JAIST Repository
https://dspace.jaist.ac.jp/

Title	Study on Blind Method of Estimating Speech Transmission Index from Noisy Reverberant Amplitude-Modulated-Signals
Author(s)	Miyazaki, Akikazu; Morita, Shota; Unoki, Masashi
Citation	Journal of Signal Processing, 18(4): 201-204
Issue Date	2014
Type	Journal Article
Text version	publisher
URL	http://hdl.handle.net/10119/12891
Rights	Copyright (C) 2014 信号処理学会. Akikazu Miyazaki, Shota Morita, and Masashi Unoki, Journal of Signal Processing, 18(4), 2014, 201-204. http://dx.doi.org/10.2299/jsp.18.201
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Journal of Signal Processing, Vol.18, No.4, pp.201-204, July 2014

SELECTED PAPER AT NCSP'14

Study on Blind Method of Estimating Speech Transmission Index from Noisy Reverberant Amplitude-Modulated-Signals

Akikazu Miyazaki, Shota Morita and Masashi Unoki

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {miyazaki.aki, s-morita, unoki}@jaist.ac.jp

Abstract

The authors previously proposed a method for blindly estimating the speech transmission index (STI) based on the concept of the modulation transfer function (MTF). This method, however, over- or under-estimates STIs from observed signals in real environments due to the effect of background noise. In this paper, the proposed method was developed from the previous method to resolve the problem by simultaneously estimating the effects of inverse MTFs in noisy and reverberant environments. Simulations were carried out to verify that the proposed method can correctly estimate STIs in noisy reverberant environments by using noisy reverberant AM signals. The results revealed that the proposed approach could be used to effectively estimate STIs from noisy reverberant signals in various room acoustics.

1. Introduction

The speech transmission index (STI) is an important objective measure that is used to assess the quality of speech transmission in room acoustics. The quality of speech transmission in room acoustics corresponds to "bad," ranging from 0.0 to 0.3, "poor," ranging from 0.3 to 0.45, "fair," ranging from 0.45 to 0.6, "good," ranging from 0.6 to 0.75, and "excellent," ranging from 0.75 to 1.0. Recently, it is also well known that STI has a correlation with listening difficulty higher than speech intelligibility.

The method for calculating STI was proposed by Houtgast et al. [1] and was standardized in IEC 60268-16 [2]. STI is based on the concept of the modulation transfer function (MTF). This concept is aimed at accounting for the relationship between the transfer characteristics of the enclosure and temporal envelopes of input and output signals, as shown in Fig. 1. STI calculation needs direct measurements of room impulse responses (RIRs) (or MTFs). However, people must be excluded from measurements of RIR for the protection of hearing because these measurements use an input signal with a high sound pressure level. Therefore, it is difficult to calculate STI by measuring RIRs in sound environments where people cannot be excluded, e.g., in public spaces such as stations, airports, and concourses.

The authors previously proposed a specified method for blindly estimating the STI in room acoustics based on the concept of the MTF [3]. Previous results revealed that (1) the previous method could estimate STIs even if RIRs could not be approximated as with Schroeder's RIR model, (2) the previous method effectively estimated STIs from reverberant amplitude modulated (AM) and speech signals, and (3) the previous method could estimate STIs in real environments where people were present in the room [3]. However, there is a problem in that the previous method over- or under-estimates the STI from the observed signals in real environments, due to the effect of background noise because this method assumes that room acoustics can be regarded as reverberant environments without noise.

In this paper, we propose a method for blindly estimating STI from observed noisy reverberant signals. The proposed method involves estimating inverse MTF from the observed signals with the same approach we previously used [3]. An advantage in our approach enables us to estimate STI in room acoustics where people cannot be excluded, without having to measure RIRs and the signal-to-noise ratio (SNR).

2. Calculation of Speech Transmission Index

Schroeder's RIR model was used in the STI calculation method. This model was defined as

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

where a is a gain factor of RIR, T_R is a parameter of reverberation time, and $\mathbf{c}_h(t)$ is a carrier such as a random variable of Gaussian white noise. The MTF of Schroeder's RIR model can be represented as

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

The MTF $m_R(f_m, T_R)$ has the characteristics of low-pass filtering as a function of the modulation frequency f_m and T_R.

The process of calculating STI can be summarized into five steps.

