Applied Acoustics 185 (2022) 108372

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response

Suradej Duangpummet [a,b,c,*], Jessada Karnjana [b], Waree Kongprawechnon [c], Masashi Unoki [a]

[a] School of Information Science, Japan Advanced Institute of Science and Technology, Japan
[b] National Science and Technology Development Agency, Thailand
[c] Sirindhorn International Institute of Technology, Thammasat University, Thailand

ARTICLE INFO

Article history:
Received 16 April 2021
Received in revised form 25 July 2021
Accepted 16 August 2021
Available online 3 September 2021

Keywords:
Room acoustics
Room impulse response
Reverberation time
Speech transmission index
Temporal amplitude envelope
Extended RIR model

ABSTRACT

The speech transmission index (STI) and room-acoustic parameters, such as the reverberation time and clarity index, are essential to assess the quality of room acoustics. However, in everyday spaces occupied by people, it is difficult to obtain such parameters since the room impulse response (RIR) cannot be measured. Blind estimation of room acoustic parameters from observed signals without measuring RIR is therefore necessary. Although existing methods can estimate one of these parameters, a single parameter is inadequate to describe comprehensive subjective aspects. To this end, this paper proposes a method for estimating STI and five room-acoustic parameters simultaneously. The temporal amplitude envelope of a reverberant speech signal is mapped to the parameters of an RIR model for each sub-band. Instead of using Schroeder's RIR model, the extended RIR model is used to approximate an unknown RIR so that the STI and room-acoustic parameters can be derived. We performed simulations to evaluate the proposed method in unseen reverberant environments. The root-mean-square errors between the ground-truths and estimated parameters suggest that the accuracy of the proposed scheme outperformed that with Schroeder's RIR model and was close to the standard measurements.

1. Introduction

The quality of sound fields is essential to our lives in many aspects. For example, in general auditoriums, alarm sounds, emergency announcements, and lectures need to be intelligible and easily audible [1,2]. In theaters and concert halls, the sounds of live performances should be clear and transparent. Intelligibility of speech and clarity of music are subjective descriptions that need to conduct listening experiments. Unfortunately, the experiments are not only expensive but also impractical for real-time applications [3–5]. Therefore, such subjective aspects are defined through room-acoustic parameters and objective indices related to the physical properties of a sound field [6]. As a result, room-acoustic parameters and objective indices can be of great benefit to public announcements, hearing-aid devices, and speech enhancement algorithms. In addition, architects who are involved in designing auditoriums or diagnosing acoustic problems, can justify the acoustic quality from these room-acoustic parameters.

Several useful room-acoustic parameters and objective indices have been standardized [7,8]. In IEC 60268 − 16 : 2020, the speech transmission index (STI), which is an objective index, is used to predict speech intelligibility from the quality of a transmission channel [7]. The STI is calculated on the basis of the modulation transfer function (MTF) concept [9,10]. In addition, ISO 3382 − 1 : 2009 specifies methods for measuring reverberation times (T_{60} or T_{30}) and other room-acoustic parameters, such as early decay time (EDT), clarity (early-to-late-arriving sound energy ratios: C_{80} or C_{50}), Deutlichkeit (early-to-total sound energy ratio: D_{50}), and center time (T_s) [8]. These parameters are derived from measuring the room impulse response (RIR). In the time domain, the RIR can completely represent the characteristics of a sound field. Similarly, the MTF describes the effects of reverberation in the modulation-frequency domain.

In general, the RIR or MTF needs to be measured. However, it is difficult to measure them in places where people cannot be excluded, e.g., stations and department stores. To measure the RIR, the given room is excited with a high-energy sound, e.g., piston shorts, bursting balloons, or power amplifiers with loudspeakers. Such measured RIRs are valid for particular positions between sources and receivers, but they might not cover an entire area. Furthermore, in public areas, room acoustics are a time-varying

* Corresponding author at: Unoki Laboratory, School of Information Science, Japan Advanced Institute of Science and Technology, Japan.
E-mail address: suradej@jaist.ac.jp (S. Duangpummet).

https://doi.org/10.1016/j.apacoust.2021.108372
0003-682X/

system. Sound absorption and reverberation are changed by occupants and the arrangement of objects [11–13]. Acoustic parameters that were measured in compliance with particular standards might differ under the current circumstances. Hence, a few methods have been proposed for estimating an acoustical parameter without measuring the RIR, known as blind estimation methods. The blind estimation of an acoustical parameter is an ill-posed problem because the original sound source and RIR are unknown. This ill-posed problem is challenging since assumptions or complementary prior knowledge are needed to formulate an estimation.

Unoki et al. proposed methods based on the MTF concept to estimate T_{60} and STI [14–16]. In their methods, the RIR is first approximated by using Schroeder's RIR model and later by a more flexible model, namely the generalized RIR model. The concept of the MTF is used to restore the modulation spectrum from an observed signal. Then, the optimal parameters of the RIR models are calculated. The estimated STI is derived from the MTF of the generalized RIR model. The generalized RIR model was proposed by modifying Schroeder's RIR model [16]. It is more accurate and closer to the measured RIRs than Schroeder's RIR model. The model has two parameters. The first represents reverberation time, but the second has no physical meaning [16]. On the basis of statistics, Kendrick et al. proposed a maximum likelihood estimation for approximating energy decay curves from reverberant speech and music [17,18]. The energy decay curve is used for calculating the reverberation time and early decay time. Although this approach produces reasonable results, it needs a long time period in the data record, i.e., 30–60 min [12]. Because of this, it might not be suitable for time-varying environments such as in common spaces.

Techniques based on machine learning have been successfully used to estimate T_{60}, STI, and the clarity of speech [19–27]. For example, in an early method based on a neural network, principal component analysis was used to reduce the input dimensions from the envelope spectrum of a speech signal. Then, multi-layer perceptron was used to estimate the STI from that feature [19]. Since the emergence of deep learning and parallel computing, many deep neural networks have been proposed for estimating acoustic parameters. A convolutional neural network (CNN), which is trained from a massive number of reverberant speech signals, can estimate STIs efficiently without feature extraction in the so-called end-to-end model [20]. We also previously proposed a robust method that estimates STIs by using a CNN and full-band temporal amplitude envelope (TAE) of a noisy reverberant speech signal [27]. The CNN was trained from pairs of TAEs and STIs in various acoustic conditions. The acoustic conditions were synthesized by using the image-source method [28]. This method generates simulated RIRs from the positions of sound sources, receivers, and absorption coefficients. The method could overcome a mismatch problem between models and actual conditions.

For T_{60} estimation, many approaches have been evaluated in the Acoustic Characterization of Environments (ACE) challenge [21]. For example, Gamper and Tashev proposed a CNN with spectra-temporal features in the time–frequency domain [22]. Recently, a combination of a CNN and long short-term memory (LSTM) network was proposed [23].

These works showed that a method using either a CNN or CNN with LSTM is computationally efficient, and only a few seconds of an observed speech signal is sufficient for reasonable accuracy. Parada et al. proposed bidirectional LSTM for estimating the clarity index at 50 ms (C_{50}) using a spectral envelope in the modulation-domain as input [24,25]. A similar architecture, i.e., CNN-LSTM with input features in the time–frequency domain using a short-time Fourier transform, known as a spectrogram, also achieved good performance [26].

Although the aforementioned methods can estimate one of these acoustic parameters, a single parameter is inadequate to describe the characteristics of room acoustics completely. None of these methods can estimate multiple parameters and indices simultaneously. In contrast, blind estimation of multiple room-acoustic parameters and indices is more fruitful and close to the standard measurements deriving the STI and acoustical parameters from the RIR.

This paper proposes a blind method for estimating five room-acoustic parameters: T_{60}, EDT, C_{80}, D_{50}, T_s, and one objective index, STI, simultaneously. We propose a novel stochastic RIR model, namely the extended RIR model, in which all parameters of the model are associated with realistic RIRs. The proposed method incorporates the extended RIR model into the MTF-based CNNs framework to approximate unknown RIRs. The reconstructed RIR based on the extended RIR model is used to derive the STI and five room-acoustic parameters. The estimated results are then compared with the framework using Schroeder's RIR model and the standard methods on the basis of the measured RIRs.

The rest of the paper is organized as follows. Section 2 introduces the STI and five room-acoustic parameters. The proposed method is described in Section 3. The evaluations and results are shown in Section 4. The results are discussed in Section 5, and Section 6 is the conclusion.

2. Speech transmission index and room-acoustic parameters

A number of objective indices and room-acoustic parameters have been studied to justify a sound field. In this study, one objective index and five single-channel-acoustic parameters are blindly estimated. These well-defined parameters are important and often used by architects and musicians [6,29]. In this section, the objective index and acoustic parameters of interest are briefly described as follows.

