Master's Thesis

Method for blindly estimating stochastic model of room impulse response from reverberant speech

Lijun, Wang

Supervisor: Masashi Unoki

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

Jan., 2022

Abstract

Evaluation of sound quality and listening difficulty in an auditory space attracts attention in the field of room acoustics since the evaluation guides the design of the listening space and helps the acoustic engineers have a picture of the room acoustic characteristics (RAC) of a sound field. Different sound fields have different RACs, resulting from the different designs of the auditory spaces for different purposes. In the concert hall, the auditory space is designed for spacious and transparent sound to create a sense of ethereality and mystery. The lecture hall specifies clear and intelligible sounds to convey an accurate speech message. The acoustic design that meets the different purposes requires a grasp of the physical properties of auditory space to carry out corresponding acoustical treatments.

The RAC in an auditory space is strongly related to daily life since it affects the sound transmission in a sound field. When sound waveforms transmit in an auditory space, the walls and ceilings reflect the waveforms, and other sound sources infer the waveforms, resulting in reverberation and noise. The reverberation and background noise deteriorates intelligibility and clarity of the sound. Intelligible and accurate sound transmission is fundamental to the functionality of communication in an auditory space, especially for the emergency announcement system when suffering disasters (e.g., earthquakes or shootings) and public addresses. From the perspective of the system, the premise for achieving an intelligible sound transmission system is to understand the RAC of a sound field.

Room impulse response (RIR) fully describes the RAC of a sound field in the time domain. The modulation transfer function (MTF), from another perspective, represents the RAC in the frequency domain. A few room acoustic parameters (RAPs), which are derived from the RIR and MTF, have been investigated and standardized to predict the subjective perception of a sound field in spite of speech intelligibility, sound quality and listening difficulty. IEC 60268-16:2020 specifies the definition and calculation of the STI based on the concept of the MTF. ISO 3382:2009 specifies the definition and measurements of five RAPs, including reverberation time (T_{60}), early decay time (EDT), clarity (C_{80}), Deutilichkeit (D_{50}) and center time (T_s). Hence, measuring the RIR of a sound field is essential. However, since it is difficult to measure RIR in daily occupied spaces, blind estimation of RIR and further STI and RAPs without measurement must be resolved as it is an imperative and challenging issue.

Blind estimation is *de facto* the inverse problem to deduce the system from the observed signal only. In general, using some parameters models the system. Thus, the issue of how blindly estimating the system is converted into the issue of how to estimate the parameters of a model, lowering constraints of the inverse problem and making the system become deductible. Here, the RIR model is used to approximate an unknown RIR. There are two common-used RIR models. The one is the image-source RIR model that mimics the reflect paths when the sound waves transmit another in an enclosure. Another is the stochastic RIR model that modulates the RIR as the temporal amplitude envelope (TAE) and temporal fine structure (TFS). The former has limitations in modeling the RIR of a sound field where the people are included and of modeling the RIR in an irregular-shape auditory space. Hence, in the blind estimation of the RIR and further RAPs, this work chose to use the stochastic RIR model to approximate an unknown RIR. Several stochastic RIR models have been proposed to approximate an unknown RIR, including Schroeder's RIR model, the generalized RIR model, and the extended RIR model. Although existing blind methods can estimate RIR, the mismatch of the RIR model limits their performance due to the poor approximation of an unknown RIR. Additionally, the learning-based previous work is absent for traceability and hard to tune when the real environments differentiate from training data used to derive the model.

This paper proposes a deterministic method to blindly estimate an unknown RIR and further the STI and five RAPs from an observed speech signal in which the extended RIR model approximates RIR. The proposed method formulates the temporal power envelope (TPE) of an observed reverberant speech signal to obtain the optimal parameters for the RIR model based on the concept of MTF. Assuming the sound field as a linear time-invariant system, the TPE of the input signal is modeled according to the superposition principles. Then, the reverberation process is formulated by using the modeled TPE of the input signal and the modeled RIR. Here, it is clarified how the parameters of the RIR affect the sound waveforms when transmitting in a sound field and what kind of waveforms are observed at the receiver position (reverberant signals). Furthermore, the dereverberation process is modeled via constructing the formulae of the restored TPE based on the concept of the inverse filtering process. It is found that when the parameters controlling the RIR model used in dereverberation are identical to the parameters that control the RIR model used in the reverberation, the envelopes of the restored TPE are invariant with time. Instead, when the parameters used in dereverberation are not equal to the parameters in the reverberation, the envelopes are time-varying, which can be approximated by the slopes of the envelopes. Thus, the relationship between the parameters

of the RIR model and the observed signal was created. Then, the author proposed the blind estimation method by using the aforementioned relationship, called the alternating estimation strategy (AES) since the parameters controlling the RIR model alternate to be blindly estimated. The proposed method utilizes some basic tools from signal processing, including the Hilbert transform, filter design and linear prediction. The idea behind the proposed method is to cover all possible parameters to carry out the inverse filtering process to find the optimal parameters at which the slopes of the envelopes of the restored TPE are minimized.

Simulations evaluate the proposed method in reverberant environments. The AM signals and speech signals were used for evaluations. The reverberant environments come from the RIR dataset. The root-mean-square errors and Pearson correlation coefficients between the estimated and authentic results were used to evaluate the proposed method with the previous method comparatively. Evaluation results indicated that the proposed method could blindly estimate the parameters that model the RIR and the STI and RAPs effectively without any training.

Contents

List of Figures

List of Tables

List of Notations

The next list describes several notations that will be later used within the body of the thesis.

$\delta(t)$ Dirac delta function

ϵ Infinitesimal

λ Regularization coefficient

ϕ_k and θ_k Phase components at the index k used in modeling the temporal power envelope (TPE) of the original and observed signal and the envelopes of the restored TPE from the observed signal

$\psi(t, T_h, T_t)$ A function used in modeling the envelopes of the restored TPE from the observed signal

σ or $\boldsymbol{\sigma}$ Optimal predictor of the whitening filter

$\mathbf{c}(t)$ White Gaussian noise carrier

$\mathbf{I}$ Identity matrix

$\mathbf{r}$ A vector used in obtaining the optimal predictor

a Constant gain in RIR model

b_{upr} and b_{lwr} Constant factor used in approximating the upper and lower envelopes

C_k Constant gain at the index k used in modeling the TPE of the original and observed signal and the envelopes of the restored TPE from the observed signal

C_{80} Clarity index in speech

D_{50} Deulitchkeit

E Expectation

$e_h(t)$ Temporal amplitude envelope (TAE) of the RIR

$E_h(z)$ Infinite-impulse-response of the extended RIR model

$e_h^2(t)$ TPE of the RIR

$e_x^2(t)$ TPE of the original signal

$e_x^2[n]$ Restored TPE

$e_y^2(t)$ TPE of the observed signal

$E_{h,\mathrm{inv}}(z)$ Inverse filter of the extended RIR model

$e_{x,\mathrm{upr}}^2(t)$ and $e_{x,\mathrm{lwr}}^2(t)$ Upper and lower envelopes of restored TPE from the observed signal

$h(t)$ RIR

h_{ext} Extended RIR model

h_{gen} Generalized RIR model

h_{Sch} Schroeder's RIR model

i Index of p, the number of predictor order of the whitening filter

k Index of K components used in modeling the TPE of the original and observed signal and the envelopes of the restored TPE from the observed signal

$m(f_m)$ Modulation transfer function at modulation frequency f_m

n Sequence length used in whitening process

$R_{e_x^2}$ or $\mathbf{R}$ Autocorrelation of the TPE of the restored TPE

$s(t)$ Arbitrary excited signal used in RIR measurement

$s_r(t)$ recieved signal used in RIR measurement

S_{upr} and S_{lwr} Upper and lower slope used in approximating the upper and lower envelopes

T Time interval used in modeling the TPE of the original and observed signal and the envelopes of the restored TPE from the observed signal

t_0 Parameter to promise casualty of the extended RIR model

T_h Rise envelope parameter of the extended RIR model

T_R or T_{60} Reverberation time

T_s Center time

T_t Decay envelope parameter of the extended RIR model

$u(t)$ Unit-step function

$W(z)$ Whitening filter

$w_x^2(t)$ or $w_x^2[n]$ Whitened restored TPE

$w_{\text{env}}^2[m]$ Envelope of whitened restored TPE at the index m of the total length N used in slope calculation

$x(t)$ Original signal

$y(t)$ Observed signal

Chapter 1

Introduction

1.1 Background

Speech communication underlies the world's connection to maintain the relationship between people via listening and speaking. We communicate daily in venues, such as train stations, office rooms, concourses, and lecture halls. In such enclosures, people talk with each other, hold the conference, listen to addresses, appreciate the music, and deliver their thoughts to the world. The world is linked by communication to consist of the community and further the society. Speech communication and recognition are fundamental for interaction and collaboration between people and people and people and machines, such as public broadcasting, speech translation, and car navigation systems by voice. On this level, human society is composed of communication. In addition, we humanity also live in the physical world, i.e., the enclosures, consisting of walls, floors, and ceilings. Almost all communications occur in these kinds of acoustical spaces (venues as mentioned above). We, humans, convey our thoughts via the articulatory system to produce the speech signals and the listeners receive the signals using the auditory system, as well as the machines radiate the music or speech signals through the loudspeakers and the microphones gather the sounds The speech signals between the speakers and listeners transmit in the enclosures in the form of waveforms from sound sources (speakers) to sound receivers (listeners) via the vibration of air molecules. From the system engineering's perspective, the enclosure in which the sound signals transmit is abstracted as a sound field, a system connecting the input (speakers) and the output (listeners).

Since people live and work in enclosures surrounded by walls and ceilings, the physical properties of the sound field affect people's lifestyles in many aspects. The sound signals reflect and diffuse when encountering the

surface during the transmission within the enclosures, causing reverberation and reflections across the sound field [1]. Moreover, the sound from the other sources interferes to cause unwanted noise. Inappropriate echo and noise occurring in a sound field degrade the intelligibility of speech and clarity of sound when the sound reaches the position of listeners, resulting in miscommunication and affecting the enjoyment of music [2, 3]. Clear sound transmission is the basis of the functionality of communication, especially for emergency announcements and public addresses. Intelligible speech is essential to convey speech messages accurately, avoiding miscommunication.

The sound fields of different architects are designed for different purposes so that the acoustic quality of each of them is considered differently. For example, a conference room is designed for clarity and intelligible sound, while a concert hall is designed for clear and transparent sound. Thereby, different sound fields have different expectations of acoustic effects, but they all have one objective: intelligible sound and avoidance of miscommunication.

To achieve this objective, many researchers focused on designing the physical space of the architecture to accomplish the intelligible sound in a sound field [4, 5, 6]. A good sound field design can ensure intelligible speech transmission by reducing reverberation and absorbing noise via architectural acoustic treatment. For instance, architectural acoustics in concert halls and church buildings are specifically designed to ensure an excellent listening and communication experience. Furthermore, signal processing-based methods are massively applied to achieve reverberation reduction and noise suppression by employing pre- or post-processing in the speaker location or listener location to make the sound robust against the reverberation and noise or to offset the effects of them. Compared to architectural treatment, these kinds of methods are economic and flexible, ignoring the limitation of the geometry of the sound field, which is widely accepted.

Additionally, the acoustic characteristic of a sound field hauls some attraction since having a systematic picture of present room acoustics promises further de-reverberation and de-noise to a sound field. With regard to the research issues, the related works included quality and intelligibility of sound measures, echo cancellation, noise suppression, sound source localization, and source separation. The works related to these research issues are widely applied in automatic speech recognition, speech enhancement, speaker detection and verification, hearing aids, and entertainment (sound reproduction systems). Principally, acquiring the room acoustics of a sound field is the basis for solving the issues above and predicting the listening experience in a sound field.

Subjective listening experiments are generally conducted to acquire the characteristics of room acoustics and predict the listening experience, i.e., the

perception of the sound quality (clarity and transparency) and speech intelligibility in a sound field [1]. The attendees are required to score the sound they hear at each iteration in a specific auditory space. Statistics process the scores to conclude the subjective evaluation representing the listening experience in this space. However, the listening experiment is expensive and tedious, taking much time to experiment as well. Subjective scores are also unreliable and ambiguous in describing the room acoustics, bringing difficulty in giving an effective response to guide the architectural design or acoustical treatment of a sound field or diagnose the acoustic problems.

Instead of subjective feelings about a sound field, the characteristics of room acoustics can be described by objective measures of the physical attributes of room acoustics of a sound field. Commonly, an impulse signal is excited in a sound field to generate the corresponding impulse response that records the direct and reflected waveforms transmitted in a sound field originating from the impulse signal source. The impulse response signal is called room impulse response (RIR), fully reflecting the physical properties of room acoustics, including the geometry and surface characteristics such as volume, shape, and absorption coefficient [7, 8].

From the points of view of signal and system theory, RIR can be regarded as a transfer function between the source location and the receiver location in a sound field in the time domain, i.e., abstracting a sound field as the transmission system. This transfer function could be represented in the frequency domain as the modulation transfer function (MTF) [9, 10].

Many objective parameters and indices are widely used to predict the subjective perception of the sound quality and speech intelligibility of a sound field, called room acoustic parameters (RAPs) [1, 11, 12]. These RAPs can be calculated from RIR or MTF directly, which is fast and responsive compared to subjective listening experiments. The measurements and calculation of RAPs have been investigated and standardized by IEC and ISO [13, 14]. The RAPs can be used to analyze a sound field of an auditory space to provide a guidebook for optimizing this sound field via architectural treatments or signal-processing algorithms to achieve clear and intelligible sound for architects, artists, and acoustic engineers [15].

1.2 Motivation

The motivation for this study stems from my curiosity about room acoustics, especially for elucidating how humans perceive a sound field and whether it is possible to extend the boundary of human perception by utilizing the characteristics of a sound field. Furthermore, it is explicated how to model

the system of a sound field by what indices and parameters. Then, is it possible to use the model to actualize blind estimation of the room acoustic characteristics (RAC) of a sound field from an observed signal? The premise that we control and utilize the RAC of a space is the physical properties of this auditory space and the clarification of human perceptions in it.

Since RIR fully represents the RAC of a sound field and underlies the computation of the RAPs, measurement of RIR is essential for having a systematic picture of a sound field and predicting the sound quality and speech intelligibility.

Commonly, broadband excitation signals are used to be excited in a sound field to obtain the corresponding RIR because impulse signals in realistic environments are hard to realize. The measurement needs the space where the people are excluded. However, it is difficult to measure in a space where people, such as common-used public spaces, cannot be excluded. Hence, estimating RIR and RAPs from an observed signal is imperative. The recorded sound (speech or music) in space is used to derive the RIR and RAPs without measurement of RIR and RAPs, which is called *blind estimation*.

Blind estimation can be applied in the spaces where people are either excluded or included. Additionally, the blind estimation can be realized as real-time or quasi-real-time processing, which extends its application for adaptive estimation. For instance, the RAC of the sound field differs following the changes in people's location and numbers in an enclosure during the day. Direct measurement is challenging to handle this situation. Instead, the blind estimation can adaptively estimate the RAC. Because of the immense adaptability of blind estimation, it not only prevails in room acoustics but also in other fields, such as system identification [16], input recovery [17], and adaptive system control [18, 19].

1.3 Objective

Blind estimation is a challenging issue since it, in essence, is an ill-positioned inverse problem that derives a system from the output only without the constraints from the input. The common way is to model the system to create the mapping between the output and the system by either mathematics-based explicit schemes or learning-based implicit schemes.

With regard to this challenging issue, some blind estimation methods with corresponding parameters and indices have been proposed, which are detailed in Chapter 2, are insufficient to describe the RAC of an auditory space or lack the explicit explanation of the relationship between the observed signal and the sound field. In addition, the concept of MTF mentioned earlier can

regard the sound field as a transfer function from the original signal (speaker) to the observed signal (listener), which lowers the difficulty of modeling the whole system. Does it work to incorporate the concept of MTF into building the relationship between the observed signals and the sound field? This study attempts to solve the issues above, i.e., proposes a deterministic method, on the basis of the concept of MTF, to blindly estimate the RIR and several useful RAPs and indices that can describe the RAC comprehensively.

1.4 Thesis outline

The thesis is organized as illustrated in Figure 1.1. The whole organization is composed as follows.

Chapter 1 introduces the background related to the topic discussed in the thesis, including a brief introduction to room acoustics, its definition and importance, and the basic concepts used in this field. Moreover, the motivation of this study is discussed, and the current challenging issues are presented.

