IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 April 2023, accepted 2 May 2023, date of publication 15 May 2023, date of current version 24 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3276327

RESEARCH ARTICLE

Blind Estimation of Speech Transmission Index and Room Acoustic Parameters by Using Extended Model of Room Impulse Response Derived From Speech Signals

LIJUN WANG[1], SURADEJ DUANGPUMMET[2], AND MASASHI UNOKI[1], (Member, IEEE)
[1]Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan
[2]NECTEC, National Science and Technology Development Agency, Klong Luang, Pathum Thani 12120, Thailand

Corresponding author: Masashi Unoki (unoki@jaist.ac.jp)

This work was supported in part by the SCOPE Program of Ministry of Internal Affairs and Communications under Grant 201605002, and in part by the JSPS-NSFC Bilateral Joint Research Projects/Seminars under Grant JSJSBP120197416.

ABSTRACT The speech transmission index (STI) and room acoustic parameters (RAPs) are essential metrics for assessing speech quality and predicting listening difficulty in a sound field. Although STI and important RAPs, such as reverberation time and clarity, can be derived from the room impulse response (RIR), measuring the RIR in regularly occupied spaces is difficult. Hence, simultaneous blind estimation of STI and RAPs is an imperative challenge issue that must be addressed. However, most existing methods provide only a single parameter and require a massive dataset for model training. A deterministic method is presented for blindly estimating STI and five RAPs using a stochastic RIR model that approximates an unknown RIR. An algorithm is formulated that uses the temporal power envelope of a reverberant speech signal to determine the optimal parameters of the RIR model. A mathematical model of reverberation and dereverabation process was proposed based on the temporal power envelope of the signals. This model maps the parameters of the RIR model to the observed reverberant signal. The estimated RIR can then be synthesized using the optimal parameters to estimate the STI and RAPs. A simulation was conducted to evaluate the simultaneous estimation of STI and five essential RAPs from observed reverberant speech signals, in comparison to the best existing previous work. The root-mean-square error (RMSE) and Pearson correlation coefficient between the estimated and measured values were used as evaluation metrics. In terms of STI, the proposed method achieves the accuracy with an RMSE of 0.037. With regard to the reverberation time and other RAPs, the accuracy remains consistent with the previous works. The results show that the proposed method can effectively estimate STI and RAPs simultaneously without any training.

INDEX TERMS Room impulse response, modulation transfer function, reverberation, speech transmission index, room acoustic parameters.

I. INTRODUCTION

The clarity of music and the intelligibility of speech in a room play essential roles in daily life [1], [2]. An auditory space in which people are present encompasses walls, ceilings, and furnishings. It is important for sounds to be intelligible and easily audible in an auditory space. For example, general auditoriums require intelligible and easily audible sounds for emergency announcements, public addresses, and lectures [3, Ch. 11]. Concert halls are designed for delivering clear and transparent music [3, Ch. 5]. The intelligibility of speech and clarity of music can be determined by evaluating room acoustic characteristics (RACs) and diagnosing the degradation of acoustical quality. Since speech intelligibility and sound clarity are subjective perceptions, listening experiments are typically conducted to assess them. However, listening

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Kamrul Hasan.

experiments are expensive and time-consuming, which makes them impractical to apply in public spaces [4]. Hence, objective indices and room acoustic parameters (RAPs), i.e., the physical descriptions of room acoustics [5], [6], [7], have been proposed for use in the subjective evaluations of auditory spaces. RAPs have proved useful in various applications such as public address systems [8], hearing aids systems [9], and speech enhancement applications [10], [11].

Several objective indices and RAPs have been investigated and standardized [5], [12], [13], [14]. For example, in IEC 60268-16:2020, a speech transmission index (STI) based on the modulation transfer function (MTF), which is an objective index, is used to predict the speech intelligibility of a sound field [12], [13], [15]. The MTF quantifies the effect of reverberation on sound waves as they traverse an auditory space [16], [17]. The essential RAPs and their measurements are specified in ISO 3382-1:2009, including reverberation time (T_{60}), early decay time (EDT), clarity (early-to-late-arriving sound energy ratio: C_{80} / C_{50}), Deutlichkeit (early-to-total sound energy ratio: D_{50}), and center time (T_s) [14].

T_{60} is an essential parameter for representing the RACs. The STI and RAPs can be obtained from the measured room impulse response (RIR). An RIR fully represents the RACs of a sound field in the time domain while MTF describes the RACs in the frequency domain. The MTF can also be derived from the RIR. Measuring RIR requires the use of sine sweep signals or maximum length sequences as the excitation signals [5], [18]. However, it is difficult to measure RIR in spaces where people cannot be excluded, e.g., concourses and train stations, since RIR measurement requires high-energy sounds. RIR measurements are also limited to reflect changes in the RAC caused by variations in the number, location, and arrangements of the occupants and objects in a given space. Hence, the RACs in public spaces can be considered a time-varying system. The STI and RAPs measured in compliance with the specific standards may differ from non-compliant ones for the same auditory space. As a result, blind estimation methods have been proposed for obtaining STI and RAPs from an observed signal.

Blind estimation is challenging because it is an ill-posed inverse problem that derives a system solely from the output without prior knowledge of the input. Common methods include modeling the system to create a mapping from the output to the system using either mathematical derivation or machine learning techniques. For example, some analytical approaches have been proposed for blind estimation of the reverberation time and STI by Kendrick et al. [19], Unoki et al. [20], and Keshavarz et al. [21]. With the recent trend of machine learning in the field of signal processing, artificial neural networks have been widely applied to seek out the implicit mapping between the observed reverberant signal and the desired RAPs. Götz et al. combined the auditory-motivated Gammatone filterbank and convolutional neural network (CNN) to blindly estimate the reverberation time in dynamic acoustic conditions [22]. Zheng et al. proposed a robust estimation method for the reverberation time under the noisy conditions using gated convolutional recurrent network [23]. Lopez-Ballester et al. proposed the CNN-based algorithm applied in Internet of Things for blind estimation of the reverberation time [24]. Duangpummet et al. proposed the temporal-amplitude-envelope (TAE) based CNN for simultaneously estimating the STI and five essential RAPs, which is the first approach to achieve six-parameter blind estimation simultaneously [4]. Although many methods have been proposed, they provide only a single RAP or STI and require a massive amount of training data, which is rarely available in spite of the fact that the environments differs from the training data.

The aim of the work reported here was to devise a deterministic method that can derive parameters of an RIR model for estimating the STI and RAPs from an observed speech signal without any training dataset. Our study makes four important contributions to the contemporary knowledge frontier as follows:

- The effect of reverberation on the waveform of a signal transmitting in a sound field is clarified.
- An explicit closed-form solution is deterministically given using temporal power envelopes (TPEs) in the time domain.
- The TPE of an input signal is shown to be restored by using the closed-form solution based on the concept of the MTF.
- A deterministic estimation method for blindly estimating STI and five RAPs simultaneously is presented that has estimation accuracy comparable to that of the state-of-the-art method [4].

This paper is organized as follows. Section II briefly describes the STI and five essential RAPs. Section III reviews related works on the blind estimation of RIR, STI, and RAPs. Section IV describes the relationship between the RIR and observed reverberant signal to establish the corresponding mathematical derivations and the proposed blind estimation method. Section V describes the experimental setup and presents evaluation results. Section VI discusses the results, limitations, findings, and remaining works. Finally, Section VII summarizes the key points.

II. SPEECH TRANSMISSION INDEX AND ROOM ACOUSTIC PARAMETERS

Many indices and RAPs describing the acoustic characteristics of a sound field have been studied and standardized [5], [7], [13], [14], [25], [26]. The index and RAPs widely used by audio engineers and musicians are briefly introduced as follows.

A. SPEECH TRANSMISSION INDEX

The speech transmission index (STI) is used to predict the speech intelligibility and listening difficulty of a sound field. It was originally defined by Houtgast and Steeneken on the basis of the MTF. The MTF describes the characteristics of a transmission channel from the speaker to the listener by

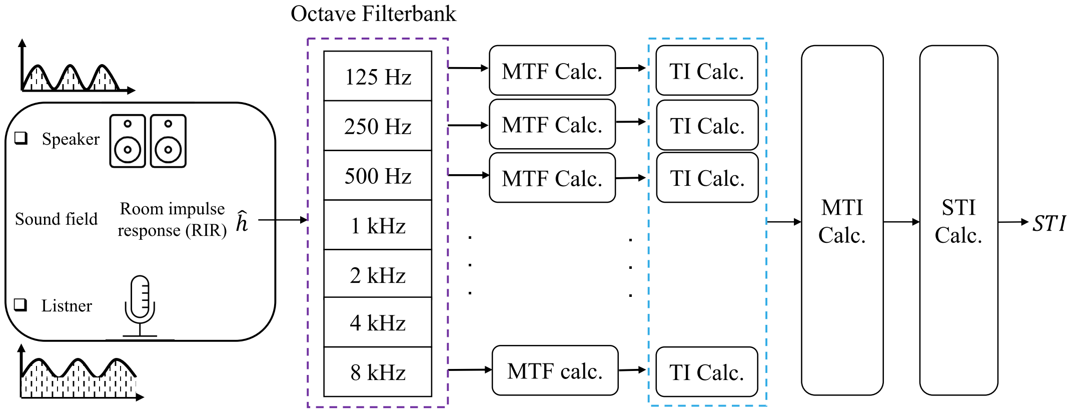

FIGURE 1. Block diagram of STI calculation.

the attenuation level of the modulation depth as a function of the modulation frequency [12], [15]. It is used to quantify the reverberation level. The higher the modulation depth, the lower the reverberation.