- The speech transmission index (STI), which is an objective index, is used to predict speech intelligibility and listening difficulty [7]. The quality of a transmission channel from a talker to a listener can be indicated by a signal number. Houstgast and Steeneken proposed an STI based on the MTF [9,10].
  The MTF is a transfer function of a linear system. It represents the characteristics of a transmission channel as a function of the modulation frequency and a decrease in modulation depth [6]. In room acoustics, the MTF concept is used to quantify the effects of reverberation. The higher the reverberation, the lower the modulation depth of modulated signals that pass through the room. The modulation distortion ratios between the input envelopes and the corresponding outputs are known as modulation indices. The magnitude of the MTF is defined as

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

  where $m(f_m)$ is the MTF at modulation frequency f_m, and $h(t)$ is a room impulse response.
  The STI method has been standardized by IEC $60268-16:2020$ [7]. Fig. 1 shows a diagram of the set-up for measuring and calculating the STI. A total of 98 modulated stimuli are used to calculate the distortion ratios between the inputs and observed signals. The stimuli are amplitude-modulated signals from seven-octave bands of carriers and 14 modulation frequencies, f_m. The modulation distortion ratio, N, is calculated as

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

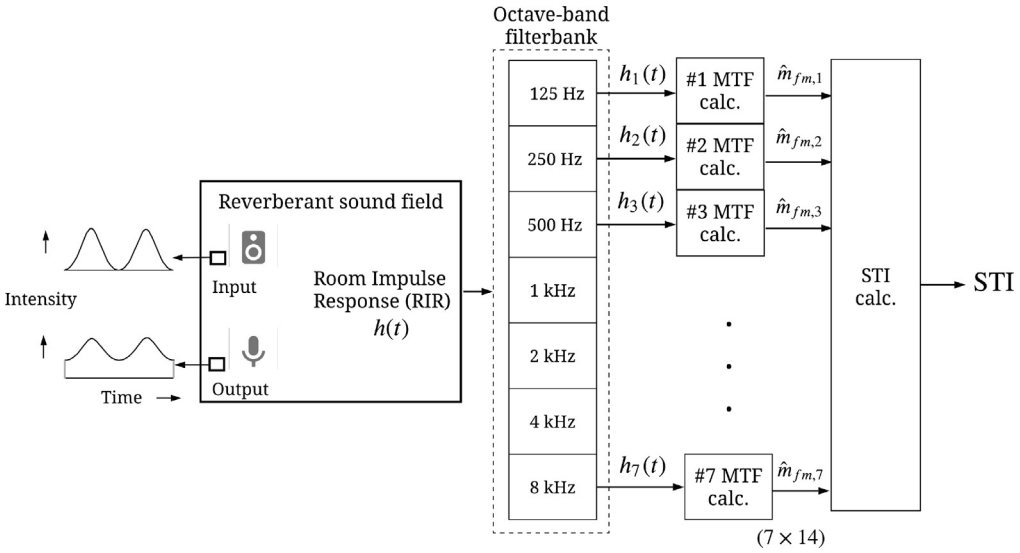


Fig. 1. Block diagram of set-up for measuring and calculating STI.

where $i = 1$ to 14, and $k = 1$ to 7. These values are limited to the range of -15 dB to $+15$ dB and are normalized. Then, the modulation transmission indices (MTIs), $M(k)$, are calculated by averaging $N(k, i)$ as

$$M_k = \frac{1}{14}\sum_{i=1}^{14} N_{k,i}. \tag{3}$$

Finally, the STI is calculated from the MTIs for each octave band as

$$\mathrm{STI} = \sum_{k=1}^{7}\alpha_k M_k - \sum_{k=1}^{6}\beta_k\sqrt{M_k M_{k+1}}, \tag{4}$$

where α_k and β_k are the weighting factors for octave band k, as shown in Table 1.
The STI is a real number on a scale between 0 and 1. Instead of a direct method based on measuring the distortion ratios of the 98 stimuli, the STI can be calculated from the RIR according to Eq. (1), known as the indirect method.

- The reverberation time (T_{60}) and early decay time (EDT) are used to characterize the duration of sound decay in seconds. They are derived from an energy decay curve of the RIR in octave bands. The curve is fitted by using a linear regression. T_{60} is the time period during which the fitted line intersects with -60 dB. In practice, a decay curve between -5 dB and -35 dB below the maximum initial level is recommended to avoid the interference of noise [8]. Similarly, EDT is the time period from the initial 10 dB of the energy decay curve. T_{60} represents reverberation in terms of a physical property, whereas EDT is strongly related to perceived reverberation [30]. Fig. 2 shows reverberation parameters derived from an energy decay curve of the RIR.
- C_{80} and D_{50} are related to the energy ratios between the early and late reflection of the RIR. C_{80} is used to characterize the transparency of music halls in dB units, while C_{50} indicates the transparency of speech. C_{80} is defined as

$$C_{80} = 10\log_{10}\frac{\int_0^{80\mathrm{ms}} h^2(t)dt}{\int_{80\mathrm{ms}}^{\infty} h^2(t)dt}. \tag{5}$$

- Deutlichkeit (D_{50}) is used to evaluate the speech intelligibility of lecture halls or classrooms (in percentage). D_{50} is defined as

$$D_{50} = \frac{\int_0^{50\mathrm{ms}} h^2(t)dt}{\int_0^{\infty} h^2(t)dt}\times 100. \tag{6}$$

- The center time, T_s, is the period at the center of gravity of the RIR. T_s shows the balance between clarity and reverberation related to speech intelligibility. T_s is defined as

$$T_s = \frac{\int_0^{\infty} h^2(t)t\,dt}{\int_0^{\infty} h^2(t)dt}. \tag{7}$$

3. Proposed method

We propose a method for blindly estimating the STI and five room-acoustic parameters: T_{60}, EDT, C_{80}, D_{50}, and T_s. The proposed method is mainly based on approximating an unknown RIR from a reverberant speech signal. Hence, the critical issues are the accuracy of the RIR model and its estimation. Schroeder's RIR model and its extended version, namely the extended RIR model, are investigated. The parameters of the RIR model are estimated on the basis of the MTF concept in octave bands. In the first subsection, the main concept previously used for Schroeder's RIR is introduced. Then, the proposed method, which is the extended RIR model incorporated into the MTF-based CNN framework, is described.

Table 1
MTI octave-band weighting factors [7].

band (Hz)	125	250	500	1k	2k	4k	8k
α	0.085	0.127	0.230	0.233	0.309	0.224	0.173
β	0.085	0.078	0.065	0.011	0.047	0.095	–

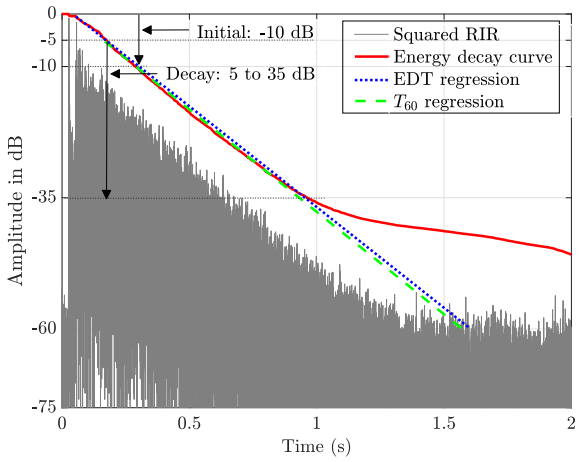


Fig. 2. Deriving reverberation time and early decay time from energy decay curve of RIR.

3.1. Previous method using Schroeder's RIR model

In a reverberant environment, we assume that an observed signal, $y(t)$, is a result of the convolution between an original speech signal, $x(t)$, and the RIR as

$$y(t) = x(t) * h(t), \tag{8}$$

where asterisk (*) indicates a convolution operation. For blind estimation of the STI and room acoustic parameters, we have only an observed signal, $y(t)$. Hence, the impulse response or system transfer function (MTF) needs to be approximated with additional assumptions. Previously, the actual RIR was assumed to be a stochastic RIR. Schroeder's RIR model is used to approximate unknown RIRs [32,31]. Schroeder's RIR model is defined as

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

where $e_h(t)$ is an exponential decay, $c_h(t)$ is a carrier of white Gaussian noise (WGN) acting as a random variable, and a is a gain factor. Thus, the MTF according to the Schroeder's RIR can be expressed as

$$m(f_m, T_{60}) = \left[1 + \left(2\pi f_m \frac{T_{60}}{13.8}\right)^2\right]^{-\frac{1}{2}}. \tag{10}$$

Fig. 3 shows the main concept of our previous method [31]. T_{60}s represent the estimated reverberation times of seven-octave bands. In accordance with the MTF, the TAE of an observed speech signal is distorted as the reverberation time increases. Thus, the parameter of Schroeder's RIR model, i.e., T_{60}, can be fitted to the observed TAE.