Chapter 2 reviews the related works in blind estimation in room acoustics, which include the measurement and modeling of room impulse response, the modulation transfer function and its concept, the room acoustic parameters, and the blind estimation methods for RIR and RAPs.

Chapter 3 proposes a blind estimation strategy for blindly estimating RIR and RAPs, starting from modeling the relationship between the observed reverberant signals and the sound field to the proposed method's details.

Chapter 4 gives the verification and evaluation of the proposed method, including the dataset used for evaluation, evaluated conditions, and estimation results for the parameters of the RIR and RAPs.

Chapter 5 summarizes the whole work conducted during the master's program, points out the significance and contributions of this work and presents the remaining works left.

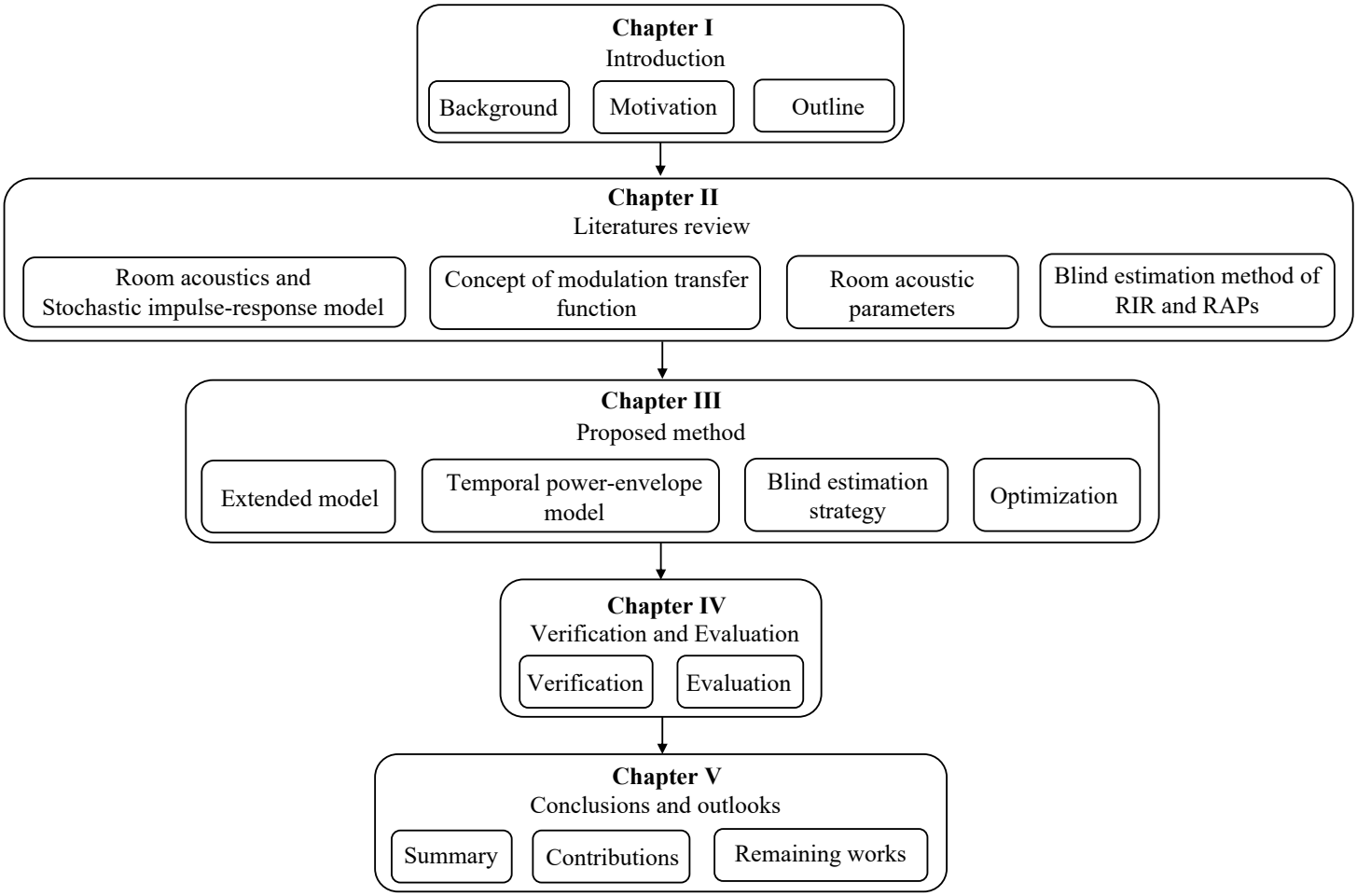

Figure 1.1: Organization of the dissertation

Chapter 2

Literatures Review

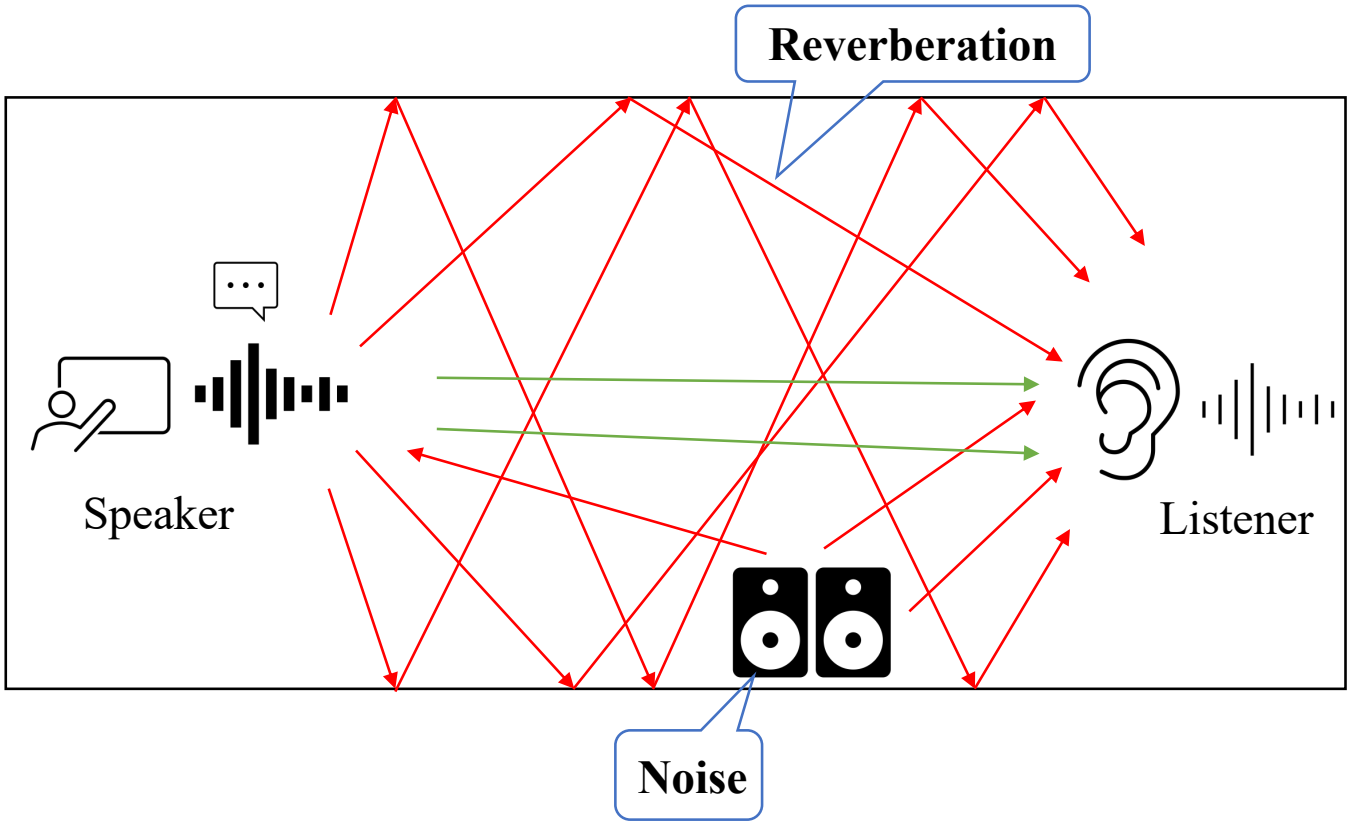

Figure 2.1: Room acoustics of a sound field in an enclosure.

2.1 Room acoustics and stochastic impulse-response model

Room acoustics is the research field that studies computation, modeling, and simulation in spite of sound fields in enclosures, leading to acoustical design of the auditory spaces and materializing the auralization of the virtual auditory spaces [1]. An illustration of room acoustics in an enclosure is shown in Figure 2.1. In this section, the measures of RAC, i.e., the RIR of a sound field and the corresponding stochastic model, are introduced.

2.1.1 Measures of room impulse response

In an enclosure, the sound waves radiate across all directions from the source to the receiver position. This process can be formulated as the convolution of the source signals and the room impulse response (RIR), which is defined as:

$$y(t) = \int_0^\infty x(t-\tau)h(\tau)d\tau \triangleq x(t) * h(t) \tag{2.1}$$

where $x(t)$ and $y(t)$ denote the original and the observed reverberant signal, respectively. $h(t)$ denotes an RIR, and " $*$ " denotes the convolution operation, respectively.

The RIR is the essential and full representation of the RAC of a sound field, recording paths of direct and reflected sound waves transmitted in the enclosures as the time-delayed impulse train. Figure 2.2 shows a conceptual illustration of how a sound field generates the RIR and which part the RIR consists of.

Literally, RIR can be obtained by gathering the impulse response signal from an impulse excitation signal. However, in realistic measurements, it is difficult to produce an ideal impulse response, and also gathering an impulse response is so challenging that modern recording techniques are limited to having a precise and excellent-resolution recorded signal of impulse response since the elapsed time of impulse response is short and sensitive to the background noise. Therefore, instead of using impulse signals, broadband signals are used as excitation to measure the RIR, such as sine sweep signals or maximum length sequences, as shown in Figure (2.3) [20, 21].

For these specific excitation signals, the measurement is specified via cross-correlation-based methods. Suppose a sound field, i.e., a physical room, is excited by arbitrary signal $s(t)$.

$$s_r(t) = \int_{-\infty}^\infty s(\tau)h(t-\tau)d\tau = \int_{-\infty}^\infty s(t-\tau)h(\tau)d\tau \tag{2.2}$$

where $s_r(t)$ denotes the received signal in measurement. The cross-correlation function of $s(t)$ and $s_r(t)$ is given by:

$$\Phi_{s,s_r}(t') = \int_{-\infty}^{\infty} h(\tau)\left[\lim_{T_0\to\infty}\frac{1}{T_0}\int_{-T_0/2}^{+T_0/2} s(t)s(t+t'-\tau)dt\right]d\tau \tag{2.3}$$

$$= \int h(\tau)\Phi_{s,s}(t'-\tau)d\tau \tag{2.4}$$

where $\Phi(t)$ represents the cross-correlation function. Eq. (2.4) implies that the measurement yields the RIR if the autocorrelation of the excitation signal $s(t)$ is equal to or at least approximated to the delta function. Sweep sine signal and maximum length signal are appropriate to the requirement.

Rewrite Eq. (2.2) as the discrete-time form as, of which the length is l:

$$s_r = \sum_{j=0}^{l-1} s_j h_{k-j} \tag{2.5}$$

or as the matrix form

$$\mathbf{s}_r = \mathbf{S}\cdot\mathbf{h}, \tag{2.6}$$

where S is a cyclic arrangement as:

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & s_0 & s_1 & \cdots & s_{l-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_{l-1} & s_0 & \cdots & s_{l-2} \end{bmatrix} \tag{2.7}$$

The solution is given by:

$$\mathbf{S}\cdot\mathbf{s}_r = \mathbf{S}\cdot\mathbf{S}\cdot\mathbf{h} = (l+1)\mathbf{h} \tag{2.8}$$

$$\mathbf{h} = \frac{1}{l+1}\mathbf{S}\cdot\mathbf{s}_r \tag{2.9}$$

by using the fact that $\mathbf{S}\cdot\mathbf{S} = (l+1)\mathbf{I}$. Then the correlation function of a room excited from excitation, i.e., $\mathbf{S}$, can be measured by the play-back method [1]. Figure 2.4 shows the block diagram of the measurement.

2.1.2 Schroeder's model

Blind estimation of the RIR prerequires the modeling of the RIR to create the mapping between the observed reverberant signal and the RIR. There are two schemes of RIR models. The one is based on image-source method to simulate an unknown RIR by regarding a reflection path as a direct-sound

path from the image source. Another is to use the statistical model to acquire the stochastic model of the RIR, viewing the RIR as the modulation of the temporal envelope and the stochastic carrier signal.

Here, the author focuses on the stochastic RIR model only since the stochastic RIR model can be used to assess quality of sound and intelligibility of speech in room acoustics, such as clarity and speech transmission index (STI) [22, 10]. Several applicable stochastic RIR models have been investigated [23, 24].

With regard to the stochastic RIR model, the most well-known model is Schroeder's RIR model, which approximates the exponential decay envelope of a realistic RIR. It uses a single parameter to control this decay envelope.

The main limitation of Schroeder's model is its shortage of approximation for the onset transition of a realistic RIR, causing the mismatch of RIR approximation. Schroeder's RIR model is defined as:

$$h_{\text{Sch}} = e_{h,\text{Sch}}(t)\mathbf{c}(t) = a\exp(-6.9t/T_R)\mathbf{c}(t) \tag{2.10}$$

where T_R is reverberation time, controlling the exponential decay envelope of RIR, $e_{h,\text{Sch}}(t)$ is the temporal amplitude envelope (TAE) of the RIR, $\mathbf{c}(t)$ is the white Gaussian noise (WGN) carrier signal [25], and a is the gain factor.

2.1.3 Generalized model

The generalized RIR model [24], is modified based on Schroeder's RIR model to assuage the limitation of Schroeder's model by introducing another parameter to approximate the onset transition of a realistic RIR additionally. The generalized RIR model is defined as:

$$h_{\text{gen}}(t) = e_{h,\text{gen}}(t)\mathbf{c}(t) = at^{(b-1)}\exp(-6.9t/T_R)\mathbf{c}(t) \tag{2.11}$$

where T_R and b are the parameters that control the exponential decay side and exponential rise envelope, i.e., the onset transition, respectively. $e_{h,\text{gen}}$ is the TAE of RIR, and a is gain factor. Figure 2.5 shows a comparison between the fits of Schroeder's RIR model and the generalized RIR model with the measured realistic RIR in terms of temporal power envelope (TPE).

