BLIND METHOD OF ESTIMATING SPEECH TRANSMISSION INDEX FROM REVERBERANT SPEECH SIGNALS

Masashi Unoki, Kyohei Sasaki, Ryota Miyauchi, Masato Akagi, and Nam Soo Kim*

School of Information Science, Japan Advanced Institute of Science and Technology 1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan * School of electrical Engineering and INMC, Seoul National University Seoul 151-742, Republic of Korea {unoki, k.sasaki, ryota, akagi}@jaist.ac.jp, nkim@snu.ac.kr

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. This paper proposes a specified method of blindly estimating the STI from observed reverberant speech signals, based on the concept of the modulation transfer function. The proposed method has been developed from a simplified method that the authors previously reported, to resolve three issues: (1) whether the method could estimate STIs even if the room impulse response could not be approximated as Schroeder's model, (2) whether the method could not only correctly estimate STIs from reverberant AM but also reverberant speech signals, and (3) whether the method could estimate STIs from observed signals in reverberant environments where people cannot be excluded. Simulations were carried out to verify the first two issues, by using both AM and speech signals in reverberant environments. Experiments were also carried out in several rooms, which a few people were in, to verify the last issue. The results revealed that the proposed approach could be used to effectively estimate STIs from reverberant speech signals in various room acoustics even if people were in the room.

Index Terms— Speech transmission index, modulation transfer function, room impulse response (RIR), Schroeder's RIR model, generalized RIR model

1. INTRODUCTION

The quality of speech transmission must be evaluated to design the required room acoustics, although many subjective experiments should be carried out to evaluate it and the costs involved are very expensive. Therefore, objective indices and measurements in room acoustics are needed to inexpensively assess the quality of speech and its intelligibility. Since examples are the articulation index (AI), the degree of contribution of early reflections, the Deutlichkeit, D_{50} , and the speech transmission index (STI) [1, 2].

The STI is a particularly important objective measurement that can be used to assess the quality of speech transmission in room acoustics [2]. It is well-known that the correlation between listening difficulty ratings and STI is the strongest of all tested objective measures [3]. Methods of calculating STI have currently been standardized by the IEC 60268-16 [4]. This standard is based on the



Fig. 1. General scheme for STI calculations based on MTF concept.

concept of the modulation transfer function (MTF) that was proposed by Houtgast and Steeneken [5, 6]. This concept has aimed at accounting 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 [5, 6], as shown in Fig. 1. Therefore, measurements of room impulse responses (RIRs) represent an important key to calculating STI. Since these measurements must be done in actual environments, it is difficult to obtain these characteristics by using typical methods of measuring RIRs in sound environments where people cannot be excluded, e.g., in common spaces such as stations, airports, and concourses.

There have been a few different approaches that can be used to blindly estimate acoustic parameters such as the reverberation time (RT) and the early decay time related to STI from received music and/or speech signals [8, 9]. These approaches have employed machine learning techniques to estimate these parameters. Although they can provide accurate estimates of these parameters, we need to have massive datasets in real environments to learn all of them. It is also very difficult to obtain a corpus of data that includes measured RIRs in which people cannot be excluded.

The authors, on the other hand, previously proposed a simplified method of blindly estimating STIs [10]. This method was used to correctly estimate STI from reverberant amplitude modulation (AM) signals in which RIR was approximated as Schroeder's model. However, three issues remained unresolved: (1) whether this could estimate STIs even if the RIR could not be approximated as Schroeder's model, (2) whether this could correctly estimate STIs from reverberant speech, and (3) whether this could estimate STIs in reverberant environments where people cannot be excluded.

This paper proposes specified blind estimates of STI from reverberant speech signals to solve the three issues. The proposed method involves the same approach we previously used [10]. An advantage in our approach enables us to estimate STI in room acoustics where people cannot be excluded, without having to measure RIRs.

This work was supported by a Grant-in-Aid for Challenging Exploratory Research (No. 23650086) and an A3 foresight program made available by the Japan Society for the Promotion of Science.

Table 1. Relationship between speech quality and STI [3].

		1	1	1 2	L J
Quality	Bad	Poor	Fair	Good	Excellent
STI	0.0	0.3	0.45	0.6	0.75
	~ 0.3	~ 0.45	~ 0.6	~ 0.74	~ 1.0

2. CALCULATION OF SPEECH TRANSMISSION INDEX

The method of calculating STI has been standardized by the IEC 60268-16 [4]. The correspondence between STI and its effectiveness in assessing the quality of speech transmission in room acoustics has been summarized in Table 1 (see Fig. 4 in [3]).