Schroeder's RIR model is an ideal exponential decay function. It is valid for representing a geometrically simple enclosure, e.g., an empty rectangular room without furniture and partitions. However, real spaces are often more complicated. For instance, a department store contains shelves of products, as depicted in Fig. 5. At some positions, a listener receives reflections from many surfaces with a delay time. Hence, a simple exponential decay model such as Schroeder's RIR model cannot represent such a complicated environment. The mismatch between the actual RIR and the model leads to inaccurately estimated acoustic parameters. Therefore, the modeling of an actual RIR with a non-exponential decay needs to be improved [14,15].

Instead of using Schroeder's RIR model, the extended RIR model is used. The proposed method, the MTF-based CNNs with extended RIR model, is shown in Fig. 4. The details of the proposed method are described as follows.

3.2. Extended RIR model

The extended RIR model is proposed to mitigate a limitation of Schroeder's RIR model. Thus, Schroeder's RIR model was modified by adding two more parameters. The extended RIR model, $h_{ext}(t)$, is defined as

$$h(t) = h_{ext}(t - T_0), T_0 \geqslant 0 \tag{11}$$

$$h_{ext}(t) = \begin{cases} a \exp(6.9t/T_h) c_h(t), & t < 0 \\ a \exp(-6.9t/T_t) c_h(t), & t \geqslant 0 \end{cases} \tag{12}$$

where T_0 denotes the peak position of the RIR. T_h and T_t are the controlling parameters for raising and decreasing the envelope of the RIR, respectively. a is a gain factor, and c_h is the WGN carrier, which is a random variable.

In Eq. (11), the time-shifting property is used to provide a causal system and stable impulse response, i.e., $h(t) = 0,\ t < 0$. In Eq. (12), the three parameters of the extended RIR model control the shape of the envelope of the RIR.

Fig. 6 shows an example of the extended RIR. The time period from the sound source ($t = 0$) to the peak position of the RIR is controlled by parameters T_h and T_0. The last parameter, T_t, represents the exponential decay of the RIR. In other words, T_t is the reverberation time, T_{60}. The envelope of the RIR is varied according to the three control parameters, as shown in Fig. 6 (a), and Fig. 6 (b) shows the RIR after the envelope is modulated by WGN. Note that if T_h and T_0 are equal to zero, the extended RIR model is the same as Schroeder's RIR model.

The extended RIR model is therefore more flexible and closer to the temporal envelope of the real RIR. Fig. 7 shows a comparison between the two RIR models to represent an actual RIR. Nevertheless, a method for estimating the parameters of the extended RIR model has not been developed. Thus, one of the main contributions to the complementary prior knowledge of the proposed method is

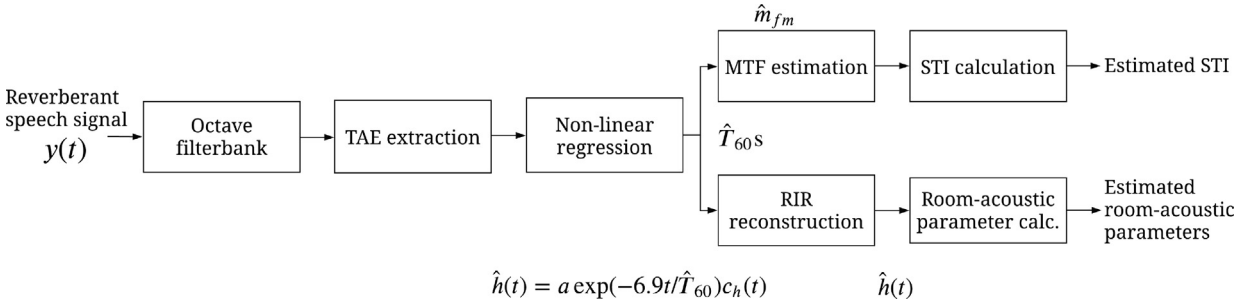


Fig. 3. Block diagram for concept of previous method [31].

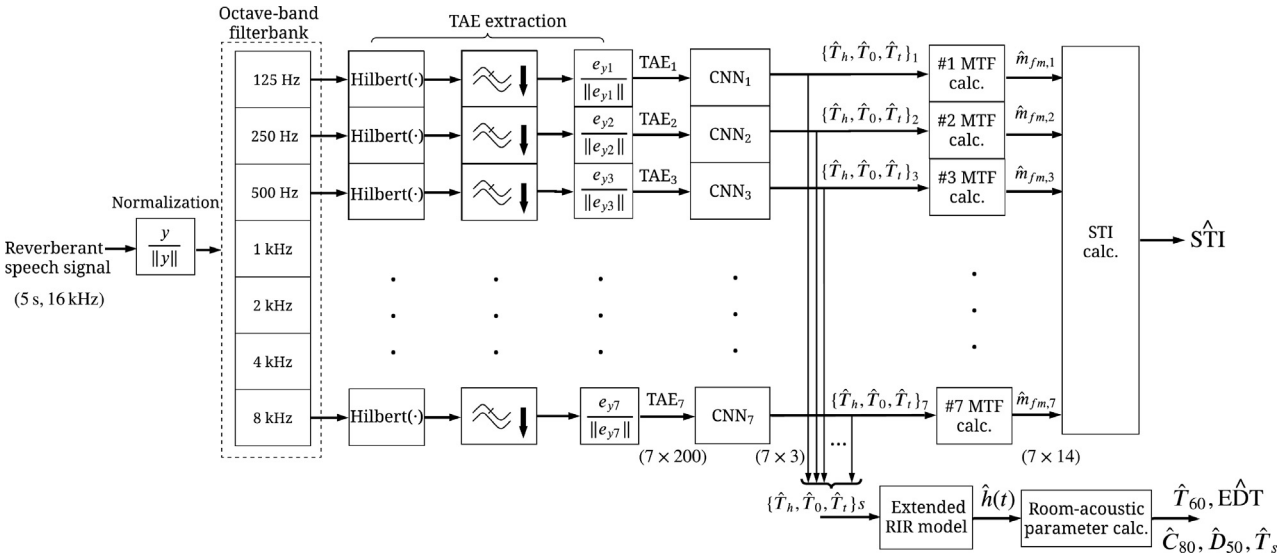


Fig. 4. Block diagram of proposed method.

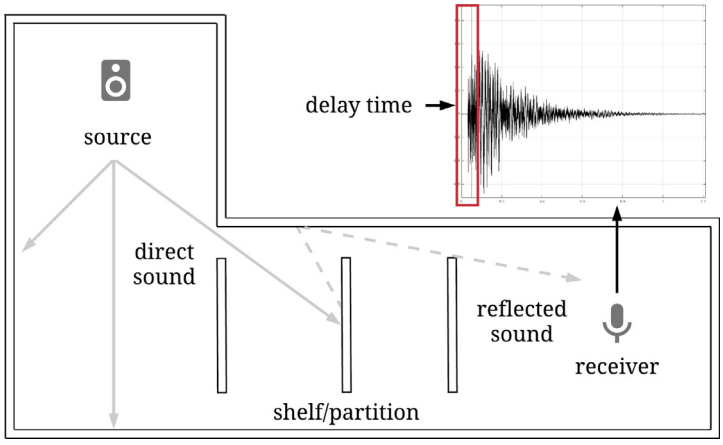


Fig. 5. Example of complex space and its impulse response.