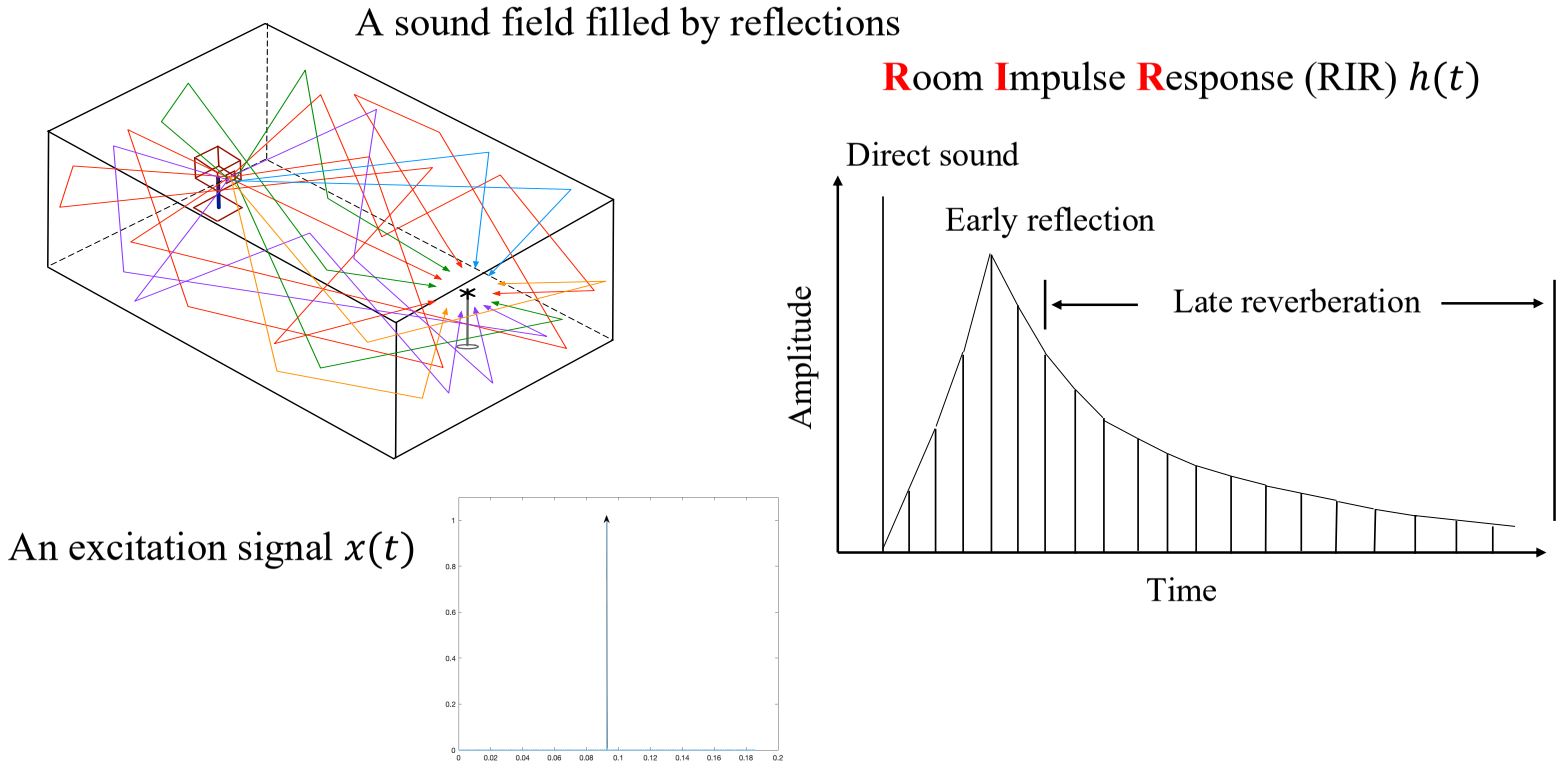

Figure 2.2: Illustration of room impulse response in a sound field.

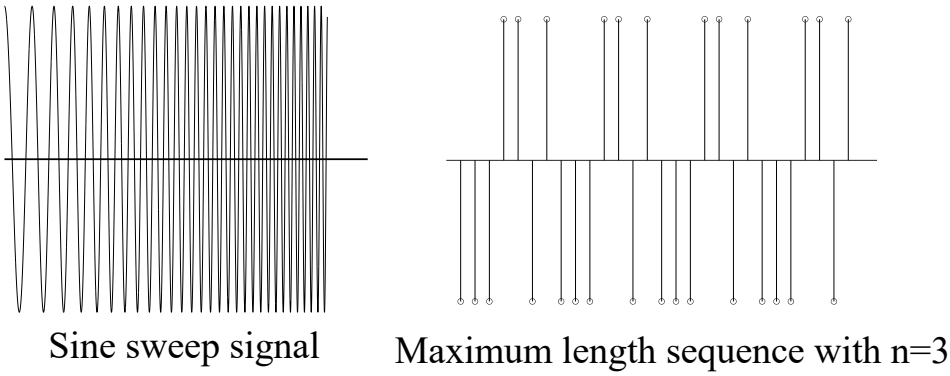

Figure 2.3: Sine sweep signal and maximum length signal as the excitation signals.

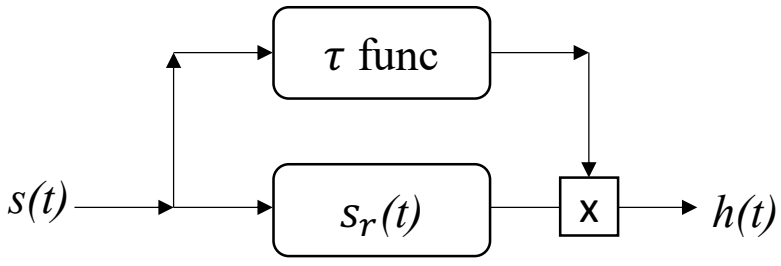

Figure 2.4: Measurement of the RIR by cross-correlation.

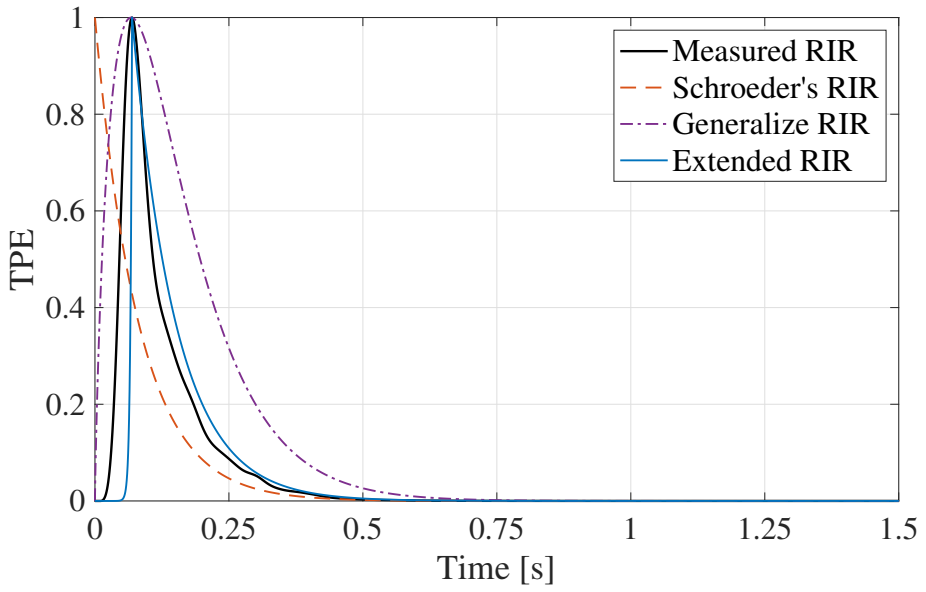

Figure 2.5: Results for fitting measured RIRs with three RIR models, Schroeder's RIR, generalized RIR, and extended RIR models.

2.2 Concept of the modulation transfer function

In this section, the modulation transfer function (MTF), which fully represents the RAC of a sound field in the frequency domain, is briefly reviewed, and the concept of the MTF, which is helpful for modeling the system, is introduced.

2.2.1 Overview of the modulation transfer function

The MTF is a concept widely applied in physical, acoustic, and optic fields [26, 27]. M. R. Schroeder defined the MTF in room acoustics to quantify the reverberation effect and noise level of a sound field. He also extended the definition of MTF to develop the concept of MTF, proving that the MTF is a useful representation of the RAC in an auditory space and abstracting a sound field in this space as the transmission system, which can be quantified as the MTF, the transfer function of a system as a function of modulation frequencies.

Then, Houstgast and Steenken proposed the objective index, i.e., the speech transmission index (STI), to evaluate the subjective perception of

speech intelligibility in an auditory space based on the concept of MTF [9]. Furthermore, a brand-new objective index, the speech intelligibility index (SII), has been developed, which is correlated significantly with the intelligibility of speech under the conditions of background noise and reverberation based on the concept of the MTF [28].

The RAC of a sound field containing the reverberation and background noise can be described as the MTF of this sound field, as shown in Figure 2.6. The MTF builds up the relationship between the signals from the speaker and the listener by the modulation depth of their modulation spectrums. The formula of the MTF can be derived as the ratio of the power envelope spectrum (modulation spectrum) of an RIR and its total energy, which is determined as:

$$m(f_m) = \frac{\int_0^\infty h^2(t) e^{-j2\pi f_m t} dt}{\int_0^\infty h^2(t) dt} \cdot \left[\frac{1}{1 + 10^{-(\frac{SNR}{10})}} \right] \tag{2.12}$$

where $m(f_m)$ denote the MTF at modulation frequency f_m. The bracketed second term accounts for noise level represented as the signal-and-noise ratio (SNR). Figure 2.6 shows the concept of the MTF.

2.2.2 Temporal envelope

The speech or music signals consist of two portions, the temporal envelope (including temporal amplitude envelope, TAE, or temporal power envelope, TPE) and the temporal fine structure (TFS), from the points of view of psychoacoustics and the psychology of the auditory system. The normal healthy cochlea decomposes the speech and music signals into narrowband signals that can be regarded as the slowly varied temporal envelope and the TFS with rapid oscillations.

The temporal envelope carries the most important information for the understanding and recognition of speech, especially in a quiet environment [29, 30]. Rather in a noisy environment, the TFS also plays an essential role in the recognition of speech [31]. B. C. J. Moore clarified that the temporal envelope information is conveyed in the auditory nerves as the short-term rate of action potentials, as the roles of auditory perception for pitch perception, sound localization, and the perception of speech in background noise [32]. The temporal envelope is significant for the auditory perception of signals. In the frequency domain, the temporal envelope can be transformed into the MTF which is highly related to speech intelligibility.

Hence, it is believed that the temporal envelopes convey important information for the perception of speech and music signals, which can be widely

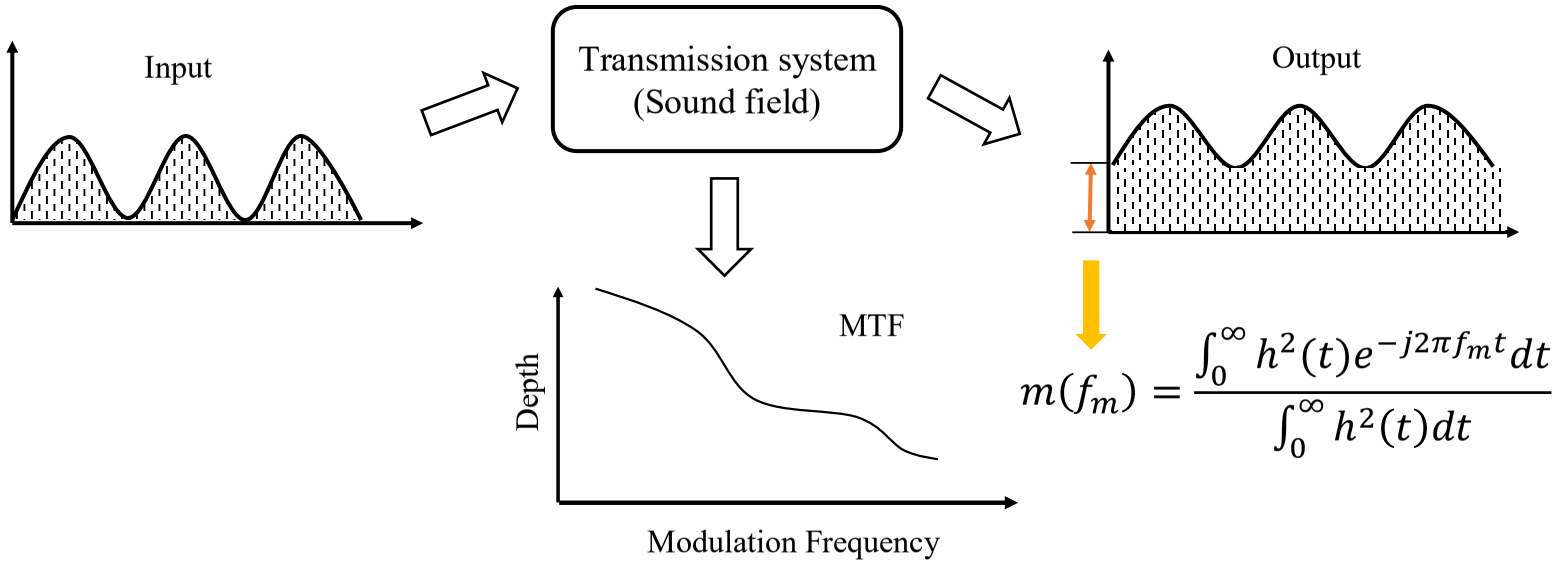

Figure 2.6: Concept of the modulation transfer function to abstract the sound field to the transmission system as a function of the modulation frequency in the reverberation and noise.

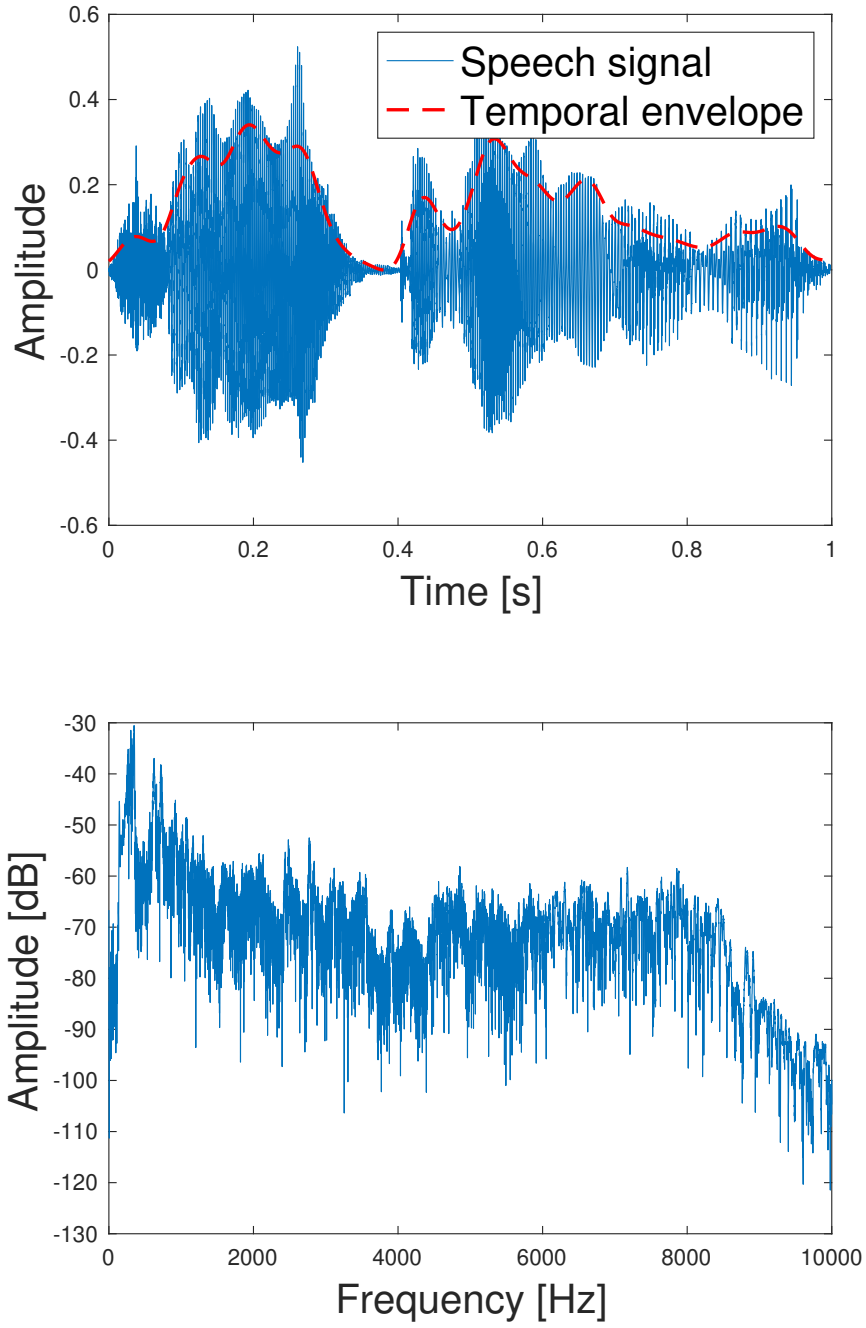

Figure 2.7: Example of a clean speech signal with its temporal envelope and corresponding spectrum.

used in room acoustics. The temporal envelope (TAE) of a signal $y(t)$ can be extracted as:

$$e_y(t) = \mathbf{LPF}\big[|y(t) + j \cdot \text{Hilbert}(y(t))|\big] \tag{2.13}$$

where $e_y(t)$ represents the temporal envelope and Hilbert denotes Hilbert tranform. **LPF** is a Butterworth lowpass filter to smooth the extracted signal. There are different ways to extract the temporal envelope, which yield similar results [32]. Figure 2.7 shows an example of the temporal envelope of a clean speech and the corresponding spectrum.