(i) Calculating MTFs in seven octave-bands: MTFs $m_k(f_i)$ are measured in seven octave-bands (center frequencies (CFs) range from 125 Hz to 8 kHz, and $k = 1, 2, 3, \cdots, 7$). In ad-

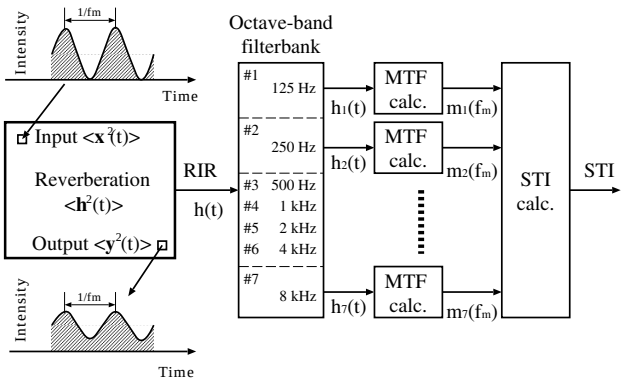

Figure 1: General scheme for STI calculations based on MTF concept

dition, these have fourteen modulation frequencies (f_i ranges from 0.63 to 12.5 Hz and $i = 1, 2, 3, \cdots, 14$).

$$m_k(f_i) = 1/\sqrt{1 + (2\pi f_i T_R/13.8)^2} \tag{3}$$

(ii) Calculating SNRs from MTFs: SNRs $N(k, i)$ are calculated from $m_k(f_i)$ as

$$N(k, i) = 10 \log_{10} m_k(f_i)/(1 - m_k(f_i)) \tag{4}$$

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

$$T(k, i) = \begin{cases} 1, & (15 < N(k, i)) \\ \frac{N(k,i)+15}{30}, & (-15 \leq N(k, i) \leq 15) \\ 0, & (N(k, i) < -15) \end{cases} \tag{5}$$

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

$$M(k) = \frac{1}{14} \sum_{i=1}^{14} T(k, i) \tag{6}$$

(v) Calculating STI: STI is calculated by $M(k)$ as

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

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

The previous method assumed that room acoustics can be regarded as reverberant environments without noise and a diffuse sound field. In addition, Schroeder's RIR model was modified as the generalized RIR model to account for the temporal envelope of the real RIR as follows.

$$\mathbf{h}(t) = at^{(b-1)} \exp(-6.9t/T_R) \mathbf{c}_h(t) \tag{8}$$

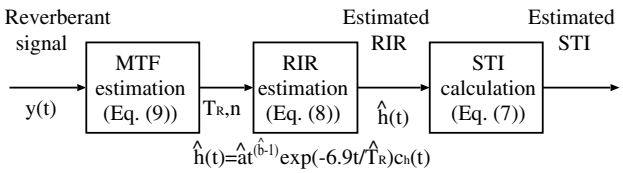

Figure 2: Block diagram of previous method

where b is an order of the RIR. The generalized RIR model has higher flexibility than Schroeder's RIR model. The MTF of the generalized RIR model is shown as

$$m_R(f_m, T_R, b) = \left[1 + \left(2\pi f_m \frac{T_R}{13.8}\right)^2\right]^{-\frac{2b-1}{2}} \tag{9}$$

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

Figure 2 shows a block diagram of the previous method. This method consists of three blocks: (I) estimating the inverse MTF from the observed signals by using the generalized RIR model, (II) estimating RIR from the estimated MTF, and (III) calculating STI by using the above method.

For estimating the MTF, the previous method had three useful characteristics: (1) the MTF at 0 Hz was 0 dB, (2) the original modulation spectrum at the dominant modulation frequency of f_m was the same as that at 0 Hz, (3) and the entire modulation spectrum of the reverberant signal was reduced as reverberation time increased, in accordance with MTF. These useful characteristics enabled us to model a strategy to blindly estimate the T_R and b of the inverse MTF $m_R^{-1}(f_m)$ that restores the original modulation spectrum from the entire modulation spectrum. Specifically, the optimal T_R and b are obtained with the minimum root mean square (RMS). These are defined as

$$\{\hat{T}_R, \hat{b}\} = \arg\min_{T_R, b} \text{RMS}(T_R, b) \tag{10}$$

$$\text{RMS}(T_R, b) = \sqrt{\frac{1}{L} \sum_{\ell=1}^{L} [|E_y(f_{m\ell})| - m_R(f_{m\ell}, T_R, b)]^2} \tag{11}$$

where $E_y(f_{m\ell})$ is the reduced modulation spectrum at a specific $f_{m\ell}$ and $m_R(f_{m\ell}, T_R, b)$ is the derived MTF of the generalized RIR at a specific $f_{m\ell}$ as a function of T_R and b, and L is 2. Then, an RIR $\mathbf{h}(t)$ is estimated on the basis of the generalized RIR model with T_R and b. Finally, the algorithm described in Section 2 is used to calculate the STI from the estimated MTF.

The previous method studied a method of blindly estimating STI in reverberant environments. Therefore, the previous method could be used to estimate the STI without having to measure RIR in reverberant environments. However, there is a problem in that the accuracy of the estimated STI is reduced

in noisy reverberant environments.