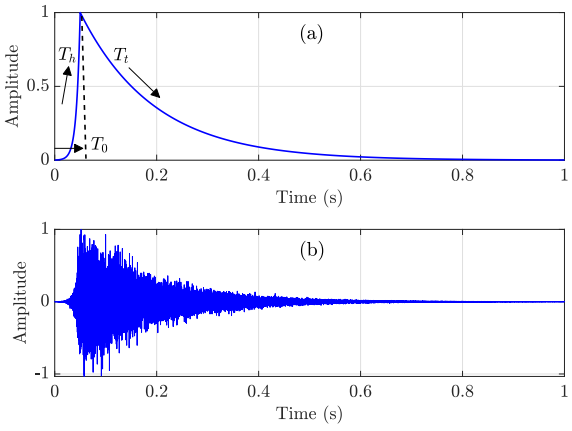


Fig. 6. Example of extended RIR model where $T_h = 0.08, T_0 = 0.05$ s, and $T_t = 1.0$: (a) temporal envelope and (b) its corresponding RIR.

that the three parameters of the extended RIR model, i.e., T_h, T_t, and T_0, are blindly estimated. Thus, according to the definition of the MTF in Eq. (1), the complex MTF of the extended RIR model can be represented as

$$m(f_m, T_h, T_0, T_t) = \frac{\exp(-j2\pi f_m T_0)}{\sqrt{\left(1 + (2\pi f_m (T_h/13.8))^2\right)\left(1 + (2\pi f_m (T_t/13.8))^2\right)}}. \tag{13}$$

3.3. TAE in seven-octave bands

Speech signals can be regarded as a summation of a temporal amplitude envelope (TAE) with a temporal fine structure. The TAE of speech plays an important role in speech intelligibility [33]. In reverberant environments, the TAE of an observed signal is a smoothed version of the original signal. The more reverberation time there is of an enclosure, the smoother the received signal, as shown in Fig. 8. The TAE of an observed signal therefore provides essential information regarding the room characteristics. In addition, an idea similar to that of envelope deconvolution techniques is used to reconstruct speech signals and ultrasound images from measured envelopes [34,35].

Thus, we can deal with this ill-posed inverse problem on the basis of two underlying mechanisms. The first is the relationship between the observed TAE and room characteristics. The second is the minimum phase transfer function of the extended RIR model (i.e., a causal and stable system). The TAE of an observed signal is

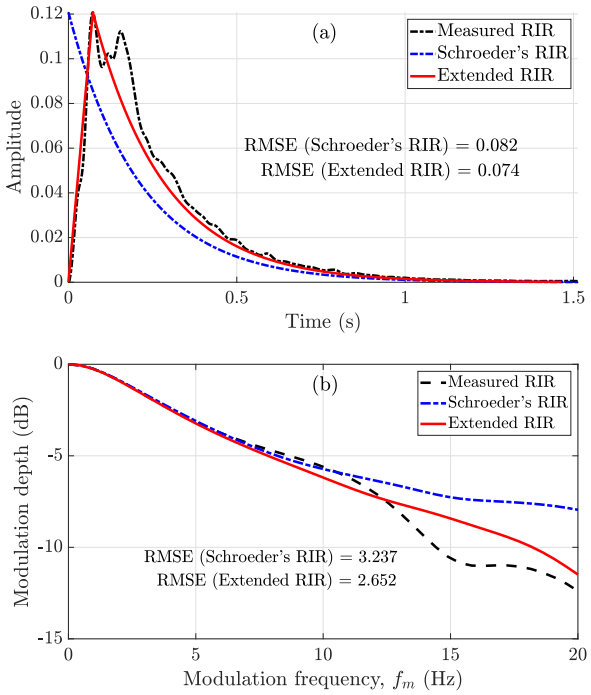


Fig. 7. Fitting results of two RIR models with temporal amplitude envelope of measured RIR: envelopes in time domain (a) and in modulation-frequency domain (b).

therefore mapped to the envelope of the RIR by using convolutional neural networks.

The idea of performing a sub-band analysis for estimating room acoustic parameters is motivated by the STI-calculation algorithm. As stated in the definition above, the STI is calculated from modulation indices in seven-octave bands. Hence, our filter bank is the same as the STI algorithm, that is, seven-octave filters. The bands have center frequencies ranging from $125-8000$ Hz. The octave filter specification conforms to the standard requirements in ANSI $S1.11-2004$ [36]. Furthermore, acoustic parameters in sub-octave bands are necessary for architects and acousticians [29].

Thus, we utilize the seven TAEs to represent the modulation distortion characteristics caused by reverberation in the bands. The reverberation attenuates the modulation depth of the observed TAEs. Moreover, the seven-band TAEs account for the enhanced accuracy of estimating the three parameters of the extended RIR model. The input is a normalized observed signal (i.e., a reverberant speech signal). The signal is then decomposed to each sub-band by using octave-band filters.

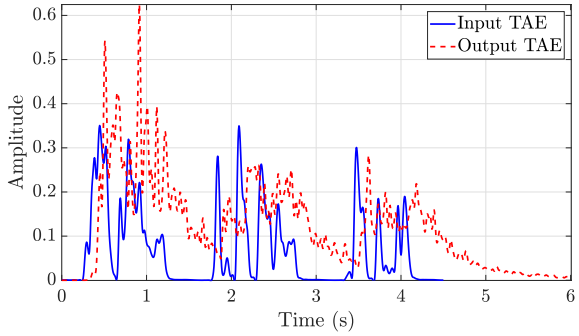


Fig. 8. Temporal amplitude envelope (TAE) of speech signal. Solid line is TAE in free field (no reverberation), and dashed line is TAE in reverberant environment ($T_{60} = 2.0$ s).

The TAE in each band is extracted according to Eq. (14). We use a Hilbert transform and a lowpass filter (LPF). The LPF is a sixth-order Butterworth filter with a cut-off frequency of 20 Hz. We downsample the signal to 40 Hz to reduce the computation complexity. Then, the TAEs are mapped to their associated parameters of the extended RIR for the seven-octave bands by using CNNs.

$$e_y(t) = \mathrm{LPF}[|y(t) + j\mathrm{Hilbert}(y(t))|]. \tag{14}$$

3.4. MTF-based CNNs

We use seven one-dimensional CNNs to map the characteristics of the TAEs to the RIR model parameters. Previously, on the basis of Schroeder's RIR model, the sub-band TAEs were mapped to their associated reverberation times, T_{60}s, whereas in the proposed method, the TAEs are mapped to the three controlling parameters of the extended RIR model: T_h, T_0, and T_t. The seven CNNs are trained from pairs of TAEs and the three parameters of the extended RIR model. The ground-truths of T_h, T_0, and T_t are the targets of the CNNs. The outputs of each CNN for each sub-band are the estimated parameters of the extended RIR model.

The CNN, which is a feed-forward neural network, performs similarly to the convolution operations between trainable filters and the input of each layer. As we decompose the TAE into seven-octave bands, there are seven identical models for each band. Each model consists of four convolutional layers with 6381 parameters. The input layer takes TAEs for convolution with the filters. The regulated linear unit (ReLU), $f(x) = \max(x, 0)$, performs nonlinear activation in every convolutional layer. Batch normalization is applied after the first convolution. Max pooling is also used to reduce the dimensions before the next layer. The dropout rate before the last layer is set to 40% to avoid the memorization problem for some dominant nodes. The fully connected layer with a linear function is the output layer. The details of the MTF-based CNN model are shown in Table 2.

3.5. RIR approximation

Previously, estimated T_{60}s were used to approximate RIR, $\hat{h}(t)$. The approximated RIR is reconstructed by using Schroeder's RIR model. Schroeder's RIR depends on only the reverberation time of a room. Hence, the estimated T_{60} for each octave-band can be used to construct the temporal envelope of the RIR, $\hat{e_h}(t)$. The envelope is modulated with a carrier signal. The carrier signal is band-limited Gaussian noise with the bandwidth of a one-third octave. The sub-band RIRs are then summed for the approximated RIR. The RIR reconstruction is defined as

$$\hat{h}(t) = \sum_{k=1}^{K} \exp\left(-\frac{6.9t}{T_{60,k}}\right) c_{h,k}(t), \tag{15}$$

Table 2
Network architecture of MTF-based CNN model.

No.	Layer Type	Parameters
1	Input	TAE shape $= 1 \times 200$
2	Conv1D^{1st}	32 filters, filter size $= 10 \times 1$, ReLU
3	Pooling	maxpooling, size $= 2$, stride $= 1$
4	Conv1D^{2nd}	16 filters, filter size $= 5 \times 1$, ReLU
5	Pooling	maxpooling, size $= 2$, stride $= 1$
6	Dropout	0.4
7	Conv1D^{3rd}	8 filters, filter size $= 5 \times 1$, ReLU
8	Pooling	maxpooling, size $= 2$,
9	Conv1D^{4th}	4 filters, filter size $= 5 \times 1$, ReLU
10	Fully Connected	3 outputs (T_h, T_0, T_t), relu
11	Regression Output	root-mean-square error (RMSE)

where $T_{60,k}$ is the estimated T_{60} in the k-th band, $K = 7$, and $c_{h,k}(t)$ is band-limited Gaussian noise. The STI can then be calculated from the estimated T_{60} s on the basis of the MTF, as in Eq. (10). Also, the T_{60}, EDT, C_{80}, D_{50}, and T_s can be calculated according to the definitions in Eqs. (5)–(7), respectively.

Similarly, the proposed method replaces Eq. (15) of the original model with Eqs. (11) and (12) of the extended RIR model. Then, the remaining calculations are the same.