2.3 Room acoustic parameters

Subjective perception of a sound field, such as speech intelligibility and music clarity, can be evaluated by objective indices and parameters. Several room acoustic parameters (RAPs) are investigated and standardized [1, 13, 14]. These RAPs can reveal the physical properties of auditory space.

2.3.1 Speech transmission index

The speech transmission index (STI) was first proposed by Houstgast and Steeneken, based on the concept of MTF, to predict speech intelligibility when speech signals transmit in a sound field [9, 33, 34]. The MTF regards the sound field as the sound transmission system to convey the sound signals from the speaker position to the listener position under the condition of reverberation and noise, as shown in Figure 2.1 and Figure 2.6. The MTF quantifies the reverberation effect and noise level of a transmission system by Eq. (3.20). The measurement of STI has been standardized in IEC60268-16:2020 [13]. Figure 2.8 shows the block diagram to calculate the STI. First, the RIR passes through the seven-octave filterbank to calculate the MTF at each band by Eq. (3.20), then SNR at each band is calculated as follows:

$$N(k,i) = 10 \log 10\Big[\frac{m_k(f_{m,i})}{1 - m_k(f_{m,i})}\Big] \tag{2.14}$$

where $k = 1, 2, \cdots, 7$ and $i = 1, 2, \cdots, 14$. The transmission index (TI) is determined as follows:

$$T(k,i) = \begin{cases} 1 & N(k,i) < 15, \\ \frac{N(k,i)+15}{30} & -15 \leq N(k,i) \geq 15, \\ 0 & N(k,i) < 15. \end{cases} \tag{2.15}$$

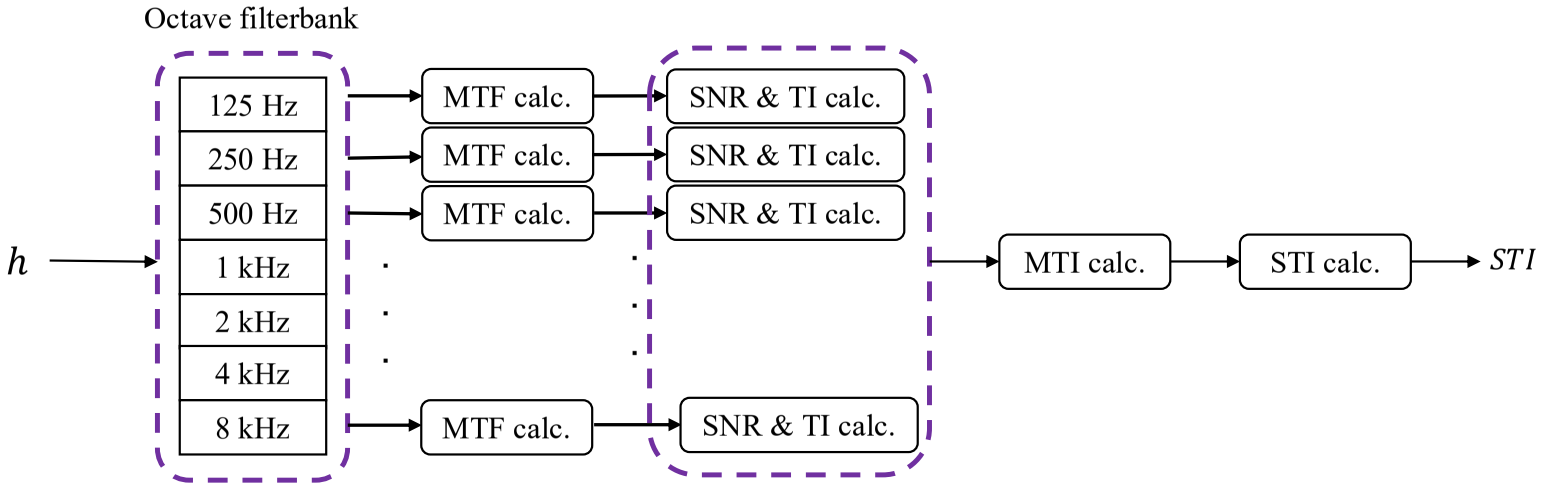

Figure 2.8: The block diagram of STI calculation

The modulation transmission index (MTI) is calculated from the TI as follows:

$$M(k) = \frac{1}{14}\sum_{i=1}^{14} T(k,i) \tag{2.16}$$

Finally, the STI is derived from the MTI:

$$\text{STI} = \sum_{k=1}^{7} \text{Wgt}(k) M(k) \tag{2.17}$$

where $Wgt(k)$ represents the weighting factor at each octave band, listed as Wgt(1) = 0.129, Wgt(2) = 0.143, Wgt(3) = 0.144, Wgt(4) = 0.114, Wgt(5) = 0.186, Wgt(6) = 0.171 and Wgt(7) = 0.143. The scale of STI is within $[0, 1]$. A higher score indicates higher speech intelligibility.

2.3.2 Reverberation time

Reverberation time (T_{60}) is the most essential RAP in room acoustics, characterizing the physical characteristics of a sound field in which energy is transmitted and distributed.

T_{60} can be calculated from the RIR directly via employing Schroeder's back integral method [35], which has been standardized in ISO 3382-1:2009. First, the early decay curve (EDC) of an RIR is calculated by the equation as follows:

$$\text{EDC} = 10\log 10\Big[\int_0^{\infty} h^2(t)dt\Big] \tag{2.18}$$

The least-square curve-fitting is used to fit the EDC at -20 dB, i.e., T_{20}, and then the fitted curve is extrapolated to -60 dB to obtain T_{60}. Figure 2.9 shows an example of calculating T_{60} using an RIR.

2.3.3 Clarity

Clarity (C_{80}) is the RAP to characterize the transparency of music signals transmitted in a sound field, representing the subjectively auditory response to sound clarity. C_{80} is the ratio of early-to-late arrival reflections, which is defined as:

$$C_{80} = 10\log 10\frac{\int_0^{80ms} h^2(t)dt}{\int_{80ms}^{\infty} h^2(t)dt} \tag{2.19}$$

where 80 ms denotes the time boundary between the early reflections and late reverberations. The calculation of C_{80} has also been standardized in ISO 3382-1:2009 [14].

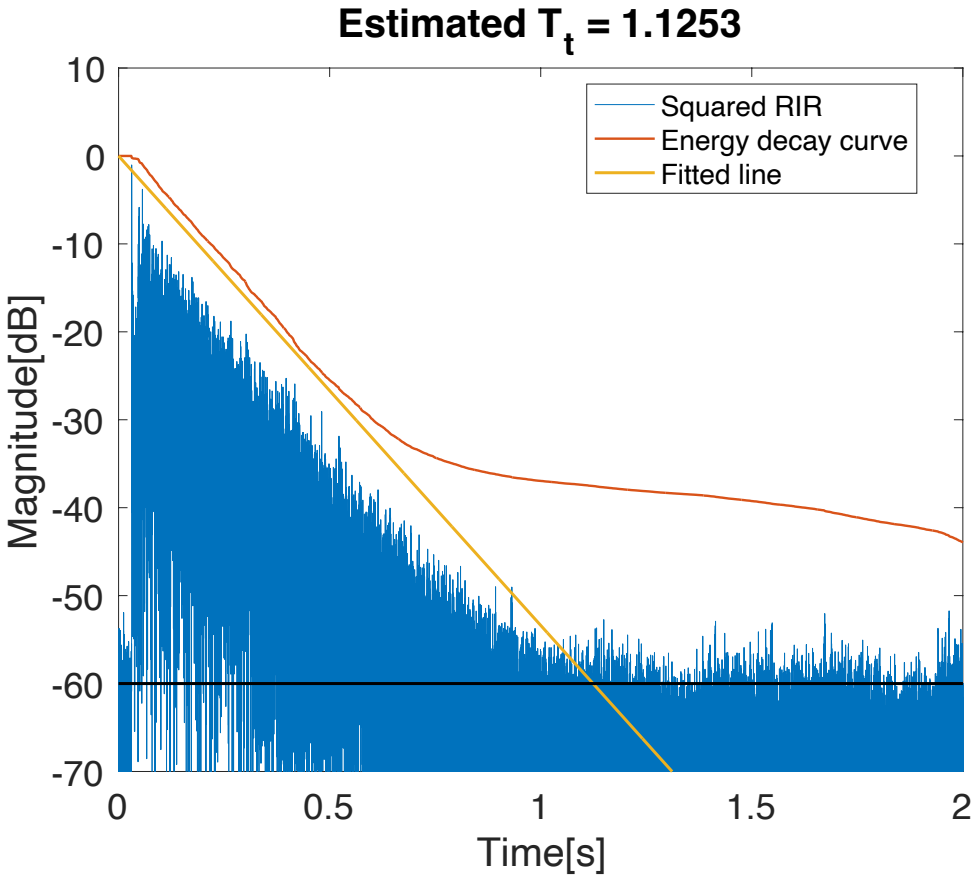

Figure 2.9: Example of calculation of T_{60}

2.3.4 Early decay time

Early decay time (EDT) is the RAP characterizing the duration of sound decay within an initial -10 dB that makes the more important contribution of direct sound and early reflection for the perception of reverberation. Similarly, EDT has been standardized in ISO 3382-1:2009 [14]. The calculation of EDT is similar to the calculation of T_{60}. Instead of fitting to -20 dB, EDC is fitted to -10 dB and extrapolated to -60 dB to obtain EDT.

2.3.5 Deulitchkeit

Deulitchkeit, also known as D_{50}, is the RAP that characterizes speech intelligibility in a sound field, representing the subjective response to speech intelligibility in a sound field. D_{50} has been standardized in ISO 3382 [14], which is the ratio of early-to-total sound energy. D_{50} is defined as follows:

$$D_{50} = \frac{\int_0^{50ms} h^2(t)dt}{\int_0^{\infty} h^2(t)dt} \times 100 \tag{2.20}$$

where the 50 ms is the boundary of the early sound energy. The measurement of D_{50} has also been standardized in ISO 3382 [14].

2.3.6 Center time

The last RAP is the center time (T_s), which characterizes the balance between clarity and reverberation related to speech intelligibility, representing the center of gravity time of decaying energy in a sound field. T_s has been standardized in ISO 3382 and is defined as [14]:

$$T_s = \frac{\int_0^{\infty} h^2(t)tdt}{\int_0^{\infty} h^2(t)dt} \tag{2.21}$$

2.4 Blind estimation methods of RIR and RAPs

Blind estimation can be realized by either analytical methods or learning-based methods. The former is to create the mapping between the observed reverberant signals and the parameters of the RIR model that approximates an unknown RIR by the mathematical derivation and formulation [24, 36, 37, 38, 39, 40]. The latter is to learn the mapping depending on the training data to derive the implicit model [41, 42, 43, 44, 45, 46].

2.4.1 Analytical methods

M. Unoki *et al.* proposed two analytical methods to blindly estimate the STI based on Schroeder's RIR model and the generalized RIR model, respectively [24, 36]. These methods utilized the relationship between the MTF of a sound field and the modulation spectrum of the observed reverberation signals.

R. Ratnam proposed a method by using the energy decay curve of a reverberant signal and combined with maximum likelihood estimators (MLE) to blindly estimate reverberation time (T_R or T_{60}) [38]. P. Kendrick *et al.* developed and revised the method Ratnam proposed to realize blind estimation of reverberation time from the speech or music signals [39].

L. Courveur *et al.* proposed a method to blindly estimate reverberation time by modeling the speech sequences and applied the MLE to optimize to derive the results [40]. Similarly, A. Keshavarz proposed the speech-model-based method based on the similarity between the autocorrelation of the reverberant speech signals and of the original speech signals [37].

2.4.2 Learning-based methods

P. Kendrick *et al.* proposed a method based on artificial neural network (ANN) to blindly estimate the reverberation time, EDT, clarity, Deulitchkeit and center time from speech signals [42]. F. Li also proposed the method based on the ANN to blindly estimate the STI [41].

J. F. Santos *et al.* proposed a method by using recurrent neural networks (RNN) to learn the features that appear in the modulation spectrum to blindly estimate the STI [43]. P. Callens proposed a blind estimation method for reverberation time, clarity and direct-to-reverberation ratio based on long short-term memory (LSTM) [45]. Convolution neural networks (CNN) are widely used in blind estimation. P. Seetharamen *et al.* proposed a blind method for the STI by using CNN. C. J. Steinmetz *et al* combined filter noise shaping and CNN to develop a method for blindly estimating reverberation time [47]. Recently, Suradej *et al.* proposed the MTF-based CNN scheme to simultaneously estimate the STI, reverberation time, clarity, EDT, Deulitchkeit and center time.

2.5 Remaining issues

However, the current works remain some issues. The current methods can estimate only a single parameter [24, 36, 37, 38, 39, 40, 41, 43]. Although the MTF-based CNN method could estimate multiple parameters, it is limited

to the training data used to derive the model, the same as the other learning-based methods [41, 42, 43, 44, 45, 46]. The efficiency of the trained models naturally decreases when the real 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. Additionally, the mismatch of the RIR model for an actual RIR limits the performance due to poor approximation. Therefore, the author proposes 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.

Chapter 3

Proposed method

A blind estimation method is proposed, namely the alternating estimation strategy (AES), shown as a block diagram in Figure 3.3. The details are given as follows.

3.1 Extended model

This section introduces a new stochastic RIR model called the extended RIR model [48, 46]. The extended RIR model, modified based on Schroeder's RIR model, presents the two mutually independent parameters to control the rise and decay sides of the envelope of an RIR. Instead of using a parameter without physical significance to control the onset transition of the RIR, the extended RIR model approximates the onset transition as the exponential rise energy envelope, enabling the modeling of the onset transition to be interpretable. Figure 3.2 shows the fit of the extended RIR model to a measured RIR and the comparison with the fit of Schroeder's RIR model. The extended RIR model is defined as:

$$h(t) = e_{h,\mathrm{ext}}(t)\mathbf{c}(t) = \begin{cases} a\exp(6.9t/T_h)\mathbf{c}(t) & t < 0 \\ a\exp(-6.9t/T_t)\mathbf{c}(t) & t \geq 0, \end{cases} \tag{3.1}$$

$$h_{\mathrm{ext}}(t) = h(t - t_0), t0 \geq 0 \tag{3.2}$$

where T_h and T_t are the parameters controlling the rise and decay exponential envelope of the RIR, respectively. $e_{h,\mathrm{ext}}$ is the TAE of the RIR. t_0 is introduced to promise the causality of the RIR model. Since its scale is close to T_h, t_0 is assumed to be equal to T_h without resulting in noticeable deviations.

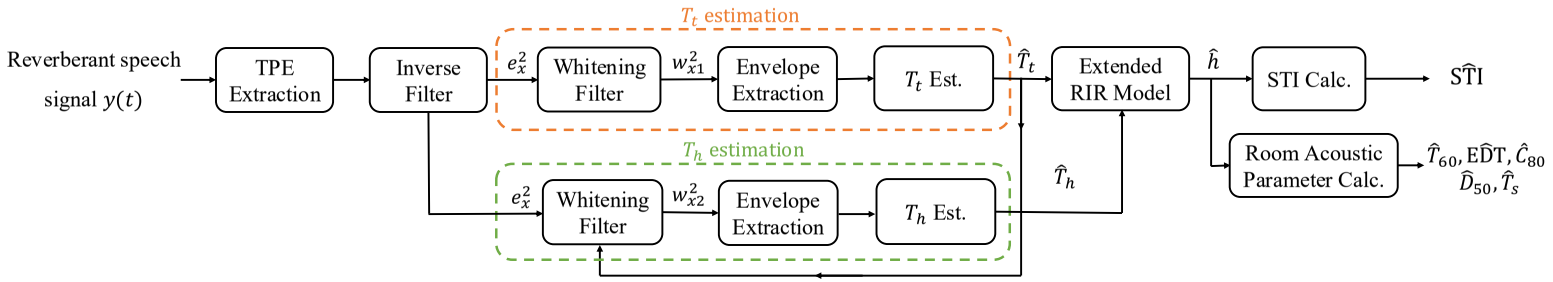

Figure 3.1: The block diagram of the proposed method

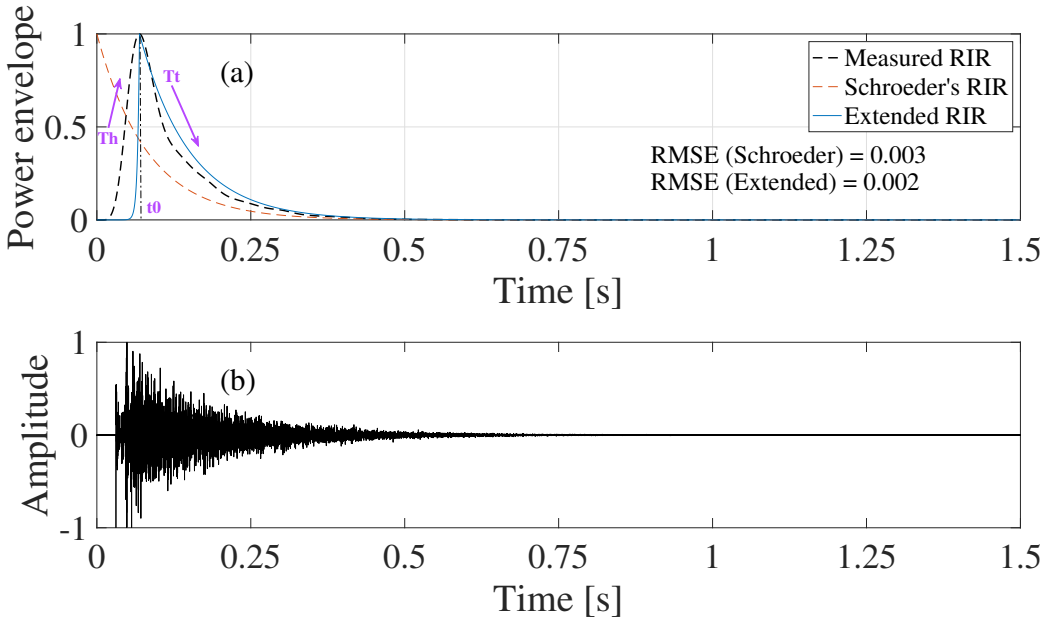

Figure 3.2: Fits of two RIR models with measured RIR: (a) temporal power envelope of RIRs and (b) the corresponding RIR.