The RIR in this method, is assumed to be the stochastic idealized RIR (Schroeder's RIR model [7]), h(t), defined as

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

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

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

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

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

where a is normalized as one. Here, T_R is a parameter of 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 the IEC 60268-16 [4], in details).

(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 $k = 1, 2, 3, \dots, 7$). This has fourteen modulation frequencies (the F_i ranges from 0.63 to 12.5 Hz and $i = 1, 2, 3, \dots, 14$).

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

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

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

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

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

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

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

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

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



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

3. PREVIOUS METHOD

3.1. Blind estimation of MTF/STI

In the previous methods, it was assumed there is no background noise. The previous method used three useful characteristics to estimate MTF: (1) the MTF at 0 Hz was 0 dB, i.e., a modulation index of 1.0, (2) the original modulation spectrum at the dominant modulation frequency, f_m , was the same as that at 0 Hz, and (3) the entire modulation spectrum of the reverberant signal was reduced as RT increased, according to 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 compensating for the reduced modulation spectrum at a dominant f_m based on the MTF being 0 dB ($m(f_m)$) was restored to 1.0 for all f_m s). Thus, T_R can be determined as

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

Figure 2 has 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 [10] 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 with AM in the bands.

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 based on 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 algorithm described in Section 2 is used to estimate STI from the estimated MTFs.

3.2. Remaining issues

The previous method could be used to estimate the MTF/STI without having to measure RIR, where there is no background noise. However, there were three issues remaining from our preliminary studies [10] as to (1) whether the method could be used to estimate STIs even if the RIR could not be approximated as Schroeder's model, (2) whether tit could correctly be used to estimate STIs from not only reverberant AM but also reverberant speech signals, and (3) whether it could be used to estimate STIs from observed signals in reverberant environments where people could not be excluded.

The estimated STI and \hat{T}_R were frequently incorrect with 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



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



Fig. 4. Block diagram for proposed method of estimating STIs.

that of the measured RIRs, since the corresponding RIRs had onsettransition in the temporal envelope, as can be seen from Fig. 3(a). Since AM signals were used to verify the concept of the previous method, issues (2) and (3) have not yet been resolved. To do this, general sounds such as speech signals should be used to reconsider these issues.

4. PROPOSED METHOD

4.1. Generalized RIR and corresponding MTF

Schroeder's RIR model was modified as a generalized RIR to represent onset-transition in the temporal envelope of RIR:

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

where a is a gain factor of RIR and n is the order of the RIR. This is the same as Schroeder's RIR at n = 1. The generalized RIR has greater flexibility than that in Schroeder's RIR.

The MTF of the generalized RIR can be derived by substituting Eq. (10) into Eq. (2) as

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

The difference between the MTFs of Schroeder's RIR and generalized RIR is an exponent of -(2n-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. Figure 3(a) indicates that the generalized RIR model could more correctly approximate the temporal envelope of the measured RIR than Schroeder's



Fig. 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.

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.

Figure 4 is a block diagram of the method we propose for blindly estimating STIs. 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).

We assumed that the input signal would have two modulation frequencies and both modulation indexes at the modulation frequencies would be 1.0. In addition, we could use the same characteristics (1)–(3) with regard to the modulation spectrum and the MTF relation as mentioned in Sec. 2.2. The root mean squared (RMS) error between the modulation indexes and the MTF of Eq. (11) at the two modulation frequencies was used to estimate the two parameters (T_R and n) of the generalized RIR model. Here, this is defined as

$$\operatorname{RMS}(T_R, n) = \sqrt{\frac{1}{L} \sum_{\ell=1}^{L} \left[|E_y(f_{m\ell})| - m(f_{m\ell}, T_R, n) \right]^2}, \quad (12)$$

where $E_y(f_{m\ell})$ is the modulation spectrum of output at specific modulation frequency $f_{m\ell}$ 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 n. Here, L is 2. The estimated T_R and n can be determined as

$$\{\hat{T}_R, \hat{n}\} = \arg\min_{T_R, n} \text{RMS}(T_R, n).$$
 (13)

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, $m(f_m, T_R, n)$, in Eq. (11). The modulation spectrum of input has peaks of 0 dB at the corresponding modulation frequencies and the corresponding peaks are reduced according to $m(f_m, T_R, n)$. Therefore, \hat{T}_R and \hat{n} are estimated from y(t) by using Eq. (13) 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 be also used to determine these two parameters, \hat{T}_R and \hat{n} in Fig. 4.