3.6. Objective function

The objective function or cost function, $J(\theta)$, is used in the optimization algorithm during training. The filters of the CNNs are convoluted with the input for each layer. The back-propagation algorithm is used to compute the error that is the difference between the estimated parameters and the targets. This kind of parameter estimation problem aims to minimize the error between the estimation and the ground-truths. The optimization algorithm of the proposed method minimizes the error of the three parameters of the extended RIR model. In addition, the algorithm takes the target acoustic parameters into account to enhance the accuracy of estimating STI and room-acoustic parameters. Therefore, the objective function is the root-mean-square error of the estimated parameters of the RIR model and the target acoustic parameters. It is defined as

$$J(\theta) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\alpha\left(T_{h_n} - \widehat{T}_{h_n}\right)^2} + \sqrt{\frac{1}{N}\sum_{n=1}^{N}\beta\left(T_{0_n} - \widehat{T}_{0_n}\right)^2} + \sqrt{\frac{1}{N}\sum_{n=1}^{N}\gamma\left(T_{t_n} - \widehat{T}_{t_n}\right)^2}, \tag{16}$$

where n is the index of the estimated parameters, N is the batch size for each iteration, α, β, and γ are weighting factors of the three controlling parameters, T_h, T_0, and T_t, respectively. Since the scale of T_h and T_0 is comparatively smaller than T_h, the weighting factors are necessary. Here, the weighting factors of T_h, T_0, and T_t are $0.1, 0.3$, and 0.6, respectively.

3.7. Generating dataset

Here, a data augmentation technique is used to generate a sufficient training set. There are several data augmentation methods used in room acoustics. For example, a geometrical acoustic technique, i.e., an image-source method, was used previously [27,28]. Here, we synthesized various RIR on the basis of Schroeder's RIR model and the extended RIR model. The simulated RIRs were synthesized by varying the parameters of the RIR models.

According to Schroeder's RIR model, the reverberation time, T_{60}, in Eq. (9), was varied from 0.3 to 4.0 s with a step size of 0.01 s. The synthesized envelope was modulated with a different random-seed WGN carrier. There are a hundred different WGN carrier seeds. The RIRs were then convoluted with speech signals. The speech signals were ten Japanese sentences uttered by five men and five women from the ATR dataset [37]. Therefore, a total of 29,000 reverberant speech signals were prepared.

Similarly, the three parameters of the extended RIR model were varied to cover the possible range of realistic RIRs. The possible range of each parameter was derived from fitting the envelope of the 43 RIRs. The rising parameter, T_h, was fitted by using nonlinear regression to fit the rising part in Eq. 12. Peak position, T_0, was peak of the envelope of the RIRs. The last parameter, T_t, was the same as T_{60}. These calculated parameters were the ground-truths for evaluating the proposed method.

From calculating the ground-truth parameters of the extended RIR model, it was found that 29 RIRs or 75% of the realistic RIRs in the SMILEdataset might fit well with a simple exponential decay. This means that such RIRs can be represented by Schroeder's model. Nevertheless, a mismatch was found for 14 RIRs, as shown in Fig. 7. Therefore, the dataset from Schroeder's RIR was added with the dataset from the extended RIR model for training the proposed method. For such signals, T_h and T_0 were set to zero. Therefore, a total of 50,000 signals could be used for the proposed method on the basis of the extended RIR. All signals had a five-second period, a sampling rate of 16 kHz, 32-bit quantization, and one channel.

3.8. Training CNNs

We trained NNs by using 80% of the total data. The rest of the data was used to validate the model and to fine-tune the hyperparameters, such as filter size as well as the number of filters and layers. Although finding the optimal parameters of the RIR model is an optimization problem, training the model is slightly different from ordinary optimization. In the training process, solutions are found for a subset from the entire dataset, known as a mini-batch. Here, we set the batch size to 64 records. We trained the model for a maximum of a hundred iterations (or epochs). An early stop was set so that the training stopped when the solution reached the global minimum. We used the RMSprop optimizer, which is an optimization algorithm based on the stochastic gradient descent algorithm [38]. The RMSprop algorithm is recommended for solving such a regression problem. We set the learning rate to start at 0.001. In training, the learning rate was gradually decreased in relation to the rate of convergence, which is called a momentum method [38]. The initial parameters for each convolutional layer were set by using the normalized values of the training set. We implemented and trained the CNN models with Python. Keras with TensorFlow 2.0 was the main library.

4. Evaluation

Since measuring the RIR requires sophisticated equipment, it is expensive. Datasets of real RIRs are limited. This study used 43 realistic RIRs from the SMILE database and 2 RIRs for benchmark algorithms for acoustical parameters [29,39]. The realistic RIRs were used in the final evaluation only. The measured RIRs are in a single channel. They were resampled equally to 16 kHz. Summarized information on the RIRs is in [16,29]. However, the MTF-based CNN framework for estimating acoustical parameters and STI needs more data. Hence, this study utilized the extended RIR model to generate RIRs. A training dataset was synthesized so that the CNNs could estimate the parameters with high accuracy without overfitting and be retrained.

In our experiments, we implemented the framework based on Schroeder's RIR and the extended RIR model. The reverberant TAEs extracted from the observed speech in the seven-octave bands were the inputs for the two models. The CNNs of the proposed method were set as close to the CNNs that were used in the previous method. The main difference is the number of estimated parameters, i.e., one parameter for Schroeder's model (T_{60}) and three for the extended RIR model (T_h, T_0, and T_t). The root-mean-square error (RMSE) and Pearson correlation coefficient were the evaluation metrics. The errors were calculated from the difference between the ground-truths and estimated parameters.

4.1. Evaluating estimated parameters of RIR models

We carried out simulations using reverberant speech signals to determine whether the proposed method can correctly estimate

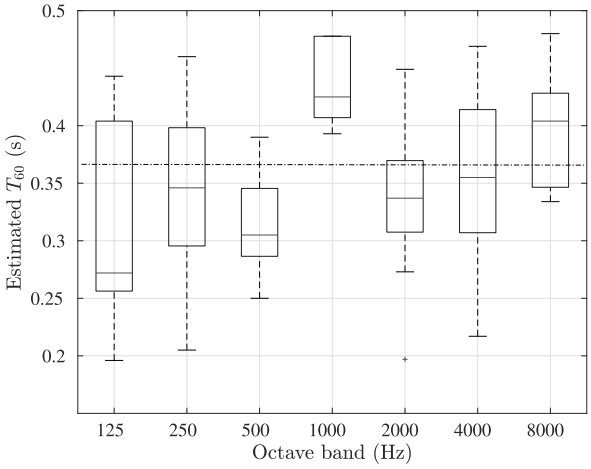


Fig. 9. Example of estimated parameter T_{60} based on Schroeder's RIR model in octave bands. Horizontal dashed line is ground-truth calculated in full-band ($T_{60} = 0.36$ s). Solid line (red) in each box is median of samples. Size of box represents distribution of estimated values, where ten reverberant speech signals were inputs. Symbol "+" is outlier.

the parameter(s) of the RIR models. For the model based on Schroeder's RIR, the reverberation time was the only estimated parameter for each band. Ten speech signals were the inputs for each real RIR. Fig. 9 shows an example of the estimated results. The results of the seven bands had different values according to the frequency-dependence of the reverberation time [29]. However, the middle bands, i.e., from 500 Hz to 2 kHz, were more consistent than the lower and upper bands since the estimated values were distributed in a smaller range.

The three parameters of the extended RIR model were simultaneously estimated for each sub-band. The results are shown in Fig. 10. Then, the estimated parameters of the RIR model were used to reconstruct the approximated RIR. Fig. 11 shows a comparison between the reconstructed envelope of the RIR and the ground-truth. The RMSE was 0.083. It was close to the reference using its fitting parameters, i.e., an RMSE of 0.074.

4.2. Evaluating estimated MTFs

Fig. 12 shows an example of the MTFs approximated from a speech signal in a simulated room ("o") and real room ("*"), where $T_{60} = 0.7$ s. The dashed lines indicate the estimated MTFs, and the solid line is the ground-truth. The estimated MTFs were derived from the MTF of the extended RIR, as in Eq. (13). We averaged the 14 MTFs of the seven-octave bands for clarity. It was found that the shapes of the approximated MTFs were similar to the ground-truths within an RMSE of 0.15 dB.

4.3. Evaluating estimated room-acoustic parameters and STI

The previous method could estimate five-room-acoustic parameters and the STI without having to measure the RIR in reverberant environments. However, the accuracy of the estimated parameters was unreliable as the RIR model did not match many realistic rooms. This critical issue was then evaluated by using the proposed method based on the extended RIR model.