3.2 Temporal power-envelope model

As mentioned in Chapter 2, the prerequisite of proposing the blind estimation is to model the transmission system blind estimation as the inverse problem so that the reverberant signal is only observed. In this section, the transmission system from the original signal to the observed signal is modeled on the level of the temporal power envelope (TPE).

3.2.1 Reverberation process

Eq. (2.1) presents the relationship between the original signals and the observed reverberant signals. Then, Eq. (2.1) derives the connection between the TPE of the original signal and of the reverberant signal. Assuming the original signal and RIR can be modeled as the modulation of the TAE and temporal fine structure (TFS) that is regarded to follow the white Gaussian noise (WGN) carrier:

$$x(t) = e_x(t)\mathbf{c}_x(t) \tag{3.3}$$
$$h(t) = e_h(t)\mathbf{c}_h(t) \tag{3.4}$$

where $e_x(t)$ and $e_h(t)$ denote the TAE of the original signal and the RIR, respectively. $\mathbf{c}_x(t)$ and $\mathbf{c}_h(t)$ are mutually independent WGN. Then, it is simple to derive that:

$$\langle \mathbf{c}(t)\mathbf{c}(t-\tau)\rangle = \delta(\tau) \tag{3.5}$$

where $\delta(\tau)$ is the Dirac delta function and $\langle\cdot\rangle$ denotes the ensemble average operation [49]. The ensemble average of $y^2(t)$ is determined as:

$$\langle y^2(t)\rangle = \left\langle \left[\int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau \right]^2 \right\rangle \tag{3.6}$$
$$= \int_{-\infty}^{\infty} e_x(\tau_1)e_h(t-\tau_1)d\tau_1 \int_{-\infty}^{\infty} e_x(\tau_2)e_h(t-\tau_2)d\tau_2 \tag{3.7}$$
$$\times \langle \mathbf{c}_x(\tau_1)\mathbf{c}_x(\tau_2)\rangle\langle \mathbf{c}_h(t-\tau_1)\mathbf{c}_h(t-\tau_2)\rangle \tag{3.8}$$
$$= \int_{-\infty}^{\infty} e_x^2(\tau)e_h^2(t-\tau)d\tau \tag{3.9}$$
$$= e_y^2(t) \tag{3.10}$$

Hence, the following equation is determined:

$$e_y^2(t) = e_x^2(t) * e_h^2(t) \tag{3.11}$$

where $e_y^2(t)$, $e_x^2(t)$ and $e_h^2(t)$ are the TPE of the reverberant signal, the original signal and the RIR, respectively.

Given the TPE of the input signal $e_x^2(t)$ according to superposition principles as:

$$e_x^2(t) = \sum_{k=0}^{K} C_k \cos(2\pi f_{m,k} t + \phi_k), \quad t \in [0, T] \tag{3.12}$$

where k is the index of K components, $f_{m,k}$ is modulation frequency at k, ϕ_k is phase, C_k is constant gain, and T is the time interval. Substituting Eq. (3.12) and Eq. (3.1) into Eq. (3.11), corresponding TPE of the reverberant signal is determined as:

$$e_y^2(t) = \sum_{k=0}^{K} \frac{C_k a^2 T_t}{\sqrt{13.8^2 + (2\pi f_{m,k} T_h)^2}} \Bigg[\cos(2\pi f_{m,k} t + \phi_k) - \exp\Big[-13.8\Big(\frac{t - T_h}{T_t}\Big)\Big] \cos(2\pi f_{m,k} T_h + \phi_k + \theta_k) \Bigg] \tag{3.13}$$

where $\theta_k = \tan^{-1}\Big(\frac{-2\pi f_{m,k} T_t}{13.8}\Big)$ and $t \in [T_h, t - T_t)$. Eq. (3.13) can model the TPE of any reverberant signal, which reveals how the RIR model parameters affect the waveform of the TPE of a reverberant signal. Figure 3.3 illustrates an example of TPE of a reverberant signal using Eq. (3.12), compared to the TPE of the same signal generated by convolution.

3.2.2 De-revereberation process

In this section, the de-reverberation process builds up the relationship between the parameters of the RIR model and observed reverberant signals. We spotlight the envelopes of TPE restored from the TPE of the reverberant signal modeled by Eq. (3.13) via an inverse filtering process.

First, the envelope of the restored TPE is constructed based on the concept of the inverse filtering process, with a set of $\tilde{T}_h$ and $\tilde{T}_t$ to obtain the corresponding $e_x^2(t)$. The upper and lower envelopes of $e_x^2(t)$ are determined as:

$$e_{x,\mathrm{upr}}^2(t) = \sum_{k=0}^{K} \frac{C_k T_t}{T_t'} \sqrt{\frac{1 + \Big(\frac{2\pi f_{m,k} T_t'}{13.8}\Big)}{1 + \Big(\frac{2\pi f_{m,k} T_t}{13.8}\Big)}} \frac{u(t) - \psi(t, T_h, T_t)}{u(t) - \psi(t, \tilde{T}_h, \tilde{T}_t)} \tag{3.14}$$

$$e_{x,\mathrm{lwr}}^2(t) = \sum_{k=0}^{K} \frac{C_k T_t}{T_t'} \sqrt{\frac{1 + \Big(\frac{2\pi f_{m,k} T_t'}{13.8}\Big)}{1 + \Big(\frac{2\pi f_{m,k} T_t}{13.8}\Big)}} \frac{-u(t) - \psi(t, T_h, T_t)}{u(t) + \psi(t, \tilde{T}_h, \tilde{T}_t)} \tag{3.15}$$

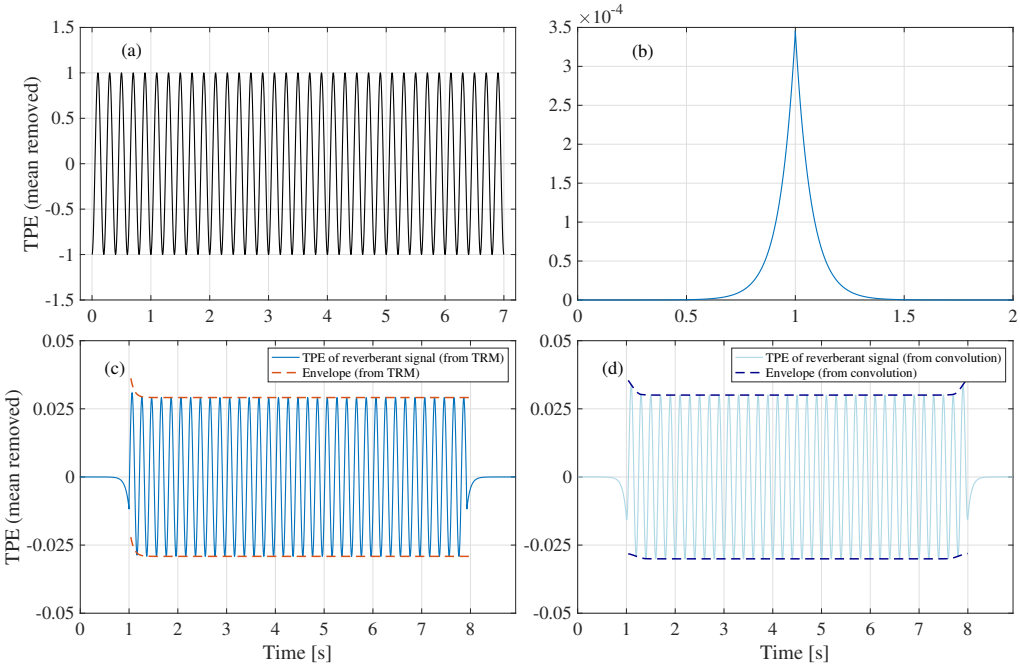

Figure 3.3: Example of TPE of a reverberation signal with discarding mean TPE (direct-current component): (a) TPE of original signal, (b) TPE of RIR, TPE of reverberant signal generated from (c) Eq. (3.13) and (d) convolution. Burnt-orange and deep-blue dashed line denote the envelopes of TPE, respectively.

where $\psi(t, T_h, T_t) = \exp\left[-13.8\ \frac{(t-T_h)}{T_t}\right]\cos(2\pi f_{m,k}T_h + \phi_k + \theta_k)$ and $e^2_{x,\mathrm{upr}}(t)$ and $e^2_{x,\mathrm{lwr}}(t)$ are upper and lower envelopes, respectively. u(t) is the unit-step function. Eqs. (3.14) and (3.15) reveal that when $T_h = \tilde{T_h}$ and $T_t = \tilde{T_t}$ hold, $e^2_{x,\mathrm{upr}}(t) = \sum_{k=0}^{K} C_k$ and $e^2_{y,\mathrm{lwr}}(t) = \sum_{k=0}^{K} -C_k$ hold. In this case, the envelopes are invariant to time, whereas when $T_h \neq \tilde{T_h}$ and $T_t \neq \tilde{T_t}$, the envelopes are time-varying. Thus, these time-varying envelopes can be approximated by first-order polynomials as:

$$e^2_{x,\mathrm{upr}}(t) = S_{\mathrm{upr}}t + b_{\mathrm{upr}} \tag{3.16}$$

$$e^2_{x,\mathrm{lwr}}(t) = S_{\mathrm{lwr}}t + b_{\mathrm{lwr}} \tag{3.17}$$

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

Until now, it has been created that the relationship between the parameters of the RIR model approximating an unknown RIR and the observed reverberant signal. It is found that the feature is related to the slope of the envelope of the restored TPE, which can be used to develop a blind estimation strategy.

3.3 Blind estimation strategy

This section proposes the blind estimation strategy based on the temporal power envelope model (TPEM) mentioned above. The block diagram of the proposed method is shown in Figure 3.3, namely, the alternating estimation strategy (AES) since the proposed method alternates to estimate the two parameters (T_h and T_t) of the extended RIR model.

First, the parameters of the extended RIR model are blindly estimated to synthesize the RIR by using the parameters. Then, using the synthesized RIRs blindly and simultaneously estimate the STI and RAPs according to the IEC 60268-16:2020 and ISO 3382:2009 standards [13, 14].

3.3.1 Blind estimation of parameters of RIR model

The proposed method estimates the parameters of the RIR model by utilizing the aforementioned slope-related feature. The TPE of an observed reverberant signal is extracted by:

$$e^2_y(t) = \mathbf{LPF}\left[|y(t) + j \cdot \mathrm{Hilbert}(y(t))|\right]^2 \tag{3.18}$$

where **LPF** is a Butterworth lowpass filter at a cut-off frequency 30 Hz. Hilbert denotes the Hilbert transform. The observed signal $y(t)$ is demodulated and smoothed by Hilbert transform and a lowpass filter. Eq. (3.1) is discretized and transformed into the z-domain to obtain the corresponding infinite-impulse-response filter as:

$$E_h(z) = \frac{a^2\ (\alpha - \beta)}{(1 - \alpha z^{-1})\ (1 - \beta z^{-1})} \tag{3.19}$$

where $\alpha = \exp(-13.8/T_t f_s)$, $\beta = \exp(13.8/T_h f_s)$ and f_s is sampling frequency.

According to the definition of the MTF in Eq. (3.20), the MTF of the extended RIR model is determined as:

$$m(f_m, T_h, T_t) = \frac{a^2}{\sqrt{\left[1 + \left(\frac{2\pi f_m T_h}{13.8}\right)^2\right]\left[1 + \left(\frac{2\pi f_m T_t}{13.8}\right)^2\right]}} \tag{3.20}$$

According to Eq. 3.19, the inverse filter is formulated as:

$$E_{h,\text{inv}} = E_h^{-1}(z) = \frac{(1 - \alpha z^{-1})(1 - \beta z^{-1})}{a^2(\alpha - \beta)} \tag{3.21}$$

The next section is about the frame-based whitening process, the key of the proposed method. The whitening transforms a complex waveform of the signal into a pulse train that consists of the even envelope to calculate the slopes, for the sake of utilizing the slope-related feature demonstrated in Section 3.2.2. Figure 3.4 illustrates how the whitening process works.

The restored TPE $e_x^2[n]$ is regarded to be autoregressive (AR) at each frame, of which frame length is n, and is rewritten as:

$$e_x^2[n] = -\sum_{i=1}^{p} \sigma_i e_x^2[n-i] + w_x^2[n] \tag{3.22}$$

$$w_x^2[n] = \sum_{i=0}^{p} \sigma_i e_x^2[n-i],\ W(z) = \sum_{i=0}^{p} \sigma_i z^{-i} \tag{3.23}$$

where σ_i is the optimal predictor, $\sigma_0 = 1$, p is the number of predictor order, $w_x^2[n]$ is the whitened restored TPE and $W(z)$ is the frame-based whitening filter [50, 51]. The optimal predictor σ_i can be obtained via *the normal equations* [50, 51, 52]. Defining $R_{e_x^2}[p]$ to be the autocorrelation sequences of $e_x^2[n]$, that is,

$$R_{e_x^2} = E\left[e_x^2[n] e_x^2[n-p]^\star\right] \tag{3.24}$$

where starry symbol "$\star$" denotes conjugate operation. The optimal predictor is given by

$$\mathbf{R}\boldsymbol{\sigma} = -\mathbf{r}, \qquad \boldsymbol{\sigma} = -\mathbf{R}^{-1}\mathbf{r} \tag{3.25}$$