The MTF of a sound field can be defined as the ratio of the modulation spectrum of the RIR of the sound field to the total energy of the RIR:

$$m(f_m) = \frac{\int_0^\infty h^2(t)\exp(-j2\pi f_m t)dt}{\int_0^\infty h^2(t)dt}, \tag{1}$$

where $h(t)$ represents the RIR, and $m(f_m)$ represents the MTF at a modulation frequency f_m.

The calculation of the STI has been standardized in IEC 60268-16:2020 [13], as shown in Figure 1. First, the RIR passes through a seven-octave filterbank to calculate the MTF for each band using (1). Then, the modulation distortion ratio at 14 specific modulation frequencies is calculated as:

$$N_{k,i} = 10\log_{10}\left[\frac{m_k(f_{m,i})}{1-m_k(f_{m,i})}\right], \tag{2}$$

where $k = 1, 2, \ldots, 7$ and $i = 1, 2, \ldots, 14$. The 14 specific modulation frequencies can be determined, as shown in Table 1. Transmission index (TI) T at each octave band is determined:

$$T(k,i) = \begin{cases} 1, & N(k,i) > 15, \\ \dfrac{N(k,i)+15}{30}, & -15 \leq N(k,i) \leq 15, \\ 0, & N(k,i) < -15, \end{cases} \tag{3}$$

where the value of TI is limited to the range of −15 dB to 15 dB and normalized to the unit.

Next, the modulation transmission index (MTI) M for each band is calculated by averaging all $N(k,i)$ along with the specific modulation frequencies:

$$M_k = \frac{1}{14}\sum_{i=1}^{14} N(k,i). \tag{4}$$

Finally, the STI is derived by summing up the MTIs for the seven bands:

$$\text{STI} = \sum_{k=1}^{7} \text{Wgt}_k M(k), \tag{5}$$

where Wgt_k represents the weighting factor for each band, as shown in Table 2. The STI is a number ranging from 0 to 1. The higher the index, the more the speech intelligibility in the sound field.

B. REVERBERATION PARAMETERS

The reverberation time (T_{60}) and EDT RAPs are related to reverberation [7]. T_{60} is the most essential RAP in room acoustics as it characterizes the physical properties of a sound field in which energy is distributed within −60 dB. EDT characterizes the duration of the sound decay in a sound field within an initial −10 dB, so it emphasizes the more important contributions of direct sound and the early reflections of perceived reverberation. Both RAPs are derived from the energy decay curve of the RIR by using Schroeder's back integration method [27].

T_{60} is the 60-dB decay time calculated by line-fitting to the proportion of the energy decay curve between −5 dB and −35 dB and a linear extrapolation to −60 dB, since directly measuring the full 60 dB of decay is a practical limitation due to the presence of the background noise [14], [28]. EDT is the 60-dB decay time calculated by line-fitting to the proportion of the energy decay curve within −10 dB and a linear extrapolation to −60 dB. Figure 2 shows an example of deriving the reverberation parameters from curve fitting of the energy decay curve.

C. ENERGY PARAMETERS

Clarity (C_{80}), Deulitchkeit (D_{50}), and center time (T_s) are related to ratios of the RIR between the early energy provided by the first reflections and the energy from the late reverberation or the whole RIR. They are strongly correlated with the clarity impression in a given sound field.

TABLE 1. Fourteen specific modulation frequencies.

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
f_m (Hz)	0.63	0.80	1.00	1.25	1.60	2.00	2.50	3.15	4.00	5.00	6.30	8.00	10.00	12.50

TABLE 2. MTI octave-band weighting factor.

Band (Hz)	125	250	500	1k	2k	4k	8k
Wgt	0.129	0.143	0.114	0.114	0.186	0.171	0.143

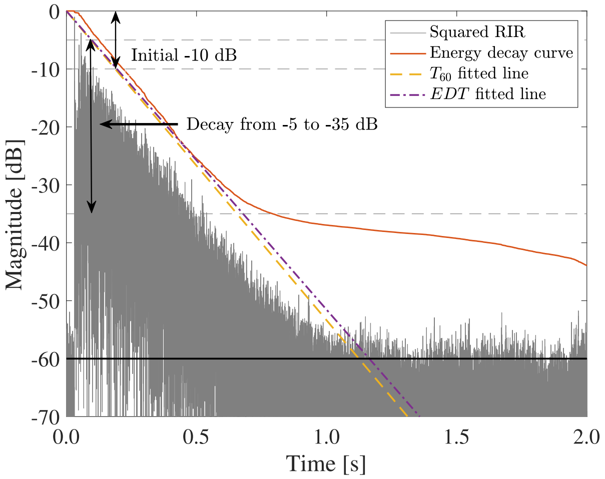

FIGURE 2. Calculation of reverberation parameters calculation from energy decay curve of RIR.

- C_{80} is the energy ratio of early-to-late arrival reflections, which characterizes the perception of transparency when music signal is transmitted in a sound field. It is defined as:

$$C_{80} = 10\log_{10}\frac{\int_0^{80\,ms} h^2(t)dt}{\int_{80\,ms}^{\infty} h^2(t)dt}, \tag{6}$$

  where $80\,ms$ denotes the time boundary between early reflections and late reverberation.
- D_{50} is another RAP correlated with clarity, it characterizes the subjective response to speech intelligibility in a sound field. It is the ratio of the sound energy in the first $50\,ms$ after the arrival of the direct sound to the total energy and is defined as:

$$D_{50} = \frac{\int_0^{50ms} h^2(t)dt}{\int_0^{\infty} h^2(t)dt} \times 100, \tag{7}$$

  where $50\,ms$ is the boundary of the early sound energy.
- T_s is the "center of gravity time" of decaying energy in a sound field. It characterizes the balance between clarity and reverberation and is related to speech intelligibility. It is expressed as:

$$T_s = \frac{\int_0^{\infty} h^2(t)tdt}{\int_0^{\infty} h^2(t)dt}. \tag{8}$$

III. RELATED WORKS

Proposed blind estimation methods for STI and RAPs are based on either analytical or learning-based approaches [4], [19], [20], [21], [22], [23], [24], [29], [30], [31], [32], [33], [34], [35], [36], [37]. Ones following the analytical approach achieve blind estimation by creating an explicit mapping between the observed reverberant signals and the desired parameters. Unoki et al. proposed two schemes based on the concept of MTF for estimating STI and T_{60} [20], [29]. They approximated an unknown RIR by using Schroeder's RIR model and then using a more precise model, namely the generalized RIR model, modified on the basis of Schroeder's RIR model [20]. They utilized the relationship between the modulation spectrum of the observed signal and the MTF to obtain the optimal parameters of the RIR model. The estimated STI and T_{60} are calculated from the MTF of the RIR model. Keshavarz et al. [21] used the proportional mapping between the autocorrelation of the original signal and the observed signal to devise a blind estimation method for T_{60}. The model of speech sequences proposed by Couvreur et al. [30] was combined with the maximum-likelihood estimation (MLE) to estimate T_{60}. A model of the energy decay curve was approximated by using the MLE described in [19], [31] to blindly estimate T_{60}.

Deep learning has been widely used in the blind estimation of STI and RAPs. Many artificial neural networks have been used to estimate a desired parameter (e.g., T_{60}, C_{80}, or STI) [32], [33], [34], [35], [36]. They proposed the idea of successive layers of representations to learn the relationship deeply between the output and the input with regard to a complicated problem. Early methods used a multi-layer perceptron (MLP) to learn the mapping between either the STI or T_{60} and the observed reverberant signal [10], [34], [35]. Santos and Folk used a recurrent neural network (RNN) to estimate STI, thereby enabling the neural network to learn more accurately the reverberant signal as the sequence vector [32]. Unfortunately, MLP-based methods suffered from insufficient training since the number of total parameters grows fairly high. They also need tedious feature extraction. Subsequent methods based on a convolutional neural network (CNN) are able to train a number of reverberant speech signals for use in efficiently estimating the STI without features extraction, which is referred to as the "end-to-end model" [33], [36]. With regard to T_{60} estimation, many learning-based approaches have been

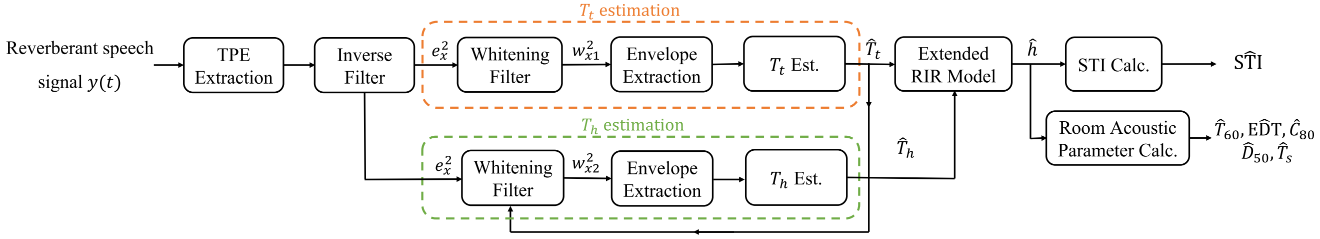

FIGURE 3. Block diagram of the proposed method.

evaluated in the Acoustic Characterization of Environment (ACE) challenge [38]. For example, Parada et al. proposed a non-intrusive method based on extracting pre-frame features by using an RNN [39]. Gamper and Tashev devised a CNN with time-frequency features appearing in a spectrogram [40].

Recently, Duangpummet et al. proposed an MTF-based scheme for simultaneously estimating STI and RAPs. A CNN incorporating the concept of the MTF learns the nonlinear mapping between the parameters of the RIR model and the reverberant speech signal [4]. They used the CNN to train the seven-octave bands of temporal amplitude envelopes (TAEs) of a massive number of reverberant speech signals synthesized using simulated RIRs. They achieved the highest accuracy for simultaneously estimating STI and five RAPs with real-time implementation.