The results of the estimated reverberation time and early decay time are plotted in Fig. 13. Note that all of the estimated parameters and STI are plotted in the same manner as follows. The horizontal axis indicates a parameter directly calculated from the measured RIRs, and the vertical axis indicates the estimated values. The symbol "o" corresponds to the estimated parameters from the

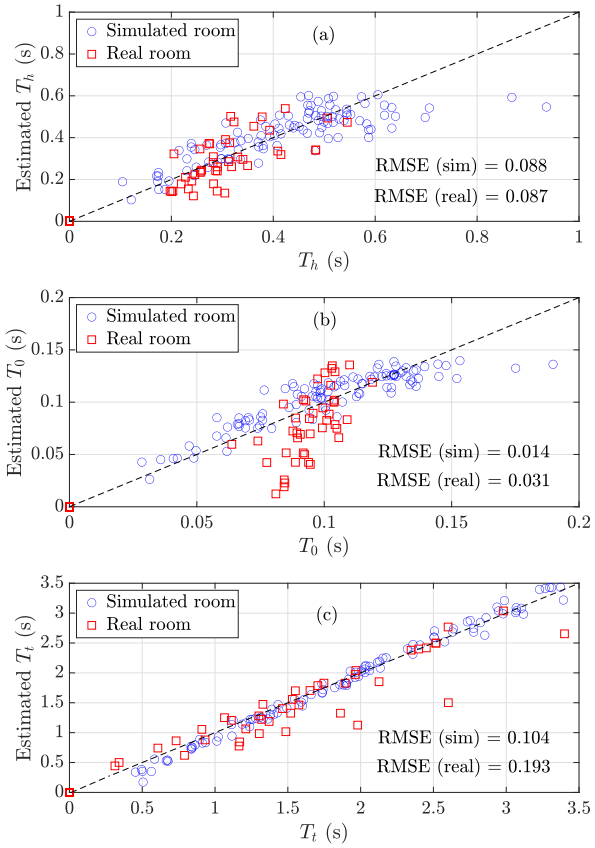


Fig. 10. Results of estimated parameters of extended RIRs: (a) raising parameter T_h, (b) peak position parameter T_0, and (c) decay parameter T_t.

previous method, and the "square" corresponds to the results from the proposed method. The dashed line represents the optimal values for each parameter. For the proposed method, the RMSEs of the estimated T_{60} and EDT were 0.393 and 0.472, respectively. In comparison, the RMSEs with the previous method were 0.440 and 0.478. These two parameters were closely related as they are derived from the same decay curve of the RIR. Therefore, the results showed the same trend. The estimated decay parameter of the two RIR models, i.e., $\widehat{T}_{60}$ and $\widehat{T}_t$, were also the same. Thus, the results of the proposed method were close to those of the previous method.

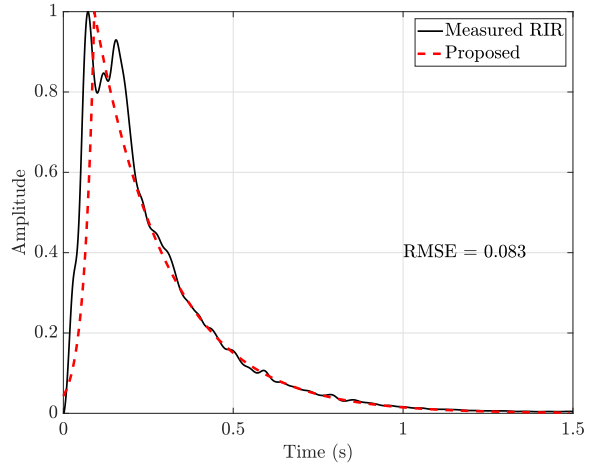


Fig. 11. Actual and reconstructed RIR using proposed method.

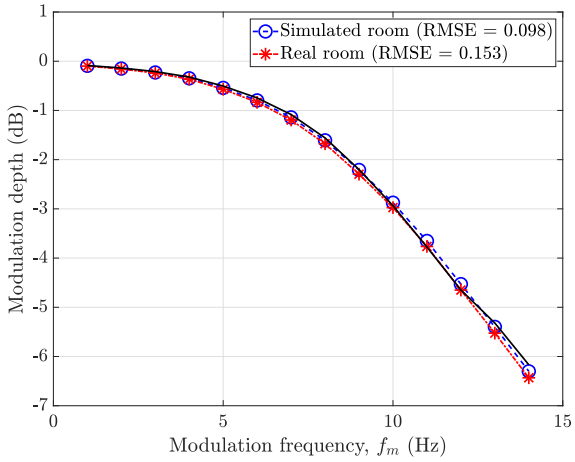
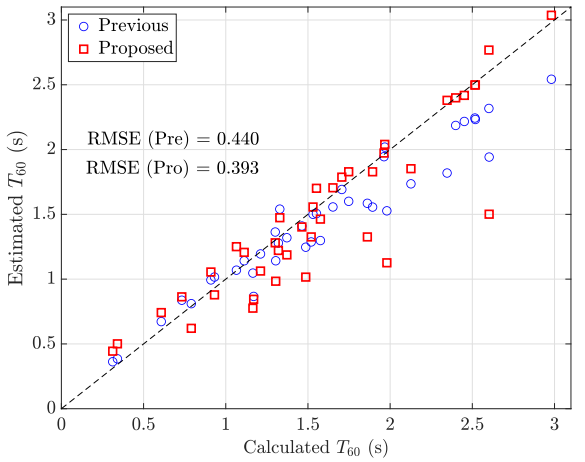


Fig. 12. Example of MTF estimated from reconstructed RIR. Dashed lines are estimated MTFs where "o" indicates MTFs estimated from simulated room and "*" is MTF estimated from real room. Solid line is ground-truth calculated from RIR.



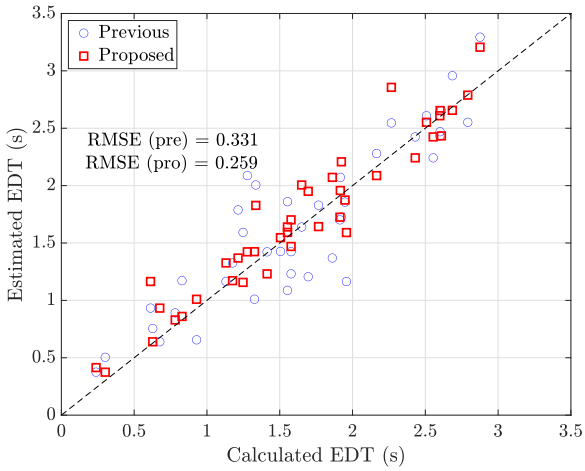


Fig. 13. Estimated reverberation time (T_{60}) and early decay time (EDT) from reverberant speech signals.

The parameters related to the energy ratio of early and late reflection, C_{80}, D_{50}, and T_s, are plotted in Fig. 14, Fig. 15, and Fig. 16, respectively. For C_{80}, the RMSEs were 2.105 with the proposed method and 3.600 with the previous method. For D_{50}, the

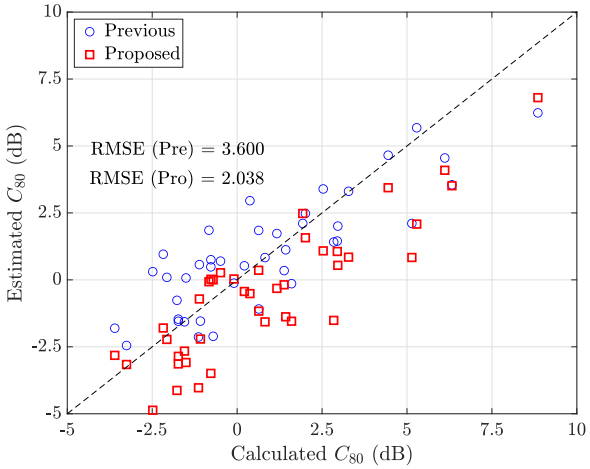


Fig. 14. Estimated clarity index, C_{80}, from reverberant speech signals.

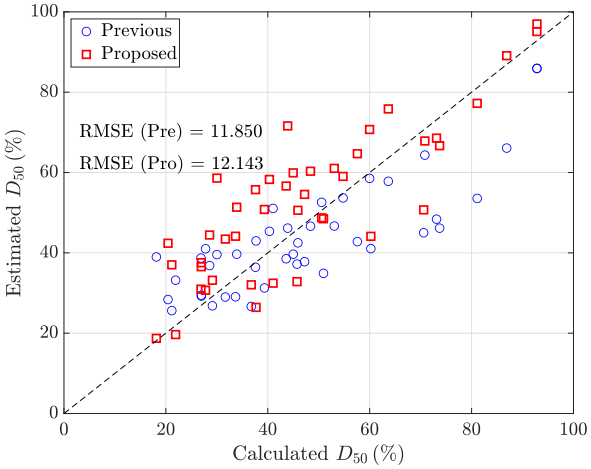


Fig. 15. Estimated Deutlichkeit, D_{50}, from reverberant speech signals.