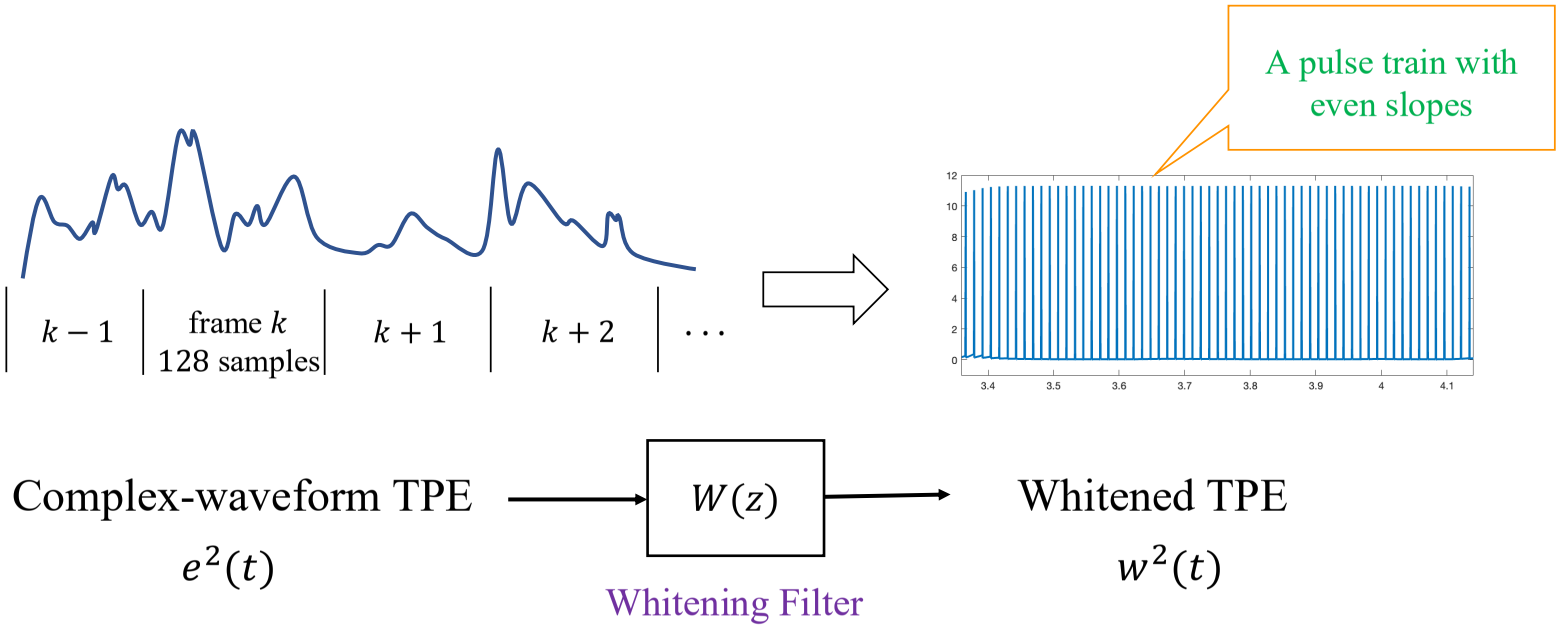

Figure 3.4: The block diagram of the whitening process

$$\mathbf{R} = \begin{bmatrix} E[|e_x^2[1]|^2] & E[e_x^2[1]e_x^2[2]^\star] & \cdots & E[e_x^2[1]e_x^2[p]^\star] \\ E[e_x^2[2]e_x^2[1]^\star] & E[|e_x^2[2]|^2] & \cdots & E[e_x^2[2]e_x^2[p]^\star] \\ \vdots & \vdots & \vdots & \vdots \\ E[e_x^2[p]e_x^2[1]^\star] & E[e_x^2[p]e_x^2[2]^\star] & \cdots & E[e_x^2[p]e_x^2[p]^\star] \end{bmatrix} \tag{3.26}$$

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_p \end{bmatrix} \tag{3.27}$$

where $\mathbf{R}$ is Toeplitz matrix of $R_{e_x^2}$, which is Hermitian, i.e., $\mathbf{R}^\dagger = \mathbf{R}$. "$\dagger$" denotes Hermitian conjugate operation. $\mathbf{r} = \begin{bmatrix} R_{e_x^2}[1] & \cdots & R_{e_x^2}[p] \end{bmatrix}^T$, is the cross-correlation vector, and $\boldsymbol{\sigma}$ is the column vector of σ_i. "T" denotes the transpose operation.

The optimal $\widehat{T}_t$ and $\widehat{T}_h$ are specifically obtained. All possible T_h and T_t sets are covered to seek out the optimal set that minimizes the slopes of envelopes of whitened TPE $w_x^2(t)$. Eq. (3.28) is derived to determine $\widehat{T}_t$, where "med" denotes the median operation. Then, substitute $\widehat{T}_t$ into Eq. (3.19) to perform inverse filtering so that the $\widehat{T}_h$ can be obtained by using Eq. (3.29).

$$\widehat{T}_t = \underset{T_t}{\text{med}}\Big\{\underset{T_h,\{\tilde{T}_t\}}{\text{argmin}}\Big[\log_{10}\big(|S_{\text{upr}}|\big) + \log_{10}\big(|S_{\text{lwr}}|\big)\Big]\Big\}. \tag{3.28}$$

$$\widehat{T}_h = \text{argmin}\Big\{\log_{10}\big(|S_{\text{upr}}|\big) + \log_{10}\big(|S_{\text{lwr}}|\big)\Big\}. \tag{3.29}$$

3.3.2 Blind estimation of RAPs

After proposing the estimation of the parameters of the extended RIR model. Then, the estimated parameters were used to synthesize the RIR according to Eq. (3.1) and Eq. (3.2) (assuming $t_0 = T_h$). Figure 3.5 shows an example of synthesized RIR by the proposed method and the measured RIR from the real environment, respectively.

The synthesized RIR is used to calculate the STI and RAPs simultaneously. The STI calculation has been standardized in IEC 60268-16:2020 [13]. The block diagram of the STI calculation is shown in Figure 2.8. The corresponding mathematical formulae are given in Eq. (2.14) - Eq. (2.17).

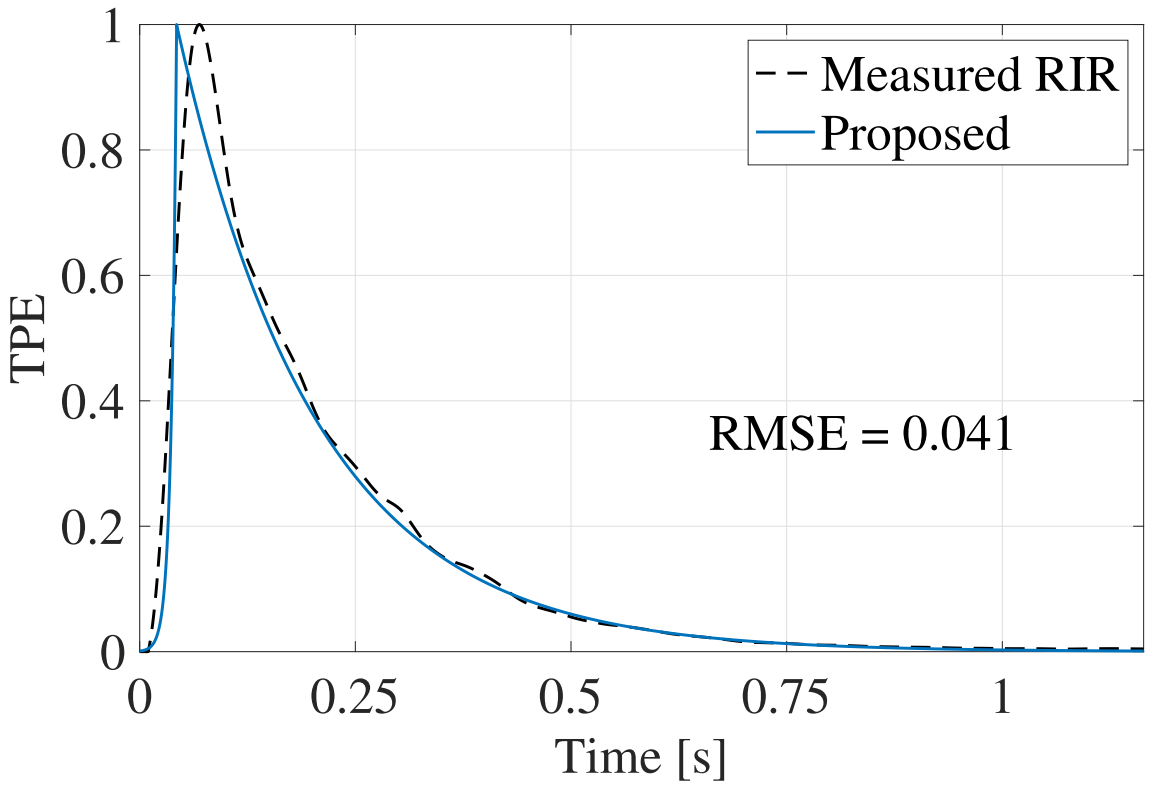

Figure 3.5: Example of realistic and reconstructed RIR by using the proposed method

The calculations of five RAPs, T_{60}, EDT, C_{80}, D_{50} and T_s have been standardized in ISO 3382:2009 [14]. The calculation of T_{60} and EDT are given in Section 2.3.2 and Section 2.3.4, respectively. C_{80} and D_{50} can be calculated by Eq. (2.19) and Eq. (2.20), respectively. T_s can be calculated by Eq. (2.21).

3.4 Optimization

The section discusses the technical details of the important signal processing techniques, including how to extract the envelope properly and how to calculate the slope.

3.4.1 Envelope extraction

The envelope extraction is of paramount importance for the proposed method, which is strongly relevant to slope calculation and further the estimation accuracy of the parameters of the RIR model and RAPs. The inverse filter and whitening filter introduce unwanted noise and outliers since designing the whitening filter is based on the linear prediction coding techniques [50]. These outliers affect severely the approximate extraction of the envelopes and further result in the inaccuracy of the calculated slopes.

Commonly, the local-extrema detection algorithm is applied to detect the local maximum and minimum of the signal to extract the upper and lower envelopes, respectively. However, this algorithm is fairly sensitive to noise and outliers, which is not the optimal choice. Hence, instead of the common-used local-extrema detection algorithm, the envelopes of a TPE are extracted using the specific peak-detection algorithm. This algorithm utilizes the alternating nature of the derivation to identify local extrema along with user-defined threshold [53]. This algorithm is robust and fast against introduced noise and outliers.

3.4.2 Slope calculation

Slope calculation is another key to the proposed method since the slope is the evaluation metric to find out the optimal T_h and T_t set. An accurate calculation of the slope is required.

As mentioned in Eq. (3.16) and Eq. (3.17), the slope is the first-order slope from the first polynomial approximation. It is intended to employ the $\mathcal{L}2$-norm least-square and $\mathcal{L}1$-norm least-absolute deviation (LAD) [54] with regularization to optimize the slope calculation, to be robust against noise.

With regard to the $\mathcal{L}2$-norm method, optimize the objective function as follows:

$$(b_{\text{env}}, S_{\text{env}}) = \operatorname{argmin} \sum_{m=1}^{N} \left| w_{\text{env}}^2[m] - \begin{bmatrix} t[m] \\ \lambda \end{bmatrix} \begin{bmatrix} S_{\text{env}} & b_{\text{env}} \end{bmatrix} \right|^2 \tag{3.30}$$

where b_{env} and S_{env} are the envelopes of whitened TPE w_{env}^2, i.e., b_{upr} and b_{upr} in Eq. (3.16), b_{lwr} and S_{lwr} in Eq. (3.17). λ is the regularization coefficient.

m is the m-th index of the total length N. The closed-form of b_{env} and S_{env} are given by vector form as:

$$\mathbf{P}_{\text{env}} = (\mathbf{H}_{\text{env}}^T \mathbf{H}_{\text{env}} + \lambda \mathbf{I})^{-1} \mathbf{H}_{\text{env}}^T \mathbf{w}_{\text{env}}^2 \tag{3.31}$$

$$\mathbf{P}_{\text{env}} = \begin{bmatrix} S_{\text{env}} \\ b_{\text{env}} \end{bmatrix} \tag{3.32}$$

$$\mathbf{H}_{\text{env}} = \begin{bmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_N & 1 \end{bmatrix} \tag{3.33}$$

where $\mathbf{w}_{\text{env}}^2$ is the vector form of $w_{\text{env}}^2[m]$.

With regard to $\mathcal{L}1$-norm method, the objective function is formulated as:

$$(b_{\text{env}}, S_{\text{env}}) = \text{argmin} \sum_{m=1}^{N} \left\| w_{\text{env}}^2[m] - \begin{bmatrix} t[m] \\ \lambda \end{bmatrix} \begin{bmatrix} S_{\text{env}} & b_{\text{env}} \end{bmatrix} \right\| \tag{3.34}$$

Instead of a closed-form solution, the numerical solution is given by the Iteratively ReWeighted Least Squares (IRWLS) algorithm [54].

We denote $\begin{bmatrix} S_{\text{env}} & b_{\text{env}} \end{bmatrix}$ as $\boldsymbol{\beta}$. Considering the initial start of $\boldsymbol{\beta}^{(0)}$, the IRWLS updates $\boldsymbol{\beta}$ at n iteration by

$$\boldsymbol{\beta}^{(n+1)} = (\mathbf{H}_{\text{env}}^T \mathbf{W}^{(n)} \mathbf{H}_{\text{env}})^{-1} \mathbf{H}_{\text{env}}^T \mathbf{W}^{(n)} \mathbf{w}_{\text{env}}^2 \tag{3.35}$$

$$\mathbf{r}^{(n)} = \mathbf{w}_{\text{env}}^2 - \mathbf{H}_{\text{env}} \boldsymbol{\beta}^{(n)} \tag{3.36}$$

$$\mathbf{W}^{(n)} = \text{diag}(\omega_1^{(n)}, \omega_2^{(n)}, \cdots, \omega_N^{(n)}) \tag{3.37}$$

where

$$\omega_m^{(n)} = \begin{cases} 1/|r_m^{(n)}|, & |r_m^{(n)}| > 10^{-6} \\ 1/10^{-6}, & |r_m^{(n)}| \leq 10^{-6} \end{cases} \tag{3.38}$$

and diag represents the diagonalization operation.

The IRWLS converges until

$$\frac{||\boldsymbol{\beta}^{(n+1)} - \boldsymbol{\beta}^{(n)}||_2}{||\boldsymbol{\beta}^{(n)}||_2} < \epsilon \tag{3.39}$$

where ϵ denotes the infinitesimal. In general, the $\mathcal{L}2$-norm method is chosen since it can handle almost situations to give a relatively accurate result of the slope without over-regularization. However, when encountering a noisy environment, the $\mathcal{L}1$-norm method can gain more robustness of calculation.

Chapter 4

Verifications and Evaluations

4.1 Verifications

The built-up TPEM model and the proposed method were verified by using the amplitude-modulation (AM) signal as the input signal to confirm the correctness of the TPEM and the proposed method.

4.1.1 AM signals

There are two types of AM signals used in verification. The first type is the single-tone AM signal used for verification of the TPEM model. The second type is the complex-tone AM signal used for verification of the proposed bind estimation strategy.

The signal-tone AM signal is the cosine AM signal at the modulation frequency 5 Hz because the modulation frequencies around 5 Hz are the most important cues for human perception of the speech signals [55, 56]. The complex-tone AM signal was synthesized by overlapping some harmonic tones from 0 to 20 Hz. The complex-tone AM signal was convolved with the RIR to obtain the corresponding reverberant AM signal. Since all the harmonic tones have identical amplitude, the synthesized complex-tone AM signal has even upper and lower envelopes, we omitted the whitening filter displayed in Figure 3.1. On the other hand, this implementation helps us to exclude the effect of whitening.