4. Proposed Method

The proposed method expands the previous method for noisy reverberant environments to resolve the above problem. The authors already proposed a method for restoring an MTF based power envelope in noisy reverberant environments [4]. This method can be used to estimate the STI in noisy reverberant environments.

Suppose that $\mathbf{x}(t)$, $\mathbf{y}(t)$, $\mathbf{h}(t)$, and $\mathbf{n}(t)$ are the original signal, noisy reverberant signal, RIR, and background noise, respectively. The signal is also assumed to be composed of the temporal envelope $e(t)$ and the carrier $\mathbf{c}(t)$ as random variables of Gaussian white noise. By assuming linear systems and the mutual independence between carriers, $e_y^2(t)$ can be represented as $e_y^2(t) = e_x^2(t) * e_h^2(t) + e_n^2(t)$, where "$*$" indicates convolution.

The MTF in the noisy reverberant environment can be represented as [4]:

$$m(f_m, T_R, b, \mathrm{SNR}) = m_R(f_m, T_R, b) \cdot m_N(f_m, \mathrm{SNR}) \tag{12}$$

Here, the MTF in the reverberant environment, $m_R(f_m, T_R, b)$, is defined in Eq. (9), and the MTF in the noisy environment, $m_N(f_m, \mathrm{SNR})$, is defined as

$$m_N(f_m, \mathrm{SNR}) = \frac{\overline{e_x^2}}{\overline{e_x^2} + \overline{e_n^2}} = \frac{1}{1 + 10^{-\frac{\mathrm{SNR}}{10}}} \tag{13}$$

Therefore, the MTF in the noisy reverberant environment $m(f_m)$ is defined as

$$\begin{aligned} & m(f_m, T_R, b, \mathrm{SNR}) \\ = & \left[1 + \left(2\pi f_m \frac{T_R}{13.8}\right)^2\right]^{-\frac{(2b-1)}{2}} \left(\frac{1}{1 + 10^{-\frac{\mathrm{SNR}}{10}}}\right) \end{aligned} \tag{14}$$

The previous method estimated STI by estimating the MTF in reverberant environments (Eq. (9)). The previous method was used in noisy reverberant environments, so estimation errors were caused by the effect of the MTF in noisy environments (Eq. (14)).

Figure 3 shows a block diagram of the proposed method. Power envelopes of observed signals $e_y^2(t)$ are calculated from the observed signals $y(t)$ as

$$\hat{e}_y^2(t) = \mathbf{LPF}\left[|y(t) + j \cdot \mathbf{Hilbert}(y(t))|^2\right] \tag{15}$$

where **Hilbert**$(\cdot)$ is the Hilbert transform and **LPF**$[\cdot]$ is a low-pass filter with a cut-off 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 [5]. The VAD algorithm consists of three blocks. The first block is an estimation of the SNR which is used to mitigate the additive noise effect on the speech power envelope. The second block is a speech power envelope dereverberation based on the MTF concept. The last block is a threshold processing on the dereverberated speech power envelope for speech/non-speech decision.

The SNR is estimated from the mean power ratio of speech sections to noise sections. However, speech sections are affected due to the additive noise effect. Therefore, the estimated SNR can be obtained by removing the additive noise effect from speech sections.

The mean power of $e_n^2(t)$, $\overline{e_n^2}$ is calculated from noise signals, and power envelope of reverberant signals is calculated by subtracting $\overline{e_n^2}$ from $e_y^2(t)$. This signal is an input signal for MTF estimation in the previous method. Therefore, the proposed method uses MTF estimation and RIR estimation from the previous method.

Next, the MTF in noisy environments $m_N(f_m)$ is calculated from Eq. (13) by using the estimated SNR of the noisy reverberant signal. Generally, calculating the STI of the proposed method can be done in the same way as with the previous method. However, MTFs in noisy reverberant environments multiply MTFs in seven octave-bands $m_{Rk}(f_m), k = 1, 2, \cdots, 7$ by $m_N(f_m)$.

Finally, the algorithm described in Section 2 is used to calculate STI from the estimated MTFs.

5. Evaluations

Simulations were carried out to confirm whether the proposed method could estimate STI in noisy reverberant environments. An AM-noise signal was used in these simulations. This 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 [6] (these RIRs' information was shown in [3]) and two types of white noise (SNR = 20 dB and 5 dB).

Figures 4 and 5 plot the estimated STIs from noisy reverberant AM-noise signals. The horizontal axis indicates STIs directly calculated from RIRs, and the vertical axis indicates estimated STIs. The symbols "$\cdot$" and "$\circ$" correspond to the STIs estimated with the previous and proposed methods. The numbers in Figs. 4 and 5 correspond to the results for the 43 realistic RIRs, where the numbers indicate over- or underestimates of STIs by 0.05. The dashed line in the figure indicates the optimal estimated values for STIs.