However, the current methods can estimate only a single parameter [19], [20], [21], [22], [24], [29], [30], [31], [32], [33], [36], [40]. Although the MTF-based CNN method can estimate multiple parameters, it is limited to the training data used to derive the model, the same as the other learning-based methods [4], [32], [33], [34], [35], [36], [37]. The efficiency of the trained models is lower when the actual environments differ from the training data. The models are also difficult to optimize because they are untraceable implicit models and have a vast number of trainable parameters. Therefore, we propose using an analytical method for blindly estimating the STI and five RAPs, T_{60}, EDT, C_{80}, D_{50}, and T_s, simultaneously. We incorporate a stochastic RIR model, namely an extended RIR model, into the relationships between the temporal power envelope (TPE) of an observed signal and the RIR model to derive the method.

IV. PROPOSED METHOD

Our proposed blind estimation method that alternates to estimate the parameters of the RIR, namely the alternating estimation strategy (AES), is shown in Fig. 3.

Figure 3 illustrates the signal processing flow of the proposed method. The input is the observed reverberant signal $y(t)$. The TPE of $y(t)$ is extracted to estimated the one parameter of the extended RIR model used to approximates an unknown RIR encompassed by orange dashed lines, called "T_t estimation" section. Then, we pass by this estimated parameter into the second stage to estimate another parameter of the extended RIR model called "T_h estimation" section encompassed by green dashed line. Next, these two estimated parameters are used to synthesize the estimated RIR according to the extended RIR model. Finally, STI and RAPs can be calculated from the estimated RIR using (1) - (8) according to IEC 60268-16:2020 and ISO 3382:2008 standards [13], [14].

The extended RIR and TPE models of input and output signals in a reverberation process are introduced. The whitening and inverse filtering and objective function that the AES is based on are then described.

A. EXTENDED RIR MODEL

For blindly estimating STI and RAPs, we can observe only reverberant signals. Thus, we model an observed signal in a reverberant room as the convolution of an original signal and RIR. Since blind estimation is an ill-posed problem, we need to model an unknown RIR that connects the original signal and the observed reverberant signal.

Schroeder's RIR model is a simple decay model and is commonly used to approximate an unknown measured RIR [16]. Schroeder's RIR model is defined as:

$$h(t) = e_h(t)c(t) = a\exp\left(-\frac{6.9}{T_{60}}\right)c(t), \tag{9}$$

where e_h is the temporal amplitude envelope (TAE) of the RIR, $c(t)$ is the carrier signal as the white Gaussian noise (WGN) that acts as random variables, and a is the gain factor. Schroeder's RIR model is sufficient to represent the room acoustics of a geometrically simple enclosure, such as a vacant rectangular-shaped room without furniture and occupants. However, realistic spaces commonly have complicated geometrical shapes and include different types of furnishings, as illustrated in Fig. 4. In realistic spaces, Schroeder's RIR model does not match the actual RIR due to the lack of modeling of the onset transition of the measured RIR.

Figure 4 illustrates an example of a complicated-shaped room with furniture. In the illustration, the sound source (loudspeakers) are placed at the main room, nearby a sofa. Two listeners locate at two smaller rooms, connecting the main room with two doors. The RIR measured at one of the two listener's locations is illustrated in Fig. 4. At some listeners' positions, the sound wave takes time to travel or reflect with walls and objects. Consequently, sharp of the RIR slightly changes depending on waves traveling in a

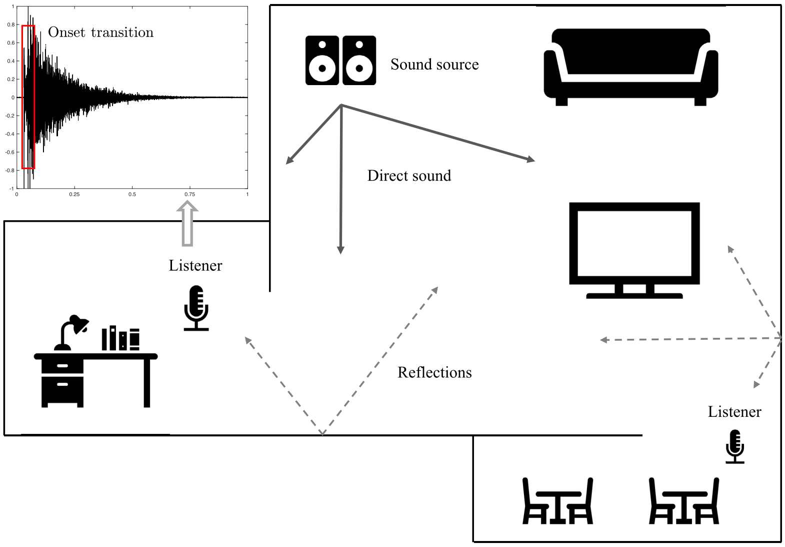

FIGURE 4. Illustration of complicated-shape room and its room impulse response.

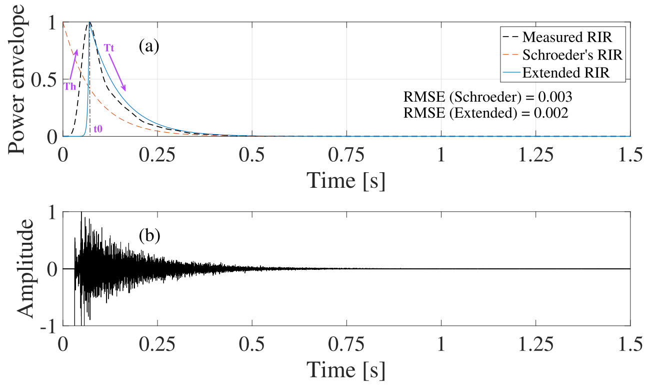

FIGURE 5. Fittings of two RIR models to measured RIR: (a) temporal power envelope of RIRs and (b) corresponding RIR.

given room. Because Schroeder's RIR model lacks the onset transition of an actual RIR in a complicated-shape room, the model mismatch occurs.

Figure 5 shows a comparison between the fits of the temporal power envelope of two RIR models with the measured RIR. The extended RIR model, the more accurate RIR model, mitigates the limitation of Schroeder's RIR model by adding a parameter to control the exponential rising envelope to approximate the onset transition of the measured RIR [4], [41]. The extended RIR model is defined as:

$$h_{ext}(t) = e_h(t)c(t) = \begin{cases} a\exp(6.9t/T_h)c(t), & t < 0 \\ a\exp(-6.9t/T_t)c(t), & t \geq 0 \end{cases} \tag{10}$$

$$h(t) = h_{ext}(t - t_0), \quad t_0 \geq 0 \tag{11}$$

where T_h and T_t denote the parameters controlling the exponential rising and decaying envelopes of the RIR; t_0 is introduced to promise a casual system and stable impulse response, i.e., $h(t) = 0, t < 0$. Here, t_0 is assumed to be equal to T_h. Therefore, the T_h and T_t parameters control the shape of the whole RIR envelope.

Table 3 shows the results of an ablation study that evaluates the suitability of both Schroeder's and the extended RIR model for realistic RIRs. We modeled realistic RIRs using both models and calculated the STI and five RAPs from the modeled and actual RIRs. The fitness of the RIR models was assessed by calculating the root-mean-squared error of STI and RAPs between the calculated and the ground-truth values from the modeled RIRs.

B. TEMPORAL POWER ENVELOPE MODEL

We mathematically model the reverberation and de-reverberation processes on the basis of the concept of the MTF. This TPE model (TPEM) is then used to develop a blind estimation strategy.

1) REVERBERATION PROCESS

Since we assume the sound field to be a linear time-invariant system, a reverberant signal observed in the sound field can be modeled as the convolution of the original signal and the RIR:

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau, \tag{12}$$

where $y(t)$, $x(t)$, and $h(t)$ denote the reverberant signal, the original signal, and the RIR, respectively. The symbol "$*$" denotes the convolution operation. $x(t)$ and $h(t)$ are assumed as the modulation of the TAEs and the carrier signals as:

$$x(t) = e_x(t)c_x(t), \tag{13}$$

$$h(t) = e_h(t)c_h(t), \tag{14}$$

$$\delta(t) = \langle c(t)c(t-\tau)\rangle, \tag{15}$$

where $e_x(t)$ and $e_h(t)$ are the TAEs of $x(t)$ and $h(t)$, respectively, $\delta(t)$ denotes the Dirac delta function, $c_x(t)$ and $c_h(t)$

TABLE 3. Accuracy (RMSE) of calculated STI and RAPs from the Schroeder's and the extended RIR model regarding the actual RIRs.

	STI	T_{60}	EDT	C_{80}	D_{50}	T_s
Schroeder's model [16]	0.166	0.844	0.789	4.066	17.836	0.075
Extended model	0.039	0.056	0.285	2.530	15.235	0.058

are mutually independent WGN carriers that act as random variables, and $\langle\cdot\rangle$ denotes the ensemble average [42].