4.1.2 For proposed reverberation process of TPEM

The single-tone AM signal with the time elapsing 5 seconds was used as the original signal. Then, the corresponding reverberant signal was obtained in two ways. The one was by the convolution with the TPE of a synthesized

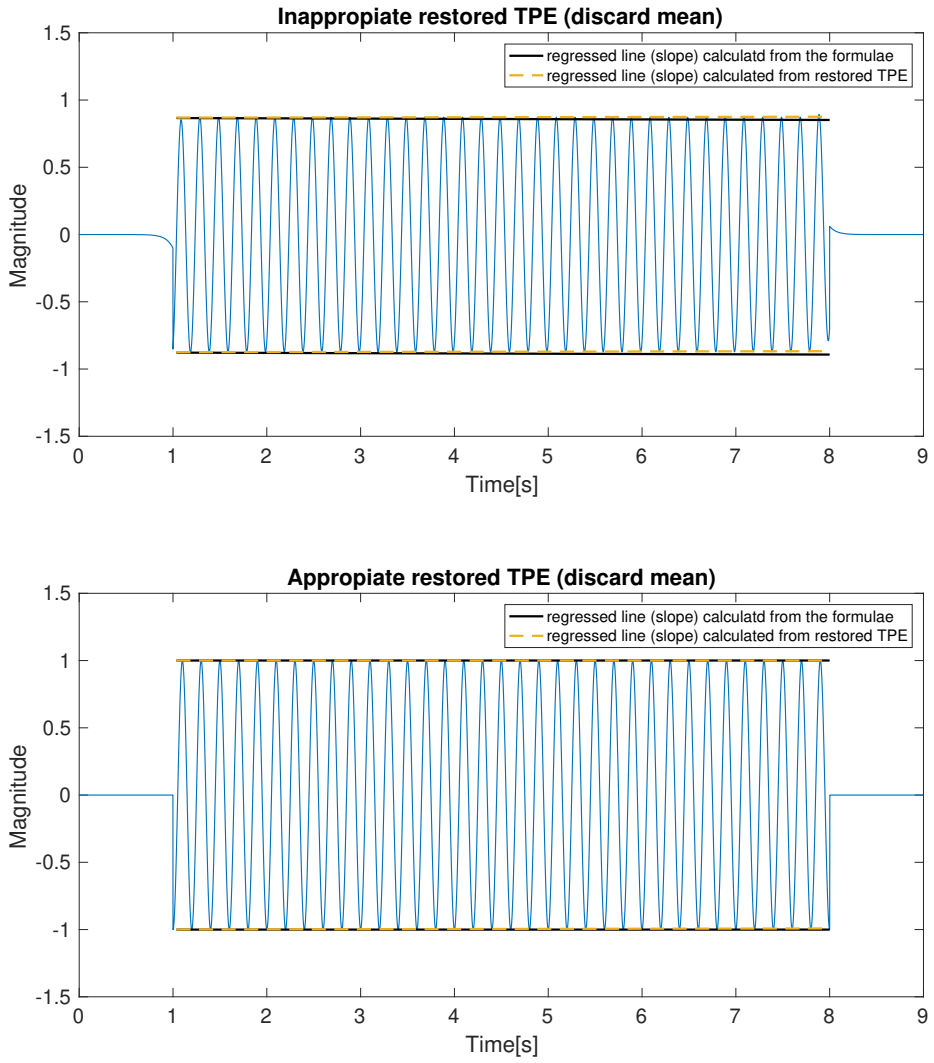

Figure 4.1: Example of envelopes of restored TPE with inappropriate restoration and appropriate restoration, respectively

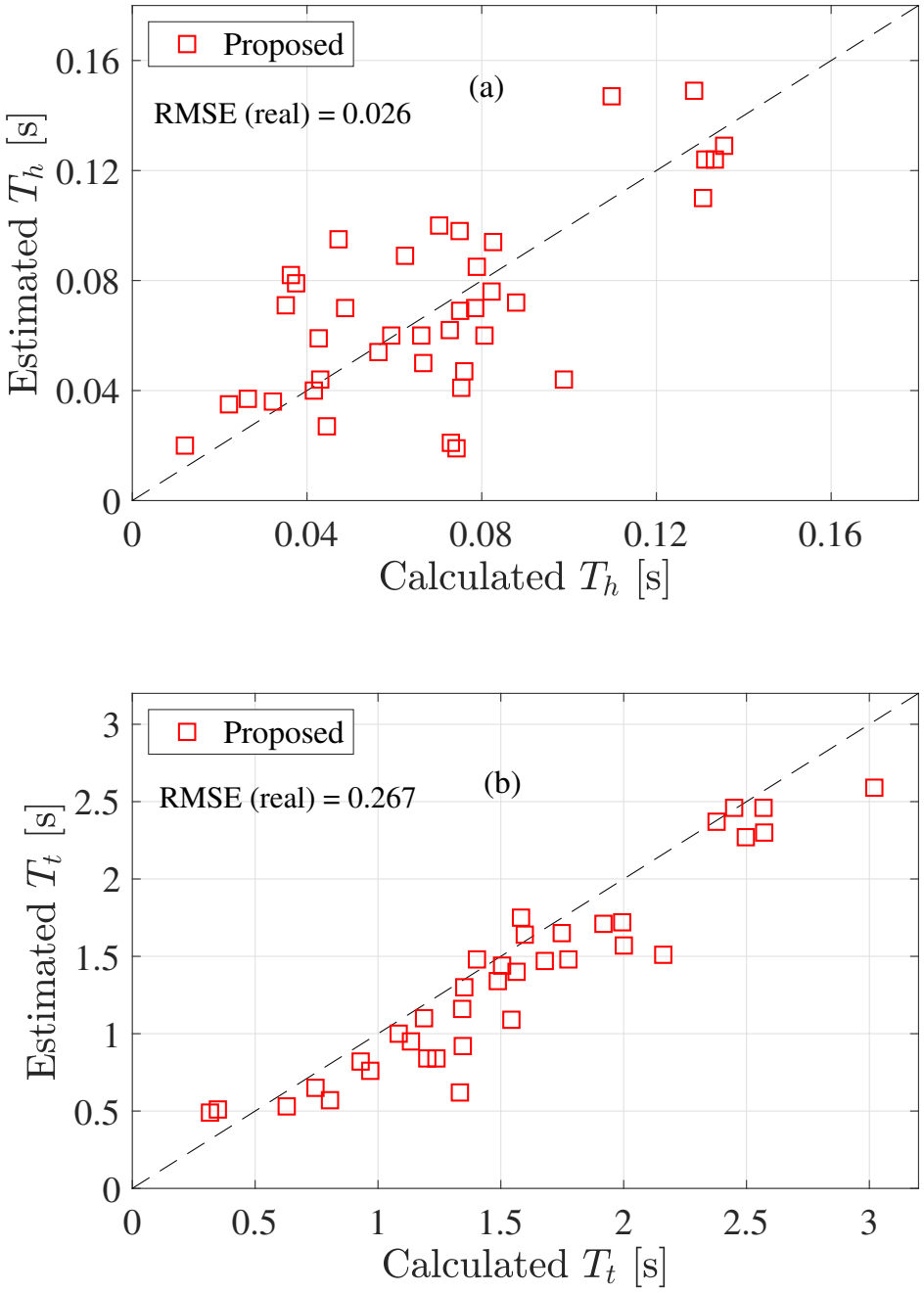

Figure 4.2: Estimated results of parameters of extended RIR model from observed reverberant speech signals: (a) T_h and (b) T_t. "□" denote the estimated value of the proposed method using AM signal.

artificial RIR by the extended RIR model in Eqs. (3.1) and (3.2) to obtain the reverberant signal. Another one was to synthesize the reverberation signal by using Eq. (3.13).

Figure 3.3 shows a verification example for the reverberation process in the TPEM. It is obvious that the TPE of the reverberant signal by using the TPEM is extremely similar to the TPE from the convolution, which indicates the built-up TPEM model can model the TPE of the reverberant signal appropriately. The Eqs. (3.13) prerequires the derivation of Eqs. (3.14) and (3.15), i.e., the derivation of the slope-related feature.

Then, the Eqs. (3.14) and (3.15) were verified by using the similar way. Figure 4.1 shows an example of the envelopes calculated using Eqs. (3.14) and (3.15) and calculated from the restored TPE using the inverse filter in Eq. (3.19), respectively. Hence, it is asserted that the TPEM model can model the reverberation and dereverberation process properly.

4.1.3 For proposed blind estimation strategy

This section verifies the proposed blind estimation strategy employing the complex-tone reverberant signal. Figure 4.2 shows the estimation results of parameters of the extended RIR model, T_h and T_t. The symbol "□" denotes estimated values of parameters. The horizontal and vertical axes indicate the parameters calculated from the RIR and parameters estimated from the AM signals, respectively. The estimation results indicate that the model the author built up is effective and the proposed method can be used for blind estimation.

4.2 Evaluations

4.2.1 Preliminaries

We evaluated the proposed method using reverberant speech signals to confirm whether or not the proposed method can estimate the parameters of the RIR model and the RAPs appropriately.

The speech signals were ten long Japanese sentences uttered by ten speakers (five males and five females) from ATR dataset [57]. We carried out simulations using reverberant speech signals synthesized by the convolution between speech signals and RIRs from the SMILE dataset, containing 43 measured RIRs [58]. The details of SMILE dataset are shown in Appendix A. We used root-mean-square error (RMSE) and Pearson correlation coefficient as the evaluation metrics. We implemented the comparative evaluations

Table 4.1: Estimation accuracy of parameters of extended RIR model by proposed and previous methods (RMSE) [46].

	T_h	T_t
TAE-CNN	0.087	0.193
Proposed	0.006	0.081

with the previous works [36, 46].

4.2.2 Evaluating parameters of RIR model

Figure 4.3 shows the estimated results of parameters of the extended RIR model from speech signals in realistic reverberant environments. The symbol "∘" represents the estimated parameters by the proposed method. The horizontal and vertical axes indicate the parameters calculated from the RIRs and estimated from the speech signals. The results show that the proposed method can effectively estimate the parameters of the extended RIR model. With regard to T_h, the RMSE was 0.006 and for T_t, the estimated results closely approach the ground-truths.

Table 4.1 shows the estimation accuracy of T_h and T_t using the proposed and previous method [46]. The results show that the proposed method can appropriately estimate the parameters of the extended RIR model.

4.2.3 Evaluating room-acoustic parameters

Figure 4.4 - 4.9 show the results of estimating STI and five-room acoustic parameters from speech signals in realistic reverberant environments. The symbols " ▫ ", " ∘ " and " ⋆ " represent the parameters estimated by the proposed and previous methods [36, 46], respectively. The horizontal axis indicates the parameters calculated from the RIRs, and the vertical axis indicates the parameters estimated from the speech signals.

Table 4.2 shows the estimation accuracy of the proposed and previous methods. The RMSEs of estimated parameters reveal that the proposed method outperforms (STI and T_{60}) or maintains the same level as the previous methods (EDT, C_{80}, D_{50} and T_s). Table 4.3 shows the correlation coefficient between the estimated and calculated results of the proposed and previous methods. However, the noticeable outliers exist for C_{80}, D_{50}, and T_s, presumably resulting from the mismatching of the carrier signal as the WGN to the realistic RIRs.

Table 4.2: Comparison between previous methods and proposed method in terms accuracy (RMSE) [36, 46].

	STI	T_{60}	EDT	C_{80}	D_{50}	T_s
MTF-based	0.060	–	–	–	–	–
TAE-CNN	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

Table 4.3: Pearson correlation coefficients between the estimated values and ground-truths.

	STI	T_{60}	EDT	C_{80}	D_{50}	T_s
TAE-CNN	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

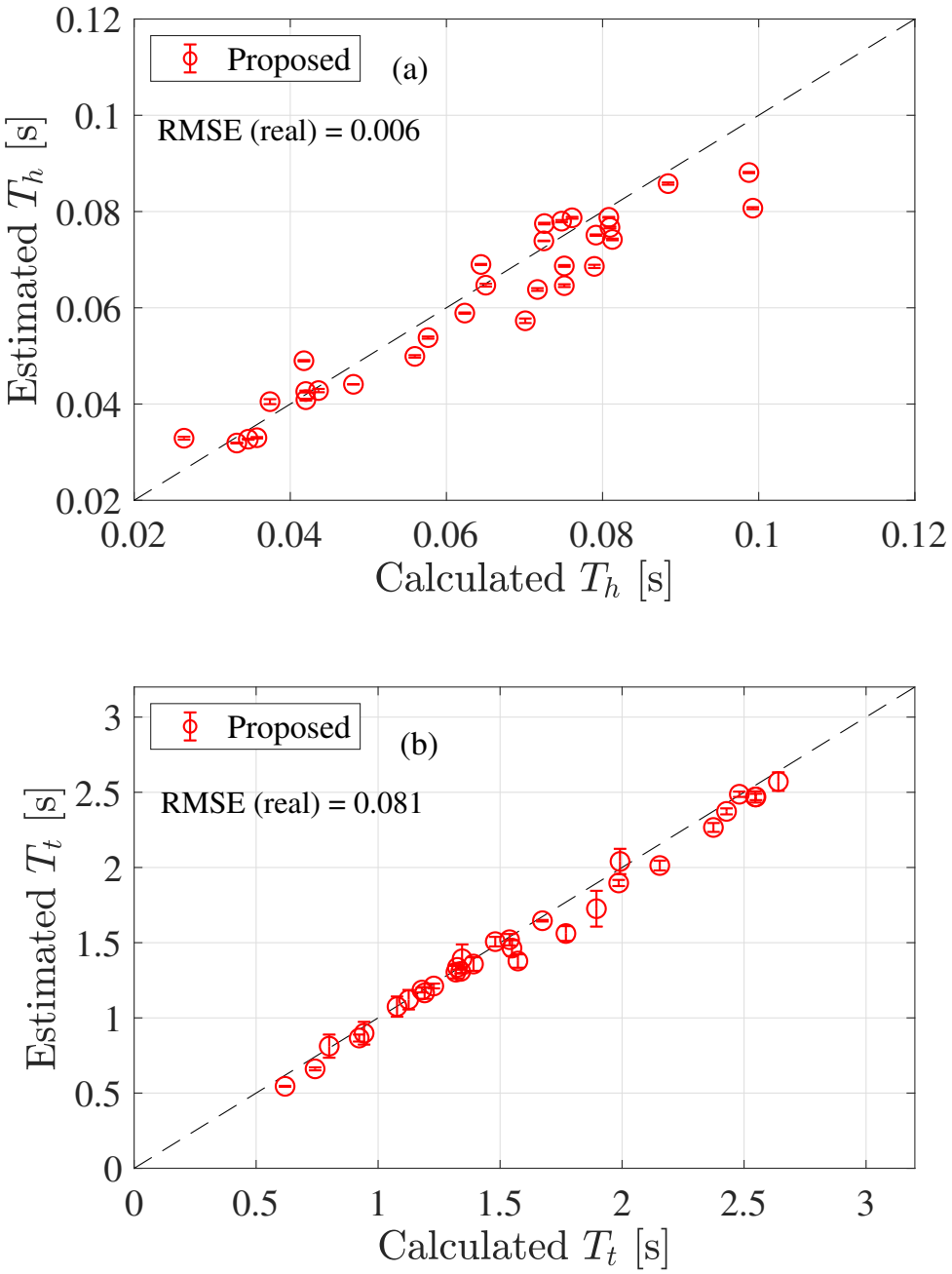

Figure 4.3: Estimated parameters for the extended RIR model using speech signals: (a) T_h and (b) T_t.

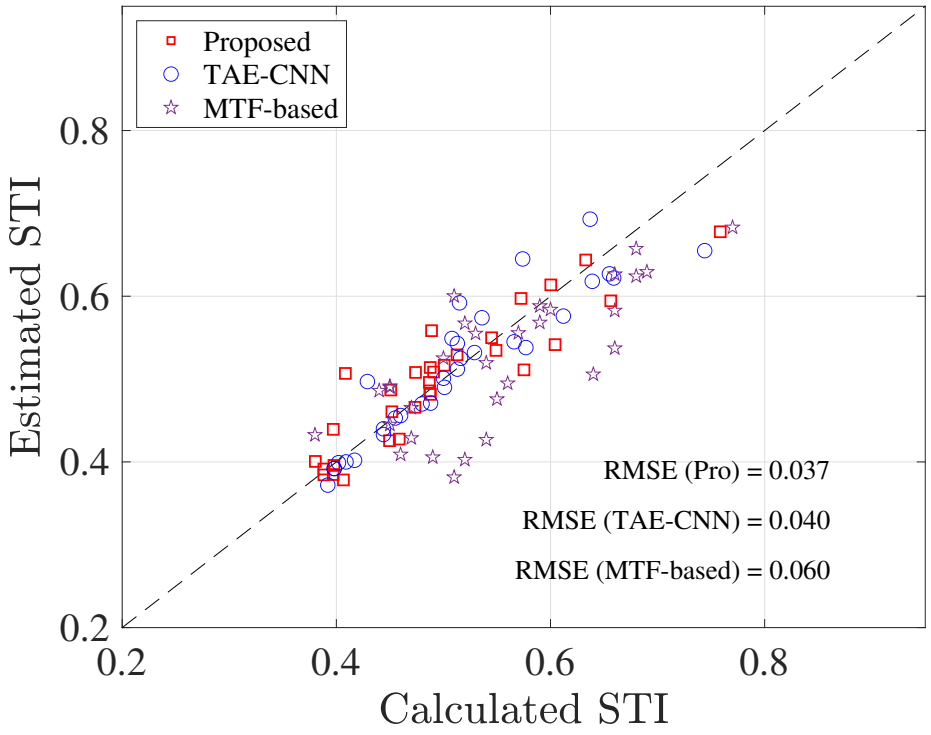

Figure 4.4: Results of estimating STI from reverberant speech signals. "▫", "∘", and "⋆" denote the estimated value of the proposed and two previous works, respectively, including the TAE-based CNN method (TAE-CNN) [46] and MTF-based method (MTF-based) [36]. The black dashed line represents the ground-truths calculated from the RIRs.