RMSEs were 2.105 with the proposed method and 3.600 with the previous method. For the estimated T_s, the RMSEs were 0.040 s with the proposed method and 0.043 s with the previous method. These results revealed that the proposed method could estimate these energy-ratios parameters with a higher accuracy Fig. 15.

Fig. 17 plots the STIs estimated from reverberant speech signals. This figure indicates that the estimated STIs were accurate for both methods because they were close to the optimal dashed line. Here, the RMSEs were 0.040 with the proposed method and 0.043 with the previous method.

Table 3 shows the correlation coefficients between the estimated parameters and ground-truths. The results show that the proposed method was closer to the ground-truths than the previous method. This means that it could effectively estimate the parameters and STI from speech signals for realistic room acoustics even if the RIR is not approximated as Schroeder's RIR model.

The accuracy of acoustical parameters related to subjective perception can be represented by the sensitivity of the listeners to a change in a given parameter, called the just noticeable difference (JND) [8]. The JNDs of all acoustical parameters are shown in Table 4. Then, the standard derivation of the estimated error was used for comparison with the JND of each parameter.

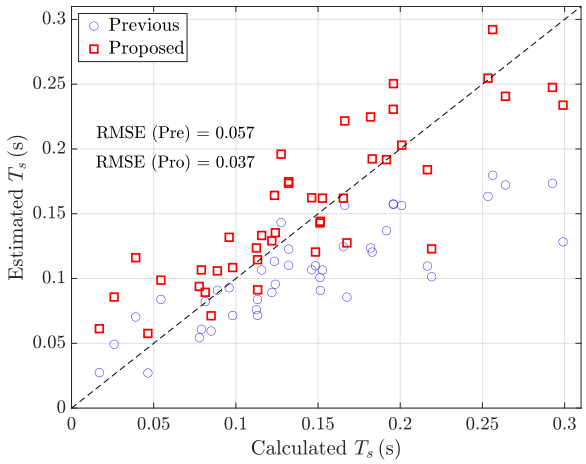


Fig. 16. Estimated center time, T_s, from reverberant speech signals.

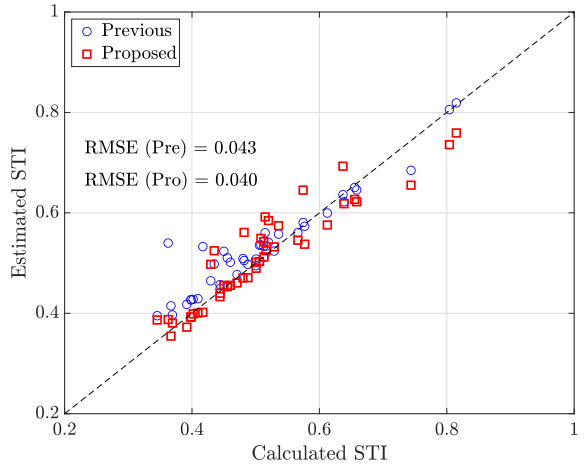


Fig. 17. Estimated speech transmission index, STI, from reverberant speech signals.

5. Discussion

In the above evaluations, the proposed method incorporating the extended RIR model was compared with the previous method based on Schroeder's RIR model. The overall results were improved. However, the advantages, limitations of the proposed method, and some remaining issues concerning the scope of this work need to be discussed.

First, the accuracy of the estimated acoustical parameters and STI depends on the accuracy of the RIR model and its estimated parameters. The estimated five room-acoustic parameters and STI were determined from the approximated RIR. We approximate the unknown RIR on the basis of the impulse response model. Hence, we need to consider which model is appropriate to represent actual RIRs. Schroeder's RIR model was used previously. The method based on Schroeder's RIR model lead to the remaining significant errors compared with the standard measurements. The errors were caused by a mismatch between Schroeder's RIR model and some realistic RIRs. In contrast, the extended RIR model was considered in terms of whether or not it could be a better model. With the fitting parameters of the RIR models from the realistic RIRs, the accuracy of the extended RIR model was higher than Schroeder's RIR model. Consequently, we hypothesize that the accuracy of the estimated room-acoustic parameters based on the extended RIR model would be better. The model mismatch problem is mitigated by incorporating the extended RIR into the proposed method.

Second, the parameters of the extended RIR were estimated simultaneously and effectively. The three parameters of the extended RIR model, i.e., T_h, T_0, T_t, were blindly estimated from the TAE of the reverberant speech signal. The one-dimensional CNNs for the seven-octave bands were used to estimate the three optimal parameters of the extended RIR model simultaneously. In terms of RMSE, the errors of the estimated T_h and T_0 were smaller than T_t. However, the accuracy of the estimated T_h and T_0 was imitated since the percentage errors were slightly higher than the estimated T_t. The percentage errors of $\widehat{T}_h, \widehat{T}_0$, and $\widehat{T}_t$, were 12%, 7%, and 5%, respectively. This is a different resolution problem. In the dataset [39], the mean values of T_h and T_0 are 0.293 and 0.079 s, respectively. In contrast, the mean value of T_t was 1.7 s. The scale of the rising time and peak position parameters (T_h and T_0) was smaller than the reverberation time (T_t). The results show that it is still difficult to estimate different scales of the parameters precisely. The high-resolution parameters of T_h and T_0 need further study.

Third, the estimated acoustical parameters and STI were improved. The parameters that are mainly related to reverberation, i.e., T_{60}, EDT, and STI, were slightly improved. This is because the extended RIR model and Schroeder's RIR model describe the reverberation time by using the same exponential decay function. The center time, T_s, which is related to the center of gravity of the RIR, was improved by about 35%. This significant improvement in estimated T_s is from the correct estimation of the peak positions of the RIRs. Also, the estimated C_{80} was improved by about 40%. It was revealed that the proposed method could overcome the previous issue. It could correctly estimate the acoustical parameters in realistic acoustic environments even though their RIRs decay in non-exponential manner. Note that since the proposed method approximated an unknown RIR in seven-octave bands, the estimated acoustical parameters could be shown for each band according to the requirements of architects, as reported in [29].

However, the estimated D_{50} was insignificantly improved. The reason for the minor improvement is still unclear. Furthermore, a

Table 3
Correlation coefficients between estimated and calculated parameters.

	T_{60}	EDT	C_{80}	D_{50}	T_s	STI
Previous	0.915	0.870	0.918	0.818	0.822	0.902
Proposed	0.918	0.873	0.943	0.903	0.836	0.913

Table 4
Comparison between just noticeable difference (JND) of acoustical parameters and standard derivation (SD) of estimated error [8,40].

Parameter	T_{60}	EDT	C_{80}	D_{50}	T_s	STI
JND	5%	5%	1 dB	5%	10 ms	0.03
SD	9.4%	10.5%	2.7 dB	14%	45 ms	0.05

few outliers of the estimated parameters remain. These outliers are caused by some complicated environments that the RIR models could not represent well. Likewise, the JNDs of all parameters compared with the standard variation of each parameter suggest that human ability might distinguish the difference between the measured and estimated values. Although the JND is required in the standard measurement, many details and specifications are impractical for blind estimation applications, e.g., accuracy of equipment, position of measurement, and level of the source. However, these limitations in accuracy and reliability need further investigation.

Fourth, the computational time and complexity of the proposed method were considered. The proposed method is prone to be applicable for nearly real-time assessments and applications based on two reasons. The first is the short period for a recorded reverberant speech signal. The proposed method needs only five seconds of a reverberant speech signal. The second is that a little computational time is required. We evaluated the computational time on common processors (i.e., Intel Core i7 processors). The STI and five acoustic parameters could be calculated within 0.26 s. Also, the one-dimensional CNNs we used needed significantly less computing power than general two-dimensional CNNs, such as images and spectrogram features. Note that a graphic processing unit was used only in the training process for faster optimization.

Last, we introduce the MTF-based CNN framework because the convolution operation of the CNNs is similar to the basis of the operation between the input signal and the RIR. Since the extended RIR model for approximating an unknown RIR is the main idea, it means that other sophisticated architectures might be used.