The ensemble average of the square of the reverberant signal $y^2(t)$ is determined by [43]:

$$\begin{aligned}\langle y^2(t)\rangle &= \left\langle\left[\int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau\right]^2\right\rangle\\ &= \int_{-\infty}^{\infty} e_x(\tau_1)e_h(t-\tau_1)d\tau_1 \int_{-\infty}^{\infty} e_x(\tau_2)e_h(t-\tau_2)d\tau_2\\ &\quad\times\langle c_x(\tau_1)c_x(\tau_2)\rangle\langle c_h(t-\tau_1)c_x(t-\tau_2)\rangle\\ &= \int_{-\infty}^{\infty} e_x^2(t)e_h^2(t-\tau)d\tau = e_y^2(t). \qquad (16)\end{aligned}$$

Hence, the temporal power envelope (TPE) of a reverberation process can be modeled as:

$$e_y^2(t) = e_x^2(t) * e_h^2(t), \qquad (17)$$

where $e_y^2(t)$ is the TPE of the reverberant signal, $e_x^2(t)$ is a TPE of the input signal, and the asterisk symbol "$*$" denotes the convolution operation. The TPE of an observed reverberant signal $y(t)$ is extracted using:

$$e_y^2(t) = \mathrm{LPF}\big[|y(t) + j\cdot\mathrm{Hilbert}(y(t))|\big]^2, \qquad (18)$$

where LPF is a Butterworth low-pass filter with a cut-off frequency of 30 Hz, and Hilbert denotes the Hilbert transform.

Lemma 1: We consider the TPE of the original signal $e_x^2(t)$ from the simplest case, as $e_x^2(t)$ is a single-tone amplitude-modulation (AM) signal:

$$e_x^2(t) = C\cos(2\pi f_m t + \phi), \quad t\in[0,T] \qquad (19)$$

where C is the constant gain, f_m is the modulation frequency, ϕ is the phase, and T is the time interval. By using (10), (11), (17), and (19), the corresponding TPE of the reverberant signal can be expressed as a closed form:

$$\begin{aligned}e_y^2(t) &= \frac{Ca^2T_h\exp\left[13.8\left(\frac{t}{T_h}-1\right)\right]}{13.8^2+(2\pi f_mT_h)^2}\Bigg\{13.8\Big[\exp(13.8)\cos(\phi)\\ &\quad - cos(2\pi f_mT_h+\phi_k)\Big]+2\pi f_mT_h\Big[\sin(2\pi f_mT_h+\phi)\\ &\quad -\exp(13.8)\sin(\phi)\Big]\Bigg\}, \quad t\in[0,T_h) \qquad (20a)\end{aligned}$$

$$\begin{aligned}e_y^2(t) &= \frac{Ca^2T_t}{\sqrt{13.8^2+(2\pi f_mT_h)^2}}\\ &\quad\times\Bigg[\cos(2\pi f_mt+\phi)-\exp\left[-13.8\left(\frac{t-T_h}{T_t}\right)\right]\\ &\quad\times\cos(2\pi f_mT_h+\phi+\theta)\Bigg], \quad t\in[T_h, T-T_t] \qquad (20b)\end{aligned}$$

$$\begin{aligned}e_y^2(t) &= \frac{Ca^2T_t\exp\left(-13.8\frac{t}{T_t}\right)}{13.8^2+(2\pi f_mT_t)^2}\Bigg\{13.8\Big[\exp(13.8)\cos(\phi)\\ &\quad -\cos(2\pi f_mT_t-\phi)\Big]+2\pi f_mT_t\Big[\sin(2\pi f_mT_t-\phi)\\ &\quad +\exp(13.8)\sin(\phi)\Big]\Bigg\}, \quad t\in(T-T_t, T] \qquad (20c)\end{aligned}$$

where $\theta = \tan^{-1}\left(\frac{-2\pi f_mT_t}{13.8}\right)$.

Proof: The detailed derivation is presented in Appendix A. □

We use the superposition principle of an LTI system to extend (19) and (20) to model the TPEs of signals in actual environments. Equation (19) is extended to:

$$e_x^2(t) = \sum_{k=0}^{K} C_k\cos(2\pi f_{m,k}t+\phi_k), \quad t\in[0,T] \qquad (21)$$

where k is the index of K components. This equation can be used to model the TPE of a random original signal. The TPE of a clean signal $x(t)$ can be extracted as:

$$e_x^2(t) = \mathrm{LPF}\big[|x(t)+j\cdot\mathrm{Hilbert}(x(t))|\big]^2. \qquad (22)$$

Since (22) filters out the noise, we assume that the TPE of a clean signal is affected only by reverberation. Thus, the TPE of an observed reverberant signal $e_y^2(t)$ can be determined as:

$$\begin{aligned}&e_y^2(t)\\ &= \sum_{k=0}^{K}\frac{C_ka^2T_h\exp\left[13.8\left(\frac{t}{T_h}-1\right)\right]}{13.8^2+(2\pi f_{m,k}T_h)^2}\\ &\quad\times\Big\{13.8\Big[\exp(13.8)\cos(\phi_k)-\cos(2\pi f_{m,k}T_h+\phi_k)\Big]\\ &\quad+2\pi f_{m,k}T_h\\ &\quad\times\Big[\sin(2\pi f_{m,k}T_h+\phi_k)-\exp(13.8)\sin(\phi_k)\Big]\Big\}, \quad t\in[0,T_h)\\ &\qquad (23a)\end{aligned}$$

$$\begin{aligned}&e_y^2(t)\\ &= \sum_{k=0}^{K}\frac{C_ka^2T_t}{\sqrt{13.8^2+(2\pi f_{m,k}T_h)^2}}\\ &\quad\Bigg[\cos(2\pi f_{m,k}t+\phi_k)-\exp\left[-13.8\left(\frac{t-T_h}{T_t}\right)\right]\\ &\quad\times\cos(2\pi f_{m,k}T_h+\phi_k+\theta_k)\Bigg], \quad t\in[T_h, T-T_t] \qquad (23b)\end{aligned}$$

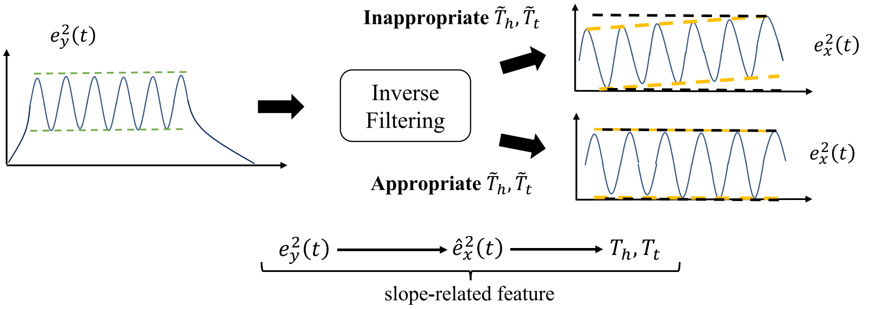
$e_y^2(t)$
Inverse Filtering
Inappropriate $\tilde{T}_h, \tilde{T}_t$
Appropriate $\tilde{T}_h, \tilde{T}_t$
$e_x^2(t)$
$e_x^2(t)$
$e_y^2(t) \longrightarrow \hat{e}_x^2(t) \longrightarrow T_h, T_t$
slope-related feature

FIGURE 6. Illustration of concept of slope-related feature described in Section IV-B.

$$
\begin{aligned}
&e_y^2(t)\\
&=\sum_{k=0}^{K}\frac{C_k a^2 T_t \exp\left(-13.8\frac{t}{T_t}\right)}{13.8^2+(2\pi f_{m,k}T_t)^2}\\
&\quad\times\left\{13.8\left[e^{13.8}\cos(\phi_k)-\cos(2\pi f_{m,k}T_t-\phi_k)\right]+2\pi f_{m,k}T_t\right.\\
&\quad\times\left.\left[\sin(2\pi f_{m,k}T_t-\phi_k)+e^{13.8}\sin(\phi_k)\right]\right\},\ t\in(T-T_t,T]
\end{aligned}
\tag{23c}
$$

where $f_{m,k}$, C_k, and ϕ_k are the modulation frequency, gain factor, and phase at k-th component, respectively. θ_k is equal to $\tan^{-1}\left(\frac{-2\pi f_{m,k}T_t}{13.8}\right)$. The value 13.8 comes from 2 × 6.9, the value used in Schroeder's RIR model and the extended RIR model to control the exponential rate.

Equations (20a) and (23a) represent the part the reverberation commences, as the modulation depth increases gradually. In this case, only T_h affects the waveform of the TPE. Equations (20b) and (23b) represent the part the reverberation constantly impose the effect on the waveform of the TPE, as the modulation depth keeps constant. Both T_h and T_t affect the waveform of the TPE together. Equations (20c) and (23c) represent the part the reverberation terminates, as the modulation depth declines gradually. Only T_t has the effect on this part.

2) DEREVERBERATION PROCESS

The TPE of an observed reverberant signal $e_y^2(t)$ can be restored by inverse filtering. The restored TPE $e_x^2(t)$ can be determined in the z-domain as:

$$
E_x(z)=\frac{E_y(z)}{E_h(z)}. \tag{24}
$$

Given the TPE of a reverberant signal $e_y^2(t)$ expressed by (20), the envelope of the recovered TPE from $e_y^2(t)$ can be expressed in closed form based on the concept of inverse filtering. When $e_y^2(t)$ is restored, the transition regions of the modulation depth in the waveform, i.e., the signal modeled by (20a) and (20c), are completely vanished. The restored TPE $e_x^2(t)$ can be constructed according to (20b), as expressed in (24):

$$
e_{x,upr}^2(t)=\frac{CT_t}{\tilde{T}_t}\sqrt{\frac{1+\left(\frac{2\pi f_m\tilde{T}_t}{13.8}\right)}{1+\left(\frac{2\pi f_m T_t}{13.8}\right)}}\,\frac{u(t)-\psi(t,T_h,T_t)}{u(t)-\psi(t,\tilde{T}_h,\tilde{T}_t)}, \tag{25a}
$$

$$
e_{x,lwr}^2(t)=\frac{CT_t}{\tilde{T}_t}\sqrt{\frac{1+\left(\frac{2\pi f_m\tilde{T}_t}{13.8}\right)}{1+\left(\frac{2\pi f_m T_t}{13.8}\right)}}\,\frac{-u(t)-\psi(t,T_h,T_t)}{u(t)+\psi(t,\tilde{T}_h,\tilde{T}_t)}, \tag{25b}
$$

where $\psi(t,T_h,T_t) = \exp\left[\frac{-13.8(t-T_h)}{T_t}\right]\cos(2\pi f_m T_h+\phi+\theta)$, $u(t)$ is the unit-step function, and $e_{x,upr}^2(t)$ and $e_{x,lwr}^2(t)$ are upper and lower envelopes, respectively. The $\tilde{T}_h$ and $\tilde{T}_t$ parameters are used to carry out the restoration.

