datasets for room impulse responses (RIRs) using simulations and experiments on blindly estimating STIs. RIR Nos. (ID. Nos. 1 – 43) are File Nos. in SMILE2004 [21]. ID Nos. 44 – 47 are Nos. in our recordings.

ID No.	Room condition	RIR No.	T_{60} [s]
1	Multi-purpose hall 1 (with reflex board)	301	1.09
2	Multi-purpose hall 1 (without reflex board)	302	0.80
3	Multi-purpose hall 2 (with reflex board)	303	1.44
4	Multi-purpose hall 2 (without reflex board)	304	1.04
5	Multi-purpose hall 3 (with reflex board)	305	1.93
6	Multi-purpose hall 3 (without reflex board)	306	1.35
7	Multi-purpose hall 4 (with absorption board)	307	1.42
8	Multi-purpose hall 4 (without absorption board)	308	1.54
9	Multi-purpose hall 5 ($14,000$ m^3)	319	1.47
10	Multi-purpose hall 6 ($19,000$ m^3)	321	2.16
11	Classic concert hall 1 ($5,600$ m^3)	309	2.35
12	Classic concert hall 1 ($d = 6$ m)	310	2.34
13	Classic concert hall 1 ($d = 11$ m)	311	2.35
14	Classic concert hall 1 ($d = 15$ m)	312	2.39
15	Classic concert hall 1 ($d = 19$ m)	313	2.38
16	Classic concert hall 2 ($6,100$ m^3)	314	1.14
17	Classic concert hall 3 ($20,000$ m^3)	315	1.96
18	Classic concert hall 4 (with absorption curtain)	316	1.92
19	Classic concert hall 4 (without absorption curtain)	317	2.55
20	Classic concert hall 5 ($17,000$ m^3)	323	2.32
21	Classic concert hall 6 (1F front)	324	1.77
22	Classic concert hall 6 (2F side)	325	1.74
23	Classic concert hall 6 (3F)	326	1.69
24	Lecture room with flatter echoes	201	1.36
25	Theater hall ($3,900$ m^3)	318	0.85
26	Meeting room (130 m^3)	401	0.62
27	Lecture room (400 m^3)	402	1.12
28	Lecture room ($2,400$ m^3)	403	1.09
29	General speech hall ($11,000$ m^3)	404	1.54
30	Church 1 ($1,200$ m^3)	405	0.71
31	Church 2 ($3,200$ m^3)	406	1.30
32	Event hall 1 ($28,000$ m^3)	407	3.03
33	Event hall 2 ($41,000$ m^3)	408	3.62
34	Gym 1 ($12,000$ m^3)	409	2.82
35	Gym 2 ($29,000$ m^3)	410	1.70
36	Living room (110 m^3)	411	0.36
37	Movie theater (560 m^3)	412	0.38
38	Atrium ($4,000$ m^3)	413	1.57
39	Tunnel ($5,900$ m^3)	414	2.72
40	Concourse in train station	415	1.95
41	General speech hall 2 (1F front)	416	1.53
42	General speech hall 2 (1F center)	417	1.49
43	General speech hall 2 (1F balcony)	418	1.40
44	Seminar Room (I-95) ($T = 15.9$ °C, $H = 43$)	—	0.45 (0.55)
45	AV Laboratory (I-94) ($T = 21.0$ °C, $H = 39$)	—	0.54 (0.38)
46	IS Lecture Hall ($T = 12.7$ °C, $H = 50$)	—	0.53 (0.57)
47	IS Lecture Room (I3-4) ($T = 12.3$ °C, $H = 49$)	—	0.63 (0.47)