5. EVALUATION

We carried out simulated evaluations using reverberant signals to confirm whether they worked on blind estimates based on our concept as well as to consider the remaining issue (1). 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 be 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 [12] summarized in Table 2.

Figure 6 plots the estimated STIs from reverberant AM signals. The horizontal axis indicates STIs directly calculated from RIRs and the vertical axis indicates estimated STIs. The symbols "·" and "o" correspond to the estimated STIs using the previous and proposed methods. The numbers in Fig. 6 correspond to the results for 43 realistic RIRs. The red numbers indicate over- or under-estimates of STIs by 0.1 with the proposed method and the blue numbers indicate those of STIs with the previous method. The dashed line in the figure indicates the optimal estimated values for STIs. The root-mean-squared error, RMS, with the proposed method is 0.049 while it was 0.059 with the previous method. This means all STIs should be on this line if the method can be used to accurately estimate them.

We then carried out subsequent simulations using the reverberant speech signals to reconfirm the remaining issue (2). The speech signals were ten long Japanese sentences uttered by ten speakers (five males and five females) from the ATR database [11]. We used the reverberant speech signals that were generated by convolving speech signals with 43 realistic RIRs from the SMILE database.

Figure 7 plots the estimated STIs from reverberant speech signals. The figure format is the same as that for Fig. 6. This figure indicates that most of the estimated STIs are located on accurate estimates because most of the plots are on the optimal line. Here, RMS with the proposed method is 0.060 while it is 0.077 with the previous method. The results for realistic RIRs indicate that the proposed approach could effectively be used to estimate STIs from the observed reverberant speech signals (long sentences) even if the RIR could not be approximated as Schroeder's RIR model.

We finally carried out subsequent experiments using RIR measuring systems to reconfirm the remaining issue (3). 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 are listed in Table 2 from ID. Nos. 44 to 47. These were measured in rooms at our university by using an RIR measuring system [13] (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). In this case, we measured RIRs under two conditions where (i) people were not in the rooms and where (ii) two 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 by Hand-held analyzer, to estimate STI without having to measure RIRs.

Figure 8 plots the estimated STIs from reverberant speech signals. The figure format is the same as that for Fig. 6. The symbols "." and "x" indicate the STIs estimated with the proposed method where people were not in the rooms and where people were in them. The "o" indicates 0.1-correct estimates of STI with the proposed method. This figure indicates that most of the estimated STIs are located on accurate estimates while they are somewhat under-estimated in several cases. This is because the corresponding

estimated T_R s are not located on the high-correlation line and most of them tend to be extremely under- and over-estimated due to background noise (effect of flooring noise). It is important for the MTF in Eq. (11) to be close to the measured MTF in estimating STIs. We intend to do this in future work to promote further developments.

6. CONCLUSIONS

This paper proposed a specified method of blindly estimating STIs from observed speech signals, based on the MTF concept. We carried out simulations using both AM and speech signals in realistic environments and experiments using speech signals where people were in rooms to take the three issues into consideration. The results we obtained from the simulations revealed that (1) the proposed method could estimate STIs even if RIRs could not be approximated as Schroeder's model and (2) the proposed method correctly estimated STIs from reverberant AM and/or speech signals. The results from the experiments revealed that the proposed approach could effectively be used to estimate these STIs in realistic situations where people could not be excluded. We intend to reconsider optimal estimates of MTFs/SITs due to background noise in future work.

7. REFERENCES

- ISO 3382, Acoustics-Measurement of the Reverberation Time of Rooms with Reference to Other Acoustical Parameters, 2nd ed. Géneve, 1997.
- [2] Kuttruff, H. Room Acoustics, 3rd ed. (Elsevier Science Publishers Ltd., Lindin), 1991.
- [3] Sato, H., Morimoto, M., Sato, H., Wada, M. "Relationship between listening difficulty and acoustical objective measures in reverberation fields," *J. Acoust. Soc. Am.*, **123**(4), 2087–2093, 2008.
- [4] IEC 60268-16:2003. Sound system equipment Part 16: Objective rating of speech intelligibility by speech transmission index.
- [5] Houtgast, T. and Steeneken, H. J. M., "The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility," *Acustica.*, 28, 66–73, 1973.
- [6] Houtgast, T., Steeneken, H. J. M., and Plomp, R., "Predicting speech intelligibility in rooms from the Modulation Transfer Function. I. General Room Acoustics," *Acustica*, 46, 60–72, 1980.
- [7] Schroeder, M. R., "Modulation transfer functions: definition and measurement," *Acustica*, **49**, 179–182, 1981.
- [8] Li, F. F. and Cox, T. J., "Speech transmission index from running speech: A neural network approach," *J. Acoust. Soc. Am.* 113, 1999-2008, 2003.
- [9] Kendrick, P., Cox, T. J., Li, F. F., Zhang, Y. and Chambers, J. A., "Monaural room acoustic parameters from music and speech," J. Acoust. Soc. Am. 124(1), 278–287, 2008.
- [10] Unoki, M. Ikeda, T., and Akagi, M., "Blind Estimation Method of Speech Transmission Index in Room Acoustics," *Prof. Forum Acousticum 2011*, CDROM, 2011.
- [11] Takeda, T., Sagisaka, Y., Katagiri, K., Abe, M., and Kuwabara, H. "Speech Database User's Manual," *ATR Technical Report*, TR-I-0028, 1988.
- [12] Architectural Institute of Japan, "Sound library of architecture and environment," Gihodo Shuppan Co., Ltd., Tokyo, 2004.
- [13] Room acoustics measurements DIRAC. http://www.bksv.com/Products/analysis-software/acoustics/ building-acoustics/