Equations (25a) and (25b) indicate that when $T_h = \tilde{T}_h$ and $T_t = \tilde{T}_t$ hold, $e_{x,upr}^2(t) = C$ and $e_{y,lwr}^2(t) = C$ hold. In that case, the envelopes are invariant with time, whereas when $T_h \neq \tilde{T}_h$ and $T_t \neq \tilde{T}_t$, the envelopes are time-varying. These time-varying envelopes can be simply approximated as a first-order polynomial:

$$
e_{x,upr}^2(t)=S_{upr}t+b_{upr},\ \text{and}\ e_{x,lwr}^2(t)=S_{lwr}t+b_{lwr}, \tag{26}
$$

where S_{upr} and S_{lwr} are the slopes of the envelopes, and b_{upr} and b_{lwr} are constant factors.

We extend (25a) and (25b) by using the superposition principle to model the envelopes of the restored TPE of any signal. They are given as:

$$
\begin{aligned}
&e_{x,upr}^2(t)\\
&=\sum_{k=0}^{K}\frac{C_k T_t}{\tilde{T}_t}\sqrt{\frac{1+\left(\frac{2\pi f_{m,k}\tilde{T}_t}{13.8}\right)}{1+\left(\frac{2\pi f_{m,k}T_t}{13.8}\right)}}\,\frac{u(t)-\psi_k(t,T_h,T_t)}{u(t)-\psi_k(t,\tilde{T}_h,\tilde{T}_t)},
\end{aligned}
\tag{27a}
$$

$$
\begin{aligned}
&e_{x,lwr}^2(t)\\
&=\sum_{k=0}^{K}\frac{C_k T_t}{\tilde{T}_t}\sqrt{\frac{1+\left(\frac{2\pi f_{m,k}\tilde{T}_t}{13.8}\right)}{1+\left(\frac{2\pi f_{m,k}T_t}{13.8}\right)}}\,\frac{-u(t)-\psi_k(t,T_h,T_t)}{u(t)+\psi_k(t,\tilde{T}_h,\tilde{T}_t)},
\end{aligned}
\tag{27b}
$$

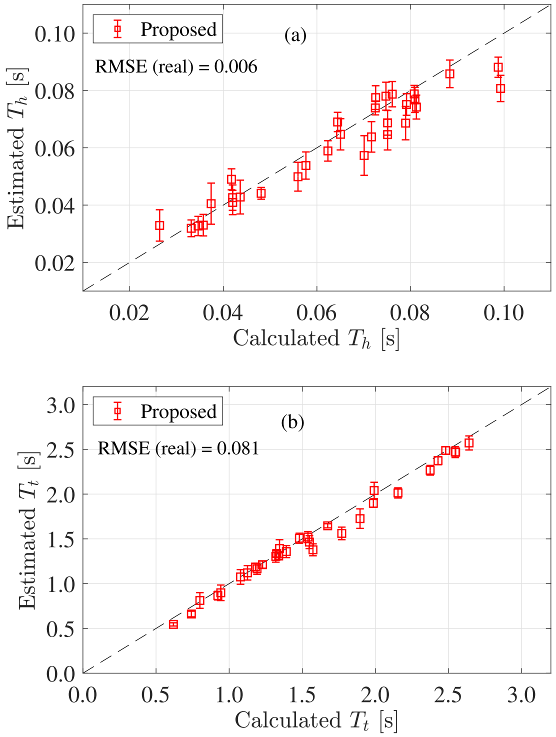

FIGURE 7. Estimated vs. calculated parameters of extended RIR model: (a) T_h, (b) T_t.

TABLE 4. Estimation accuracy (RMSE) of parameters of extended RIR model by proposed and previous methods [4].

	T_h	T_t
TAE-CNN	0.087	0.193
Proposed	0.006	0.081

where $\psi_k(t, T_h, T_t) = \exp\left[\frac{-13.8(t-T_h)}{T_t}\right]\cos(2\pi f_{m,k} T_h + \phi_k + \theta_k)$. When using appropriate $\tilde{T}_h$ and $\tilde{T}_t$ values to carry out the restoration, i.e., $T_h = \tilde{T}_h^*$ and $T_t = \tilde{T}_t^*$ hold, each single-tone component k in (27) stays constant so that $e_{x,upr} = \sum_{k=0}^{K} C_k$ and $e_{x,lwr} = -\sum_{k=0}^{K} C_k$ hold. In this case, the summation of the slopes for each component is minimized; i.e., $\sum_{k=0}^{K} S_{k,upr} = 0$ and $\sum_{k=0}^{K} S_{k,lwr} = 0$.

if inappropriate $\tilde{T}_h$ and $\tilde{T}_t$ values are used for restoration, the summation of the slopes for each component must be a time-varying function; i.e., $\sum_{k=0}^{K} S_{k,upr} = f_{upr}(t)$ and $\sum_{k=0}^{K} S_{k,lwr} = f_{lwr}(t)$. Since a higher-order polynomial has many degrees of freedom, which makes it difficult to optimize, we use the first-order polynomials, as in (26), is used to approximate the time-varying envelopes in (27).

We thus far have established the relationship between the observed reverberant signal and the parameters of the RIR. Figure 6 illustrates the concept of this relationship and corresponding *slope-related feature*. A blind estimation strategy based on the slope-related feature is presented in Section IV-C to IV-E.

C. INVERSE FILTER

We estimate the parameters of the RIR by using the model as described in Sec IV-B that generates a TPE of the signals. The discrete TPE of the RIR model (Eqs. 10 and 11) is obtained by sampling the continuous-time RIR. Then, z-transform is applied. Thus, an infinite-impulse response (IIR) of the RIR can be defined as:

$$E_h(z) = \frac{a^2(\alpha - \beta)}{(1 - \alpha z^{-1})\,(1 - \beta z^{-1})} \tag{28}$$

where $\alpha = \exp(-13.8/T_t f_s)$, $\beta = \exp(13.8/T_h f_s)$, and f_s is the sampling frequency. The IIR-inverse filter can be defined as:

$$E_{h,inv}(z) = E_h^{-1}(z). \tag{29}$$

We use all possible $\tilde{T}_h$ and $\tilde{T}_t$ sets to implement the inverse filtering. $\tilde{T}_h$ ranges from 0.01 to 0.15 s at step size 0.001 s. $\tilde{T}_t$ ranges from 0.35 to 3.5 s at step size 0.01 s.

D. WHITENING FILTER

A whitening filter, the key of the AEM, was designed based on linear prediction coding [44], [45]. As mentioned in Section IV-B, we use first-order polynomials to approximate the time-varying envelopes of the restored TPE, which requires using relatively even envelopes to calculate the slopes. However, the complex waveform of the TPE of an actual reverberant signal makes it difficult to meet this requirement. Hence, we employ the whitening filter to acquire even envelopes of the waveform. The frame-based whitening filter whitens a restored TPE with a complex waveform into a pulse train that has even envelopes suitable for calculating the slopes from (26).

The restored TPE at each frame (frame length is n) is regarded as autoregressive (AR) mode and rewritten as:

$$e_x^2[n] = -\sum_{i=1}^{p} \sigma_i e_x^2[n-i] + w_x^2[n], \tag{30}$$

$$w_x^2[n] = \sum_{i=0}^{p} \sigma_i e_x^2[n-i],\ \ W(z) = \sum_{i=0}^{p} \sigma_i z^{-i}, \tag{31}$$

where σ_i is the optimal predictor, $\sigma_0 = 1$, p is the number of the predictor order, $w_x^2[n]$ is the whitened restored TPE, and $W(z)$ is the frame-based whitening filter [44], [46]. Here, $n = 128$. Since (27a) and (27b) imply that the reverberation smears over all frequencies, where $w_x^2[n] \in e_x^2[n]$, the reverberation also smears into $w_x^2[n]$. Therefore, we assert that whitening preserves the reverberation information, which can be used for blind estimation.

The optimal predictor σ_i can be determined by using *the normal equations*, as used elsewhere [47]. $R_{e_x^2}(p)$ is defined to be the autocorrelation sequences of $e_x^2[n]$ as:

$$R_{e_x^2}(p) = \mathrm{E}[e_x^2[n]\overline{e_x^2}[n-p]], \tag{32}$$

where E denotes the expectation operation, and “$\bar{\cdot}$ ” denotes conjugation. The optimal predictor is given by:

$$\mathbf{R}\boldsymbol{\sigma} = -\mathbf{r}, \quad \text{and} \quad \boldsymbol{\sigma} = -\mathbf{R}^{-1}\mathbf{r}, \tag{33}$$

TABLE 5. Comparison of accuracy (RMSE) between previous and proposed methods [4], [20].