In cases where $\mathrm{SNR} = 20$ dB (Fig. 4), the RMS error with the proposed method was 0.04, while it was 0.05 with the previous method. This means that all STIs should be on this line if the method can be used to accurately estimate them. In cases where $\mathrm{SNR} = 5$ dB (Fig. 5), the RMS error with the proposed method was 0.05, while it was 0.11 with the previous method. The results for $\mathrm{SNR} = 5$ dB indicate that the proposed method could be used to accurately estimate STIs from the observed noisy reverberant AM-noise signals.

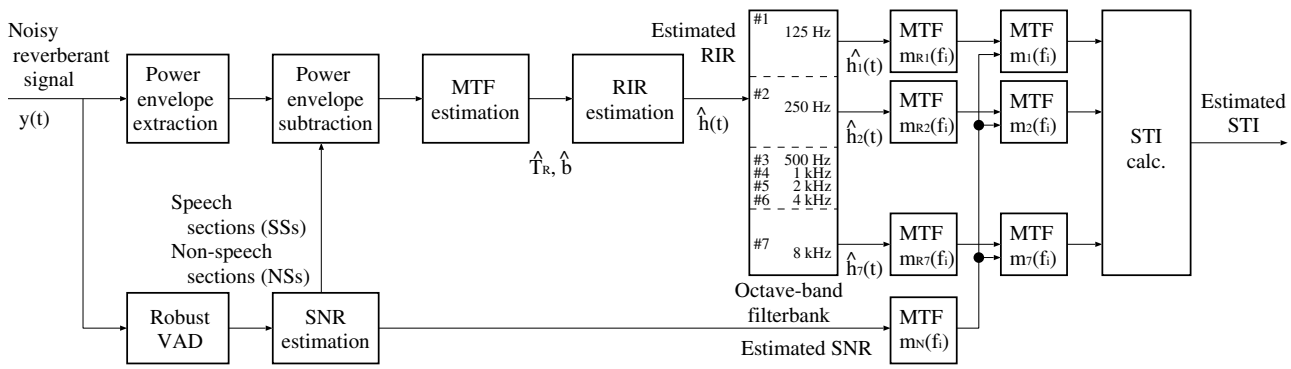

Figure 3: Block diagram of proposed method

6. Conclusion

In this paper, we proposed a method for blindly estimating STIs in noisy reverberant environments based on the MTF concept. The proposed method resolved the problem of the previous method over- or under-estimating the STI by simultaneously estimating the inverse MTF in noisy and reverberant environments. Results from simulations revealed that the proposed approach could be used to accurately estimate these STIs in noisy reverberant environments.

Acknowledgments

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.

References

[1] 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.

[2] IEC 60268-16:2003. Sound system equipment - Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index.

[3] 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, 2013.

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

[5] S. Morita, M. Unoki and M. Akagi: Study on robust voice activity detection in noisy reverberant environments, 2013 Autumn Meeting Acoustical Society of Japan, 1-P-22b, 2013.

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

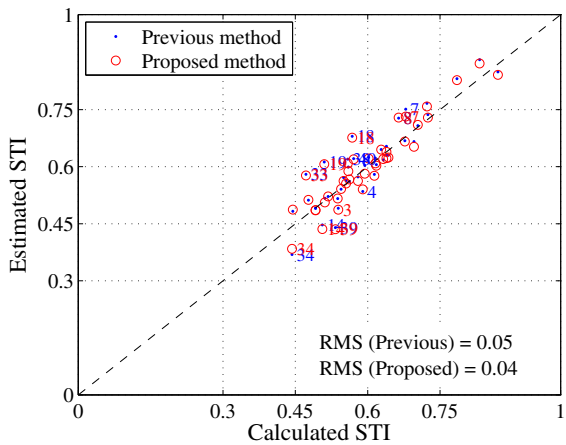

Figure 4: STI estimated from noisy reverberant AM signals (SNR = 20 dB)

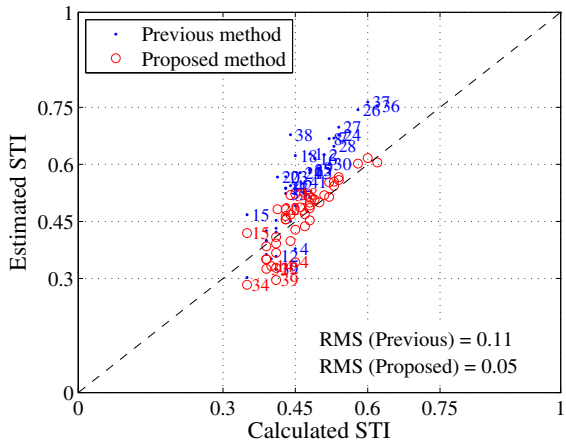

Figure 5: STI estimated from noisy reverberant AM signals (SNR = 5 dB)