6. Conclusion

This paper proposed a blind method for simultaneously estimating the STI and five room-acoustic parameters: T_{60}, EDT, C_{80}, D_{50}, and T_s. The TAE of a reverberant speech signal was used as the input feature according to the MTF concept. One-dimensional CNNs were used to estimate the parameters of an RIR model for seven-octave bands. A more accurate RIR model, the extended RIR model, was used rather than Schroeder's RIR model. Then, unknown RIRs were reconstructed from the estimated parameters of the extended RIR model. Thus, the STI and five room-acoustic parameters could be derived from the reconstructed RIR in each band. Simulations were carried out to determine whether the proposed method could correctly estimate these acoustic parameters in unseen and realistic reverberant environments. The results in terms of RMSEs and correlation coefficients showed that the proposed method could correctly estimate the STI and five room-acoustic parameters. The proposed method outperformed our previous method using Schroeder's RIR model. Moreover, all the estimated parameters were close to the standard measurements. In the future, the accuracy and robustness against background noise and more complicated sound fields will be further investigated.

CRediT authorship contribution statement

Suradej Duangpummet: Conceptualization, Methodology, Software, Investigation, Validation, Visualization, Writing - original draft, Writing - review & editing. **Jessada Karnjana:** Writing - review & editing. **Waree Kongprawechnon:** Writing - review & editing, Funding acquisition. **Masashi Unoki:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by Shibuya Science Culture and Sports Foundation; JSPS-NSFC Bilateral Joint Research Projects/Seminars [JSJSBP120197416]; SCOPE Program of Ministry of Internal Affairs and Communications [Grant No.: 201605002]; Thammasat University Research Fund, Contract No. TUGR 2/35/2562; and SIIT-JAIST-NSTDA Dual Doctoral Degree Program.

References

[1] Escobar VG, Morillas JB. Analysis of intelligibility and reverberation time recommendations in educational rooms. Appl Acoust 2015;96:1–10. https://doi.org/10.1016/j.apacoust.2015.03.001.
[2] Murphy WJ, Xiang N. Room acoustic modeling and auralization at an indoor firing range. J Acoust Soc Am 2019;146(5):3868–72.
[3] Nozaki K, Ikeda Y, Oikawa Y, Fujisaka Y-I, Sunohara M. Blind reverberation energy estimation using exponential averaging with attack and release time constants for hearing aids. Appl Acoust 2018;142:106–13. https://doi.org/10.1016/j.apacoust.2018.08.010.
[4] Ullah R, Shohidul Islam M, Imran Hossain M, Wahab FE, Ye Z. Single channel speech dereverberation and separation using RPCA and SNMF. Appl Acoust 2020;167:. https://doi.org/10.1016/j.apacoust.2020.107406107406.
[5] Tsilfidis A, Mporas I, Mourjopoulos J, Fakotakis N. Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing. Comput Speech Language 2013;27(1):380–95.
[6] Kuttruff H. Room acoustics. Crc Press; 2016.
[7] IEC 60268–16:2020, Sound system equipment – part 16: Objective rating of speech intelligibility by speech transmission index.
[8] ISO 3382:2009, Acoustics – measurements of room acoustics parameters – part 1: Performance spaces.
[9] Houtgast T, Steeneken H. The modulation transfer function in room acoustics as a predictor of speech intelligibility. Acta Acust United with Acust 1973;28(1):66–73.
[10] Steeneken HJ, Houtgast T. A physical method for measuring speech-transmission quality. J Acoust Soc Am 1980;67(1):318–26.
[11] Kendrick P. Blind estimation of room acoustic parameters from speech and music signals, Ph.D. thesis, University of Salford (2009).
[12] Kendrick P. Blind estimation of reverberation time in classrooms and hospital wards. Appl Acoust 2012;73(8):770–80. https://doi.org/10.1016/j.apacoust.2012.02.010.
[13] Choi YJ. Effects of the distribution of occupants in partially occupied classrooms. Appl Acoust 2018;140:1–12. https://doi.org/10.1016/j.apacoust.2018.05.015.
[14] Unoki M, Hiramatsu S. MTF-based method of blind estimation of reverberation time in room acoustics. In: European Signal Processing Conference (EUSIPCO). IEEE; 2008. p. 1–5.
[15] Unoki M, Sasaki K, Miyauchi R, Akagi M, Kim NS. Blind method of estimating speech transmission index from reverberant speech signals. In: European Signal Processing Conference (EUSIPCO). IEEE; 2013. p. 1–5.
[16] Unoki M, Miyazaki A, Morita S, Akagi M. Method of blindly estimating speech transmission index in noisy reverberant environments. J Inf Hiding Multimedia Signal Process 2017;8(6):1430–45.
[17] Kendrick P, Li FF, Cox TJ, Zhang Y, Chambers JA. Blind estimation of reverberation parameters for non-diffuse rooms. Acta Acust united with Acust 2007;93(5):760–70.
[18] Kendrick P, Cox TJ, Li FF, Zhang Y, Chambers JA. Monaural room acoustic parameters from music and speech. J Acoust Soc Am 2008;124(1):278–87.
[19] Li FF, Cox TJ. A neural network model for speech intelligibility quantification. Appl Soft Comput 2007;7(1):145–55.
[20] Seetharaman P, Mysore GJ, Smaragdis P, Pardo B. Blind estimation of the speech transmission index for speech quality prediction. In: Int Conf on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 591–5.
[21] Eaton J, Gaubitch ND, Moore AH, Naylor PA. Estimation of room acoustic parameters: The ace challenge. IEEE/ACM Trans Audio Speech Language Process 2016;24(10):1681–93.
[22] Gamper H, Tashev I. Blind reverberation time estimation using a convolutional neural network, in: 16th International Workshop on Acoustic Signal Enhancement (IWAENC), IEEE, 2018, pp. 136–140.
[23] Deng S, Mack W, Habets EAP. Online blind reverberation time estimation using CRNNs, International Speech Communication Association (INTERSPEECH). IEEE; 2020. p. 5061–5.

[24] Parada PP, Sharma D, Naylor PA. Non-intrusive estimation of the level of reverberation in speech. In: Int Conf on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2014. p. 4718–22.
[25] Parada PP, Sharma D, Lainez J, Barreda D, Waterschoot Tv, Naylor PA. A single-channel non-intrusive C50 estimator correlated with speech recognition performance. IEEE/ACM Trans Audio Speech Language Process 2016;24 (4):719–32. https://doi.org/10.1109/TASLP.2016.2521486.
[26] Gamper H. Blind C50 estimation from single-channel speech using a convolutional neural network. In: International Workshop on Multimedia Signal Processing (MMSP). IEEE; 2020. p. 1–6, 10.1109/MMSP48831.2020.9287158.
[27] Duangpummet S, Karnjana J, Kongprawechnon W, Unoki M. A robust method for blindly estimating speech transmission index using convolutional neural network with temporal amplitude envelope, In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE; 2019. p. 1208–14, 10.1109/APSIPAASC47483.2019.9023209.
[28] Allen JB, Berkley DA. Image method for efficiently simulating small-room acoustics. J Acoust Soc Am 1979;65(4):943–50.
[29] Architectural Institute of Japan, Benchmark problems for acoustical parameters. URL:http://news-sv.aij.or.jp/kankyo/s24/benchmark/.
[30] Barron M. Interpretation of early decay times in concert auditoria. Acta Acust united with Acust 1995;81(4):320–31.
[31] Duangpummet S, Karnjana J, Kongprawechnon W, Unoki M. Blind estimation of room acoustic parameters and speech transmission index using MTF-based CNNs. In: European Signal Processing Conference (EUSIPCO). IEEE; 2021. p. 181–5.
[32] Schroeder MR. Modulation transfer functions: Definition and measurement. Acta Acust united with Acust 1981;49(3):179–82.
[33] Unoki M, Zhu Z. 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 2020;41(1):233–44.
[34] Mourjopoulos J, Hammond J. Modelling and enhancement of reverberant speech using an envelope convolution method, 8. IEEE; 1983. p. 1144–7.
[35] Yu C, Zhang C, Xie L. An envelope signal based deconvolution algorithm for ultrasound imaging. Signal Process 2012;92(3):793–800. https://doi.org/10.1016/j.sigpro.2011.09.024.
[36] ANSI S1.11-2004, Specification for octave-band and fractional-octave-band analog and digital filters.
[37] Takeda T. Speech database user's manual ATR techical report, TR-I-0028.
[38] Goodfellow I, Bengio Y, Courville A. Deep learning. The MIT Press; 2018.
[39] Architectural Institute of Japan, Sound library of architecture and environment, in: Gihodo Shuppan Co., Ltd, Tokyo, 2004.
[40] Bradley JS, Reich R, Norcross S. A just noticeable difference in C50 for speech. Appl Acoust 1999;58(2):99–108. https://doi.org/10.1016/S0003-682X(98)00075-9.