	STI	T_{60}	EDT	C_{80}	D_{50}	T_s
MTF-based [20]	0.060	–	–	–	–	–
TAE-CNN [4]	0.040	0.393	0.259	2.038	12.143	0.037
Proposed	0.037	0.067	0.256	2.309	14.303	0.052

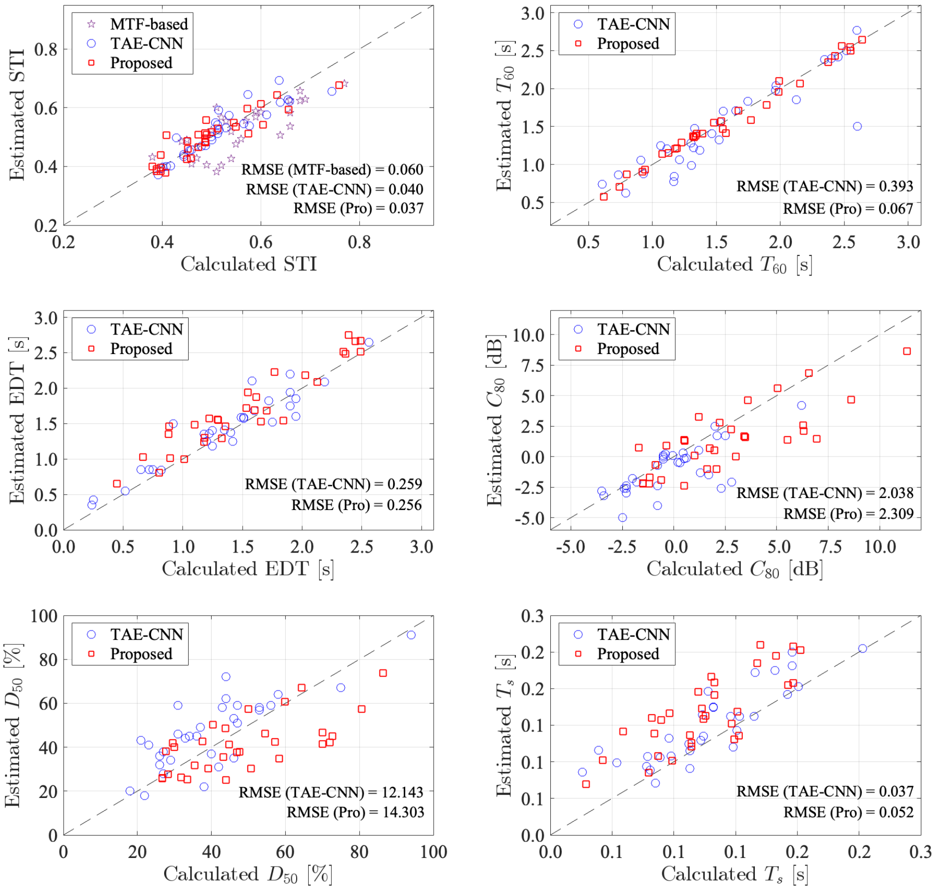

FIGURE 8. STI and RAPs from reverberant speech signals: (a) STI, (b) T_{60}, (c) EDT, (d) C_{80}, (e) D_{50}, and (f) T_s. "□", "∘", and "⋆" denote the estimated values with proposed and two previous methods: TAE-based CNN method (TAE-CNN) [4] and MTF-based method (MTF-based) [20], respectively. The Black dashed line represents the ground truths calculated from the RIRs.

where **R** is a Toeplitz matrix of $R_{e_x^2}$, and **r** is the cross-correlation vector; $\left[R_{e_x^2}[1]\, R_{e_x^2}[2] \cdots R_{e_x^2}[p]\right]^T$. " T " denotes transposition. We increase the order of the predictor to flatten the whitened signal so that the actual signals comply with the AR mode. Here, the order of the predictor p is set to 20. The Levinson-Durbin algorithm is then utilized to optimize the computation of the normal equations [44], [48].

E. OBJECTIVE FUNCTION

The optimal $\widehat{T}_t$ and $\widehat{T}_h$ are obtained by optimizing the following functions. $\widehat{T}_t$ is determined from all possible sets of $\tilde{T}_h$ to satisfy (34) as:

$$\widehat{T}_t = \underset{T_t}{\text{med}}\Big\{\underset{T_h,\{\tilde{T}_t\}}{\text{argmin}}\big[\log_{10}\big(|S_{upr}|\big) + \log_{10}\big(|S_{lwr}|\big)\big]\Big\}, \quad (34)$$

where "med" denotes the median operation. Thus, $\widehat{T}_h$ is obtained by substituting $\widehat{T}_t$ into (29) to satisfy (35) as:

$$\widehat{T}_h = \text{argmin}\big\{\log_{10}\big(|S_{upr}|\big) + \log_{10}\big(|S_{lwr}|\big)\big\}. \quad (35)$$

We synthesized the estimated RIR $\widehat{h}$ by modulating the WGN carrier with the extended RIR model (10) from the optimal $\widehat{T}_t$ and $\widehat{T}_h$. Fig. 9 shows the TPE of the reconstructed RIR and the measured RIR. The estimated STI and five RAPs are thus derived from the estimated RIR by using (1) - (8) in accordance with IEC 60268-16:2020 and ISO 3382:2008 standards [13], [14].

V. EXPERIMENTS AND RESULTS

We evaluated the proposed method using reverberant speech signals to determine whether it can estimate STI and RAPs appropriately.

TABLE 6. Correlation coefficients between the estimated values and ground-truths.

	STI	T_{60}	EDT	C_{80}	D_{50}	T_s
TAE-CNN [4]	0.913	0.918	0.873	0.943	0.903	0.836
Proposed	0.908	0.993	0.945	0.794	0.680	0.797

A. EXPERIMENTAL SETUP

We carried out simulations by using reverberant speech signals synthesized by convoluting speech signals with RIRs from the SMILE dataset, which contains 43 measured RIRs [49]. The measured RIRs are single-channel recorded. Complete information about the measured RIRs is available elsewhere [20]. The speech signals were taken from the ATR dataset [50]. They were ten long Japanese sentences uttered by ten speakers (five males and five females) from the ATR dataset [50] and had been single-channel recorded with 16-bit quantization and 20-kHz sampling frequency. The experimental framework was based on the extended RIR model and conventional signal processing techniques. The root-mean-square error (RMSE) and Pearson correlation coefficient were used as the evaluation metrics. The error was calculated from the difference between the estimated results and the ground truths. The performance of the proposed method was compared with that of previous methods by using the same RIR and speech datasets [4], [20].

B. EVALUATION FOR PARAMETERS OF RIR MODEL

We conducted the simulations using reverberant speech to determine whether the proposed method can effectively estimate the parameters of the RIR model. Detailed information about evaluation conditions can be found in [20], including various environments such as concert halls, lecture halls, and office rooms where people occupy in daily lives.

Table 4 shows the accuracy (RMSE) of T_h and T_t estimated using the proposed and previous methods [4]. Fig. 7 shows the estimated versus calculated parameters (T_h and T_t) of the extended RIR model. The horizontal axis indicates the parameters fitted from the measured RIRs directly, and the vertical axis indicates the parameters estimated with the proposed method.

The results show that the proposed method can appropriately estimate the parameters of the extended RIR model. Then, estimated parameters and (10) and (11) were used to synthesized the approximated RIR. Fig. 9 shows a comparison between the approximated RIR reconstructed using the proposed method and the meaured RIR in terms of their TPEs, in which the RMSE was 0.041.

C. EVALUATION FOR ROOM ACOUSTIC PARAMETERS

Figure 8 shows the estimated STI and five room-acoustic parameters from reverberant speech signals in realistic reverberant environments. The horizontal axis indicates the parameters calculated from the RIRs, and the vertical axis indicates the parameters estimated from the speech signals.

Table 5 compares estimation accuracy between the previous and proposed methods. The RMSEs of the estimated STI and T_{60} reveal that the proposed method outperformed the previous methods. With regard to EDT, the estimated results closely approach the ground-truth results calculated using the standard method [14]. With regard to C_{80}, D_{50}, and T_s, the RMSEs were 2.31 dB, 14.30 %, and 0.052, respectively. Table 6 shows the correlation coefficient between the estimated and calculated values for the previous and proposed methods, in which correlation coefficient results keep consistent with the RMSE accuracy. The proposed method had virtually the same accuracy as the previous methods for STI, T_{60}, EDT, and T_s, whereas, it had lower accuracy for C_{80}, and D_{50}. The just noticeable difference (JND) and standard deviation of the RAPs are shown in Table 7. There were noticeable outliers for C_{80}, D_{50}, and T_s, possibly because the carrier signal of the measured RIRs did not strictly match the WGN [52], [53]. Since the method proposed by Duangpummet [4] is considered the state-of-the-art for blindly and simultaneously estimating STI and RAPs, we assert that the proposed method is similarly effective.

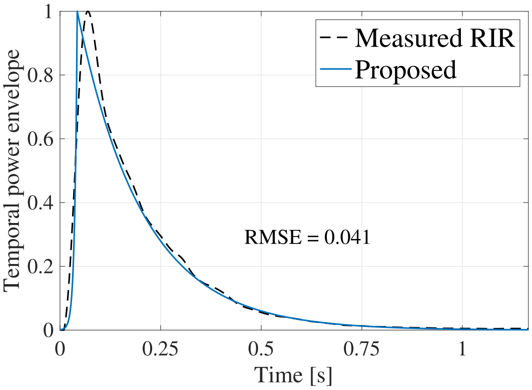

FIGURE 9. TPE of approximated RIR reconstructed using the proposed method and measured RIR.

VI. DISCUSSION

In the previous section, we described our evaluation of the proposed method and compared its performance with that of the previous methods [4], [20]. The results show that the proposed method outperforms or performs at the same level as the previous ones. In this case, we discuss the advantages and disadvantages of the proposed method and remaining issues concerning the scope of this work.