Fig. 6. Estimated STIs from reverberant AM signals.



Fig. 7. Estimated STIs from reverberant speech signals.



Fig. 8. Estimated STIs from observed speech in real environments.

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 [12]. ID Nos. 44 – 47 are Nos. in our recordings. Reverberation time T_{60} was determined as the average of all T_{60} s on the transfer function at 125 Hz to 8 kHz in octave frequencies. MPH, CCH, and GSH are abbreviations for Multi-purpose hall, Classic concert hall, and General speech hall. RB, AB, and AC are those for "reflex board", "absorption board", and "absorption curtain". RIRs in ID Nos. 44 – 47 were measured in which people were in or out of rooms. T_{60} in branches indicates the reverberation time where people were in rooms.

ID No.	Room condition	RIR No.	T ₆₀ [s]
1	MPH 1 (with RB)	301	1.09
2	MPH 1 (without RB)	302	0.80
3	MPH 2 (with RB)	303	1.44
4	MPH 2 (without RB)	304	1.04
5	MPH 3 (with RB)	305	1.93
6	MPH 3 (without RB)	306	1.35
7	MPH 4 (with AB)	307	1.42
8	MPH 4 (without AB)	308	1.54
9	MPH 5 $(14,000 \text{ m}^3)$	319	1.47
10	MPH 6 (19,000 m ³)	321	2.16
11	$CCH \ 1 \ (5, 600 \ m^3)$	309	2.35
12	CCH 1 (d = 6 m)	310	2.34
13	CCH 1 ($d = 11 \text{ m}$)	311	2.35
14	CCH 1 ($d = 15 \text{ m}$)	312	2.39
15	CCH 1 ($d = 19 \text{ m}$)	313	2.38
16	CCH 2 (6, 100 m ³)	314	1.14
17	CCH 3 $(20, 000 \text{ m}^3)$	315	1.96
18	CCH 4 (with AC)	316	1.92
19	CCH 4 (without AC)	317	2.55
20	CCH 5 $(17,000 \text{ m}^3)$	323	2.32
21	CCH 6 (1F front)	324	1.77
22	CCH 6 (2F side)	325	1.74
23	CCH 6 (3F)	326	1.69
24	Lecture room	201	1.36
25	Theater hall $(3,900 \text{ m}^3)$	318	0.85
26	Meeting room (130 m ³)	401	0.62
27	Lecture room (400 m^3)	402	1.12
28	Lecture room $(2, 400 \text{ m}^3)$	403	1.09
29	$GSH(11,000 \text{ m}^3)$	404	1.54
30	Church 1 $(1, 200 \text{ m}^3)$	405	0.71
31	Church 2 $(3, 200 \text{ m}^3)$	406	1.30
32	Event hall 1 (28, 000 m^3)	407	3.03
33	Event hall 2 $(41,000 \text{ m}^3)$	408	3.62
34	$Gym 1 (12, 000 \text{ m}^3)$	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 $(4,000 \text{ m}^3)$	413	1.57
39	Tunnel (5, 900 m^3)	414	2.72
40	Concourse in train station	415	1.95
41	GSH 2 (1F front)	416	1.53
42	GSH 2 (1F center)	417	1 49
43	GSH 2 (1F balcony)	418	1.40
44	Seminar Room (I-95)		0.45 (0.55)
45	AV Laboratory (I-94)		0.54(0.38)
46	IS Lecture Hall	_	0.53(0.57)
47	IS Lecture Room (I3-4)	_	0.63(0.47)