Our aims is to clarify the effects of the RIR on the envelope of an original signal and to reveal the relationship between the RIR and the observed signal by formulating closed-form solutions for the reverberation and dereverberation processes based on the MTF concept. We approximate an unknown RIR and the observed signal to identify the connection between them. We thus need to consider which model is best for modeling the RIR and how to investigate the effects of the

TABLE 7. Comparison between just noticeable difference (JND) of RAPs and standard deviation (SD) of estimated error [14], [51].

	STI	T_{60}	EDT	C_{80}	D_{50}	T_s
JND	0.03	5.0%	5.0%	1.0 dB	5%	10 ms
SD (TAE-CNN) [4]	0.05	9.4%	10.5%	2.7 dB	14%	45 ms
SD (Proposed)	0.03	4.0%	4.0%	1.8 dB	9%	28 ms

RIR model's parameters on the waveform of the reverberant signal.

We used an extended impulse response model modified on the basis of Schroeder's impulse response model to represent an unknown RIR. It overcomes the limitation of Schroeder's model in terms of the modeling of the onset transition of the RIR in a complicated-shape room. We formulate the closed-form expression for an observed signal using the extended model based on the basis of its TPE. Hence, we understand how the RIR affects the waveform of a signal transmitted in a sound field. Moreover, we found the connection between the parameters of the RIR model and the TPE of a restored signal restored from the reverberant signal, i.e., the slope-related feature.

We also presented a deterministic method based on the slope-related feature for blindly estimating STI and five RAPs. Instead of the learning-based approach used in the previous method [4], an analytical approach is used. The proposed method achieved estimation accuracy comparable to that of the best existing blind method. First, the parameters of the RIR model, i.e., T_h and T_t, were estimated from the observed signal by extracting the TPE. The estimated results indicate that the estimation error of T_h varies more than that of T_t, which is consistent with the previous method. Because T_h has a substantially smaller scale than T_t, it is still a challenge to estimate the parameters at different resolutions. Next, the estimated parameters were used to synthesize the approximated RIR used to blindly estimate the STI and RAPs. The results show that the proposed method is effective in blindly estimating these acoustical parameters.

Additionally, we found that the estimated C_{80}, D_{50}, and T_s deviated more than the other acoustical parameters. We hypothesized that the carrier signal used for synthesis, i.e., the WGN carrier, did not completely match the temporal-fine structure of an actual RIR, as discussed elsewhere [52], [53].

Furthermore, the computational time and complexity of the proposed method need to be considered. Since the whitening filter was designed based on linear prediction coding, it takes much time to compute the optimal predictors. It thus might be inappropriate for real-time applications. Hence, the quasi-real-time implementation of the proposed method needs further investigation. In addition, some parameters controlling the proposed method are scaled empirically, such as the threshold for envelope extraction and regularization coefficient used in slope calculation. Further investigation is needed to determine explicit scaling rules.

Lastly, we have clarified the effects of the RIR on an observed reverberant signal by using a mathematical model of the temporal envelope. Since the temporal envelope plays an important role in human auditory perception, this model might be useful for elucidating the connection between the characteristics of room acoustics and the subjective perception of a reverberant sound [54].

VII. CONCLUSION

We have presented an analytical method for blindly estimating the speech transmission index and five room acoustic parameters, i.e., T_{60}, EDT, C_{80}, D_{50}, and T_s. Instead of relying on training data, a model is used to formulate the relationship between a reverberant signal and an extended model of a room impulse response. The proposed method uses the temporal power envelope of an observed signal to estimate the optimal parameters of the extended RIR model. The RIR approximated from the impulse response model was used to estimate STI and RAPs. The evaluation results demonstrate that the proposed method can blindly and simultaneously estimate the STI and RAPs with accuracy comparable to that of previous work that first achieved simultaneous estimation of six room-acoustic related parameters [4]. Future work includes evaluating the accuracy of the proposed method in spaces where people exist will be evaluated. A robust estimation against background noise will also be investigated further since the noise causes the significant effect on the estimation accuracy [23].

VIII. CREDIT AUTHORSHIP CONTRIBUTION

Lijun Wang: Survey, Conceptualization, Methodology, Algorithm, Validation, Visualization, Writing—original draft, and Writing—review and editing. **Suradej Duangpummet**: Survey, Conceptualization, Validation, and Writing—review and editing. **Masashi Unoki**: Conceptualization, Funding, Supervision, and Writing—review and editing.

APPENDIX A
PROOF OF LEMMA 1

Using (10) and (11), the TPE of a casual RIR can be determined as:

$$e_h^2(t) = \begin{cases} a^2 \exp\left(\frac{13.8t}{T_h}\right), & t \in [0, T_h) \\ a^2 \exp\left(\frac{-13.8t}{T_t}\right), & t \in [T_h, T_t] \end{cases} \tag{36}$$

During the time interval $t \in [0, T_h)$, the corresponding TPE of the reverberant signal can be determined by using (19) and (17):

$$e_y^2(t) = \int_0^{T_h} Ca^2 \exp\left[\frac{13.8(t-\tau)}{T_h}\right] \cos(2\pi f_m \tau + \phi) d\tau$$
$$= Ca^2 \exp\left(13.8\frac{t}{T_h}\right) \int_0^{T_h} \exp\left(\frac{-13.8\tau}{T_h}\right)$$

$$
\begin{aligned}
&\times \cos(2\pi f_m \tau + \phi) d\tau \\
&= Ca^2 \exp\left(13.8 \frac{t}{T_h}\right) \Re \left\{ \int_0^{T_h} \exp\left(\frac{-13.8\tau}{T_h}\right) \right. \\
&\quad \left. \times \exp\left[j(2\pi f_m \tau + \phi)\right] d\tau \right\} \\
&= \frac{Ca^2 T_h \exp\left[13.8\left(\frac{t}{T_h} - 1\right)\right]}{13.8^2 + (2\pi f_m T_h)^2} \left\{ 13.8 \left[\exp(13.8)\cos(\phi) \right.\right. \\
&\quad \left. - cos(2\pi f_m T_h + \phi_k) \right] + 2\pi f_m T_h \left[\sin(2\pi f_m T_h + \phi) \right. \\
&\quad \left.\left. - \exp(13.8)\sin(\phi) \right] \right\}.
\end{aligned} \tag{37}
$$

Similarly, with regard to $t \in [T_h, T - T_t]$, the reverberant TPE $e_y^2(t)$ can be determined:

$$
\begin{aligned}
e_y^2(t) &= \int_{T_h}^{t} Ca^2 \exp\left[\frac{-13.8(t-\tau)}{T_t}\right] \cos(2\pi f_m \tau + \phi) d\tau \\
&= \frac{Ca^2 T_t}{13.8^2 + (2\pi f_m T_h)^2} \left\{ 13.8 \cos(2\pi f_m t) \right. \\
&\quad + 2\pi f_m T_t \sin(2\pi f_m t + \phi) - \exp\left[\frac{-13.8(t - T_h)}{T_t}\right] \\
&\quad \left. \times \left[13.8 \cos(2\pi f_m T_h + \phi) 2\pi f_m T_t \sin(2\pi f_m T_h + \phi) \right] \right\} \\
&= \frac{Ca^2 T_t}{\sqrt{13.8^2 + (2\pi f_m T_h)^2}} \\
&\quad \left[\cos(2\pi f_m t + \phi) - \exp\left[-13.8\left(\frac{t - T_h}{T_t}\right)\right] \right. \\
&\quad \left. \times \cos(2\pi f_m T_h + \phi + \theta) \right]. \ \theta = \tan^{-1}\left(\frac{-2\pi f_m T_t}{13.8}\right)
\end{aligned} \tag{38}
$$

Lastly, in time interval $t \in (T - T_t, T_t]$, the TPE of the reverberant signal is terminated. $e_y^2(t)$ is given by back integration:

$$
\begin{aligned}
e_y^2(t) &= \int_{-T_t}^{0} Ca^2 \exp\left[\frac{-13.8(t-\tau)}{T_t}\right] \cos(2\pi f_m \tau + \phi) d\tau \\
&= Ca^2 \exp\left(-13.8 \frac{t}{T_t}\right) \Re \left\{ \int_{-T_t}^{0} \exp\left(\frac{13.8\tau}{T_t}\right) \right. \\
&\quad \left. \times \exp\left[j(2\pi f_m \tau + \phi)\right] d\tau \right\} \\
&= \frac{Ca^2 T_t \exp\left(-13.8 \frac{t}{T_t}\right)}{13.8^2 + (2\pi f_m T_t)^2} \left\{ 13.8 \left[\exp(13.8)\cos(\phi) \right.\right. \\
&\quad \left. - \cos(2\pi f_m T_t - \phi) \right] + 2\pi f_m T_t \left[\sin(2\pi f_m T_t - \phi) + \right. \\
&\quad \left.\left. \times \exp(13.8)\sin(\phi) \right] \right\}
\end{aligned} \tag{39}
$$

REFERENCES

[1] V. G. Escobar and J. M. B. Morillas, "Analysis of intelligibility and reverberation time recommendations in educational rooms," *Appl. Acoust.*, vol. 96, pp. 1–10, Sep. 2015.

[2] T. Schäfer, P. Sedlmeier, C. Städtler, and D. Huron, "The psychological functions of music listening," *Frontiers Psychol.*, vol. 4, p. 511, Aug. 2013.

[3] M. Barron, *Auditorium Acoustics and Architectural Design*, 2nd ed. London, U.K.: Routledge, 2009.

[4] S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response," *Appl. Acoust.*, vol. 185, Jan. 2022, Art. no. 108372.

[5] H. Kuttruff, *Room Acoustics*. New York, NY, USA: Taylor & Francis, 2016.

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

[7] S. Cerdá, A. Giménez, J. Romero, R. Cibrián, and J. L. Miralles, "Room acoustical parameters: A factor analysis approach," *Appl. Acoust.*, vol. 70, no. 1, pp. 97–109, Jan. 2009.

[8] H.-Y. Dong and C.-M. Lee, "Speech intelligibility improvement in noisy reverberant environments based on speech enhancement and inverse filtering," *EURASIP J. Audio, Speech, Music Process.*, vol. 2018, no. 1, pp. 1–13, Dec. 2018.

[9] B. Eurich, T. Klenzner, and M. Oehler, "Impact of room acoustic parameters on speech and music perception among participants with cochlear implants," *Hearing Res.*, vol. 377, pp. 122–132, Jun. 2019.

[10] F. Xiong, S. Goetze, and B. Meyer, "Estimating room acoustic parameters for speech recognizer adaptation and combination in reverberant environments," Tech. Rep., May 2014.

[11] N. Chu, Y. Ning, L. Yu, Q. Liu, Q. Huang, D. Wu, and P. Hou, "Acoustic source localization in a reverberant environment based on sound field morphological component analysis and alternating direction method of multipliers," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[12] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 54, no. 2, p. 557, Aug. 1973.

[13] *Sound System Equipment—Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index*, document IEC 60268-16, 2020.

[14] *Acoustics Measurements of Room Acoustics Parameters—Part 1: Performance Spaces*, document ISO 3382, 2009.

[15] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[16] M. R. Schroeder, "Modulation transfer functions: Definition and measurement," *Acta Acustica United With Acustica*, vol. 49, no. 3, pp. 179–182, 1981.

[17] M. Unoki and Z. Zhu, "Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech," *Acoust. Sci. Technol.*, vol. 41, no. 1, pp. 233–244, 2020.

[18] W. T. Chu, "Impulse-response and reverberation-decay measurements made by using a periodic pseudorandom sequence," *Appl. Acoust.*, vol. 29, no. 3, pp. 193–205, 1990.

[19] P. Kendrick, F. Li, T. Cox, Y. Zhang, and J. Chambers, "Blind estimation of reverberation parameters for non-diffuse rooms," *Acta Acustica United With Acustica*, vol. 93, pp. 760–770, Sep. 2007.

[20] M. Unoki, A. Miyazaki, S. Morita, and M. Akagi, "Method of blindly estimating speech transmission index in noisy reverberant environments," *J. Inf. Hiding Multimedia Signal Process.*, vol. 8, pp. 1430–1445, Nov. 2017.

[21] A. Keshavarz, S. Mosayyebpour, M. Biguesh, T. A. Gulliver, and M. Esmaeili, "Speech-model based accurate blind reverberation time estimation using an LPC filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 6, pp. 1884–1893, Aug. 2012.

[22] P. Götz, C. Tuna, A. Walther, and E. A. P. Habets, "Blind reverberation time estimation in dynamic acoustic conditions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 581–585.

[23] K. Zheng, C. Zheng, J. Sang, Y. Zhang, and X. Li, "Noise-robust blind reverberation time estimation using noise-aware time–frequency masking," *Measurement*, vol. 192, Mar. 2022, Art. no. 110901.

[24] J. Lopez-Ballester, S. Felici-Castell, J. Segura-Garcia, and M. Cobos, "AI-IoT platform for blind estimation of room acoustic parameters based on deep neural networks," *IEEE Internet Things J.*, vol. 10, no. 1, pp. 855–866, Jan. 2023.

[25] F. Toole, *Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms* (Audio Engineering Society Presents). New York, NY, USA: Taylor & Francis, 2017.

[26] *Benchmark Problems for Acoustical Parameters*, Architectural Institute of Japan, Tokyo, Japan.

[27] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, no. 3, pp. 409–412, Mar. 1965.

[28] B. Rakerd, E. J. Hunter, M. Berardi, and P. Bottalico, "Assessing the acoustic characteristics of rooms: A tutorial with examples," *Perspect. ASHA Special Interest Groups*, vol. 3, no. 19, pp. 8–24, Jan. 2018.

[29] M. Unoki and S. Hiramatsu, "MTF-based method of blind estimation of reverberation time in room acoustics," in *Proc. 16th Eur. Signal Process. Conf.*, Aug. 2008, pp. 1–5.

[30] L. Couvreur, C. Ris, and C. Couvreur, "Model-based blind estimation of reverberation time: Application to robust ASR in reverberant environments," Tech. Rep., 2001.

[31] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Amer.*, vol. 114, no. 5, pp. 2877–2892, 2003.

[32] J. F. Santos and T. H. Falk, "Blind room acoustics characterization using recurrent neural networks and modulation spectrum dynamics," *J. Audio Eng. Soc.*, to be published.

[33] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 591–595.

[34] F. F. Li and T. J. Cox, "A neural network model for speech intelligibility quantification," *Appl. Soft Comput.*, vol. 7, no. 1, pp. 145–155, Jan. 2007.

[35] 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. Amer.*, vol. 124, no. 1, pp. 278–287, Jul. 2008.

[36] P. Callens and M. Cernak, "Joint blind room acoustic characterization from speech and music signals using convolutional recurrent neural networks," 2020, *arXiv:2010.11167*.

[37] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2021, pp. 221–225.

[38] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 10, pp. 1681–1693, Oct. 2016.

[39] P. Peso Parada, D. Sharma, T. Waterschoot, and P. Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ace challenge," Tech. Rep., Oct. 2015.

[40] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, pp. 136–140.

[41] M. Unoki, Y. Kashihara, M. Kobayashi, and M. Akagi, "Study on method for protecting speech privacy by actively controlling speech transmission index in simulated room," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 1199–1204.

[42] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes With Errata Sheet*, New York, NY, USA: McGraw-Hill, 2002.

[43] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "An improved method based on the MTF concept for restoring the power envelope from a reverberant signal," *Acoust. Sci. Technol.*, vol. 25, no. 4, pp. 232–242, 2004.

[44] P. P. Vaidyanathan, *The Theory of Linear Prediction* (Synthesis Lectures on Engineering Series). San Rafael, CA, USA: Morgan & Claypool, 2008.

[45] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation* (Prentice-Hall Information and System Sciences Series). Upper Saddle River, NJ, USA: Prentice-Hall, 2000.

[46] S. Haykin, *Adaptive Filter Theory*. London, U.K.: Pearson, 2014.

[47] A. H. Sayed, *Fundamentals of Adaptive Filtering*. Hoboken, NJ, USA: Wiley, 2003.

[48] G. Heinig and K. Rost, "Fast algorithms for Toeplitz and Hankel matrices," *Linear Algebra Appl.*, vol. 435, no. 1, pp. 1–59, Jul. 2011.

[49] *Sound Library of Architecture and Environment*, Architectural Institute of Japan, Gihodo Shuppan, Tokyo, Japan, 2004.

[50] T. Takeda, Y. Sagisak, K. Katagiri, M. Abe, and H. Kuwabara, "Speech database user's manual," ATR, Blagnac, France, Tech. Rep. TR-I-0028, 1988.

[51] J. S. Bradley, R. Reich, and S. G. Norcross, "A just noticeable difference in C50 for speech," *Appl. Acoust.*, vol. 58, no. 2, pp. 99–108, 1999.

[52] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 48, pp. E7856–E7865, Nov. 2016.

[53] R. Badeau, "Common mathematical framework for stochastic reverberation models," *J. Acoust. Soc. Amer.*, vol. 145, no. 4, pp. 2733–2745, Apr. 2019.

[54] B. C. J. Moore, "The roles of temporal envelope and fine structure information in auditory perception," *Acoust. Sci. Technol.*, vol. 40, no. 2, pp. 61–83, 2019.

LIJUN WANG received the M.S. degree (Hons.) in information science from the Japan Advanced Institute of Science and Technology (JAIST), Japan, in 2023, where he is currently pursuing the Ph.D. degree. His research interests include room acoustics, statistical signal processing, psychoacoustics, and auditory systems. He is a Student Member of the Acoustical Society of Japan (ASJ).

SURADEJ DUANGPUMMET received the B.Eng. degree in mechatronics engineering from the Pathumwan Institute of Technology, in 2002, the M.Eng. degree in mechatronics engineering from the Asian Institute of Technology, Thailand, in 2011, and the Ph.D. degree (Hons.) in information science from the Japan Advanced Institute of Science and Technology (JAIST), in 2021. He has been with the National Electronics and Computer Technology Center (NECTEC), since 2002, where he is currently a Researcher with the Artificial Intelligence Research Group. His research interests include acoustic signal processing, system identification and estimation, pattern recognition, and machine learning.

MASASHI UNOKI (Member, IEEE) received the M.S. and Ph.D. degrees in information science from the Japan Advanced Institute of Science and Technology (JAIST), in 1996 and 1999, respectively. His research interests include auditory-motivated signal processing and the modeling of auditory systems. He was a Japan Society for the Promotion of Science (JSPS) Research Fellow, from 1998 to 2001. He was a Visiting Researcher with the ATR Human Information Processing Laboratories, from 1999 to 2000, and a Visiting Research Associate with the Centre for the Neural Basis of Hearing (CNBH), Department of Physiology, University of Cambridge, from 2000 to 2001. He has been a Faculty with the School of Information Science, JAIST, since 2001, where he is currently a Full Professor and the Dean of the School of Information Science. He is an IEICE Fellow. Currently, he is an Associate Editor of *Applied Acoustics*.