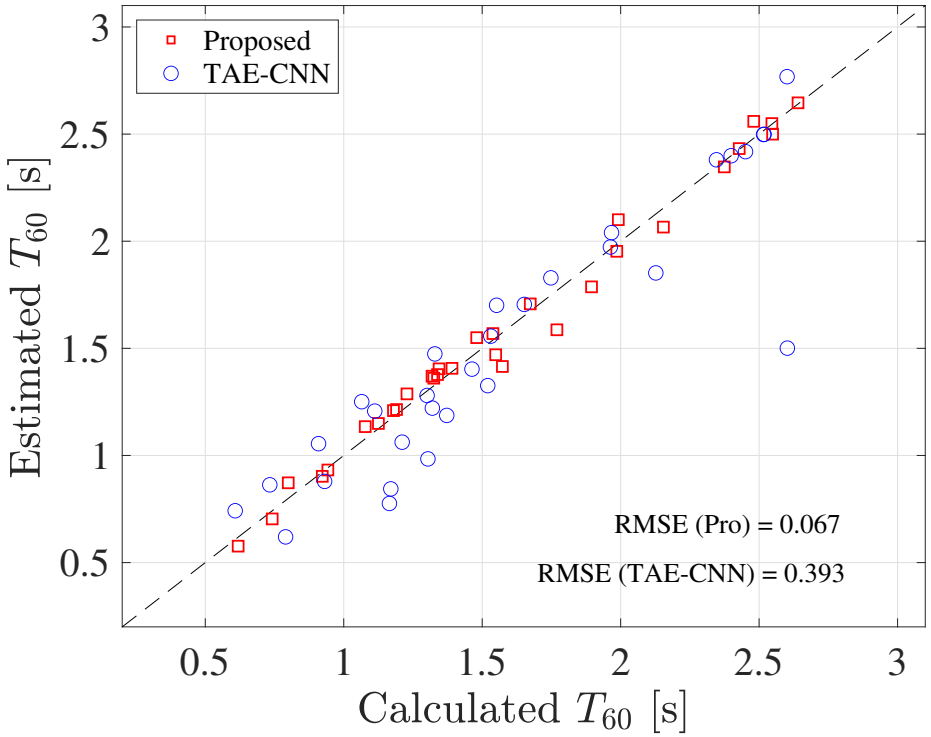

Figure 4.5: Results of estimating T_{60} from reverberant speech signals. "▫" and "∘" denote the estimated value of the proposed and two previous works, respectively, including the TAE-based CNN method (TAE-CNN) [46]. The black dashed line represents the ground-truths calculated from the RIRs.

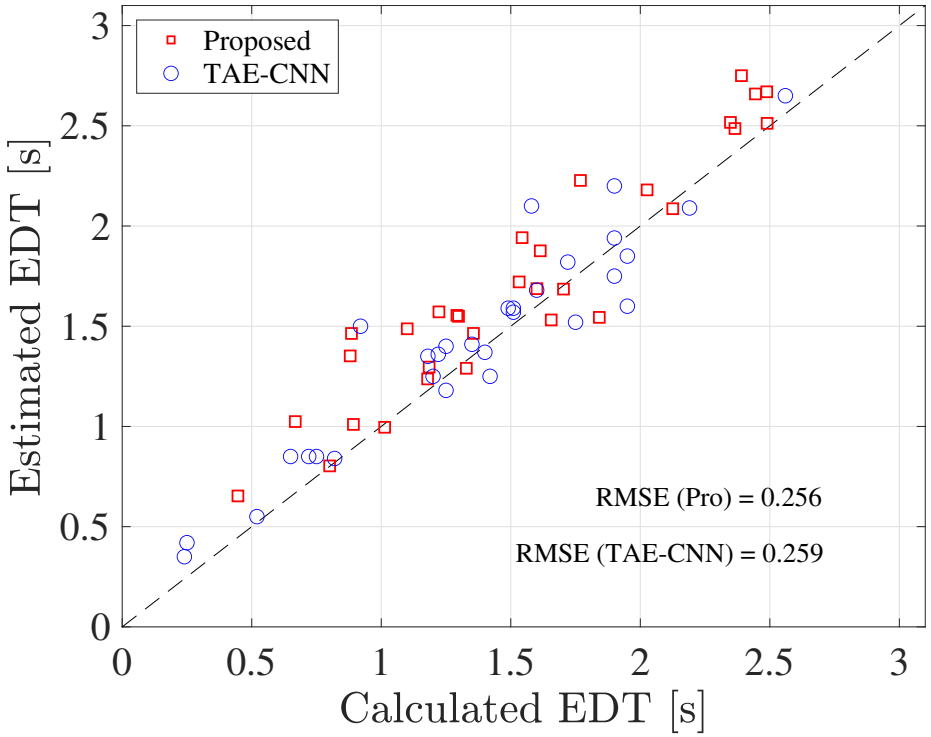

Figure 4.6: Results of estimating EDT from reverberant speech signals. "▫" and "∘" denote the estimated value of the proposed and two previous works, respectively, including the TAE-based CNN method (TAE-CNN) [46]. The black dashed line represents the ground-truths calculated from the RIRs.

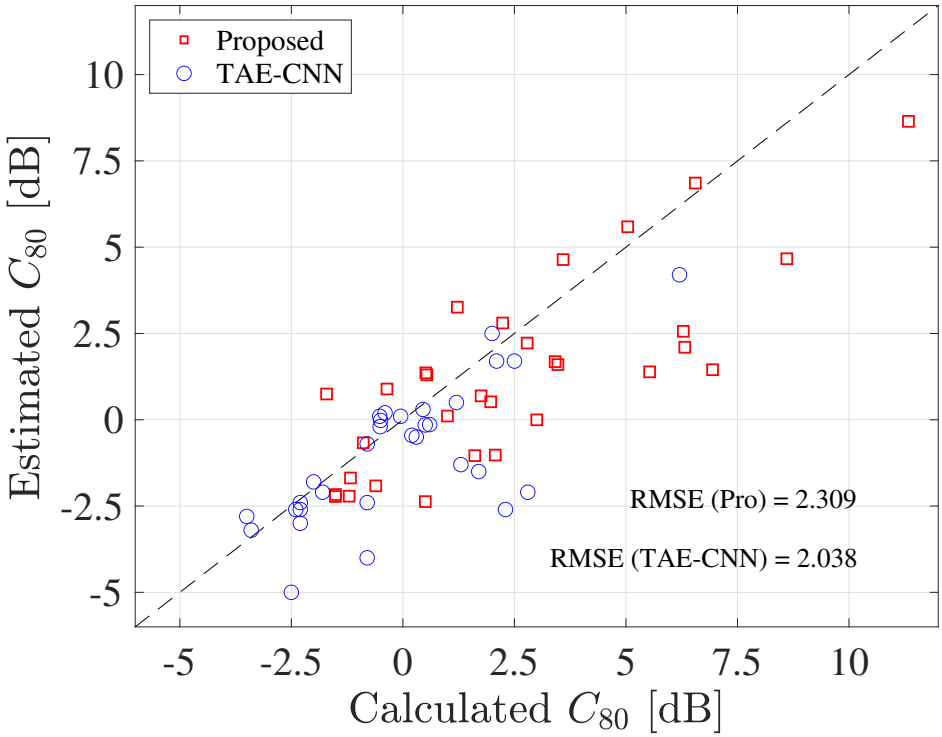

Figure 4.7: Results of estimating C_{80} from reverberant speech signals. "▫" and "∘" denote the estimated value of the proposed and two previous works, respectively, including the TAE-based CNN method (TAE-CNN) [46]. The black dashed line represents the ground-truths calculated from the RIRs.

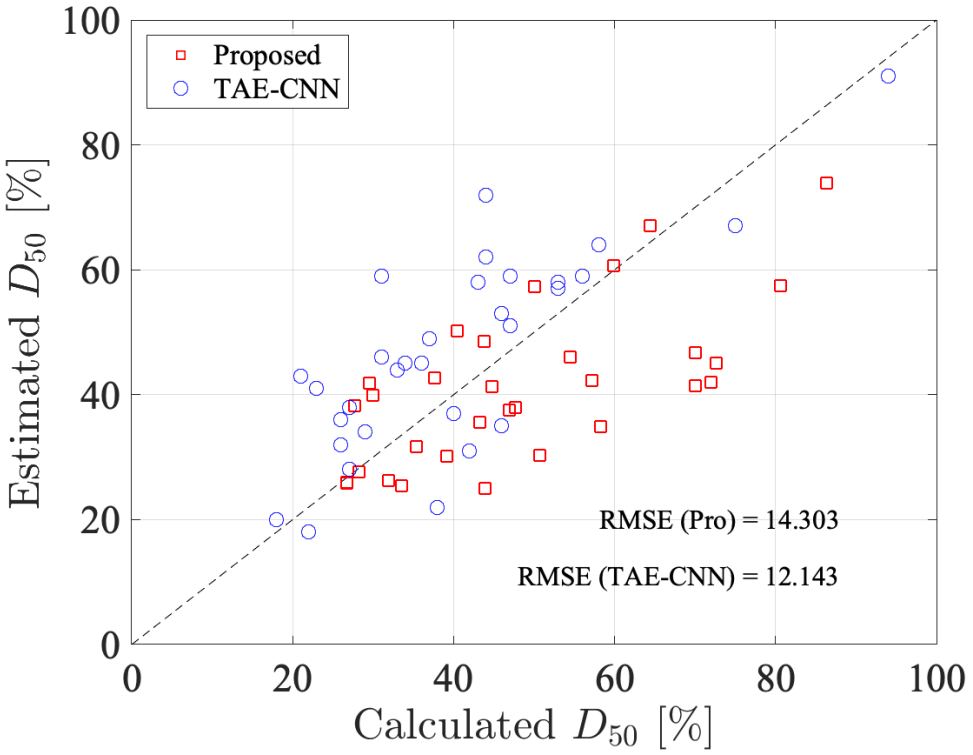

Figure 4.8: Results of estimating D_{50} from reverberant speech signals. "▫" and "∘" denote the estimated value of the proposed and two previous works, respectively, including the TAE-based CNN method (TAE-CNN) [46]. The black dashed line represents the ground-truths calculated from the RIRs.

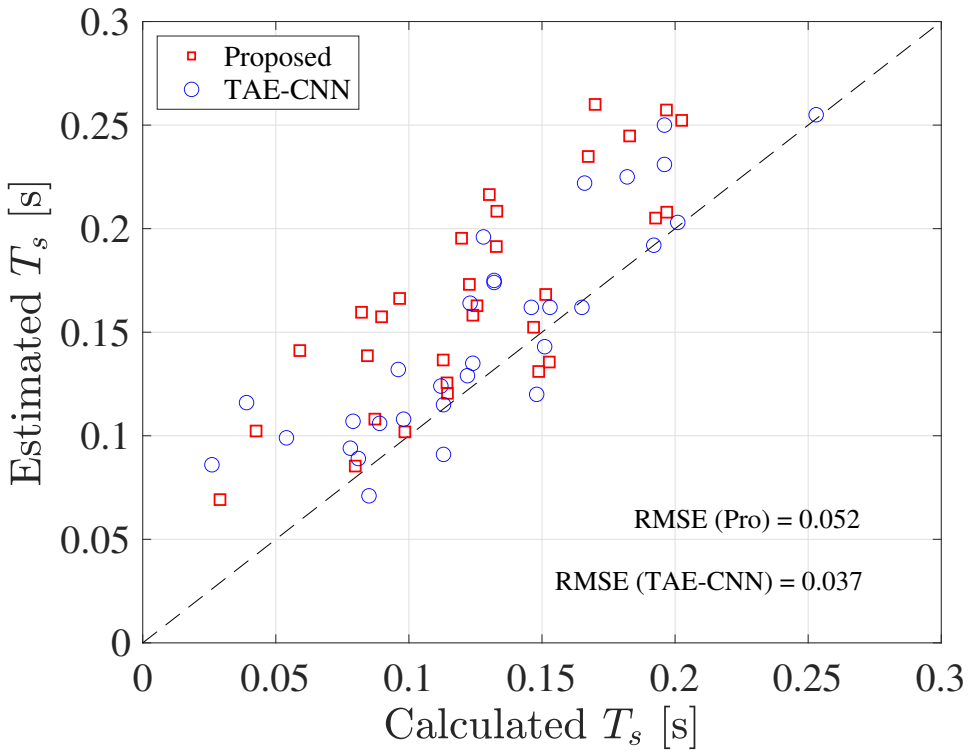

Figure 4.9: Results of estimating T_s from reverberant speech signals. "▫" and "∘" denote the estimated value of the proposed and two previous works, respectively, including the TAE-based CNN method (TAE-CNN) [46]. The black dashed line represents the ground-truths calculated from the RIRs.

Chapter 5

Conclusions and outlooks

5.1 Summary

A deterministic method was proposed for blindly estimating the parameters of the extended RIR model, as well as the STI and five RAPs, i.e., T_{60}, EDT, C_{80}, D_{50}, and T_s, simultaneously.

Using the extended RIR model to approximate an RIR mitigates the limitation of the common-used Schroeder's model and makes the RIR model physically interpretable. Furthermore, instead of relying on training data, the relationship between envelopes of the TPE of an observed reverberant signal and the RIR was created as a function of the parameters of the RIR model. This relationship was used to estimate the optimal parameters of the RIR model. Then, the STI and five RAPs were blindly estimated by using the estimated RIR synthesized using these optimal parameters.

The evaluation results concluded that the proposed method could blindly the parameters of the extended RIR model effectively. The evaluation results also indicated that the proposed method could blindly and simultaneously estimate the STI and RAPs effectively.

5.2 Contributions

This work contributes to room acoustics and audio signal processing research. The contributions can be listed as follows:

- Clarifying how the RIR affects the waveform of the signals, via modeling the RIR and reverberation/dereverberation process in the time domain based on the concept of the MTF.

- Proposing a deterministic blind estimation strategy to effectively estimate the RIR by blindly estimating the parameters of the extended RIR model.
- Blindly and simultaneously estimating the RAPs that can be used to comprehensively describe the room acoustic characteristics of a sound field.
- Providing an alternative idea to understand the room acoustic characteristics of a sound field without measuring the RIR of a sound field.

5.3 Remaining works

The following issues are recommended to be contemplated in the future:

- Possibility of real-time or quasi-real-time implementation of the proposed blind estimation strategy. The current work is still limited to depend upon the long-time speech signal, taking an overwhelming amount of time to process.
- Although the researchers have an overall picture of the room acoustics of a sound field, it is still not clear what cues that appear in the RIR or MTF are related to speech intelligibility and sound clarity.
- The effects of the variations of the distance between the estimation and sound source position should be considered since the different distances result in different sound pressure levels, which may affect the estimation accuracy.
- The proposed method needs to be evaluated in real environments where people are included because the positions and numbers of the people cause the changes in room acoustics.
- Application of the proposed method for speech enhancement and hearing aids.
- Blind estimation from the music signals attracts some interest since all the current works cannot achieve the accurate estimation of the RIR and RAPs.

Appendix A

Room Impulse Response Dataset - SMILE

Table A.1: SMILE dataset of room impulse response used in evaluation

No.	ID	Measured Condition	T_{60}
1	301	Multi-purpose hall 1 (with reflex board)	1.09
2	302	Multi-purpose hall 1 (without reflex board)	0.80
3	303	Multi-purpose hall 2 (with reflex board)	1.44
4	304	Multi-purpose hall 2 (without reflex board)	1.04
5	305	Multi-purpose hall 3 (with reflex board)	1.93
6	306	Multi-purpose hall 3 (without reflex board)	1.35
7	307	Multi-purpose hall 3 (with absorption board)	1.42
8	308	Multi-purpose hall 4 (with absorption board)	1.54
9	319	Multi-purpose hall 5 (14,000 m^3)	1.47
10	321	Multi-purpose hall 6 (19,000 m^3)	2.16
11	309	Classis concert hall 1 (5600 m^3)	2.35
12	310	Classic concert hall 1 (d = 6 m^3)	2.34
13	311	Classic concert hall 1 (d = 11 m^3)	2.35
14	312	Classic concert hall 1 (d = 15 m^3)	2.39
15	313	Classic concert hall 1 (d = 19 m^3)	2.38
16	314	Classic concert hall 2 (6,100 m^3)	1.14

Continued on next page

Table A.1: SMILE dataset of room impulse response used in evaluation (Continued)

17	315	Classic concert hall 3 (20,000 m^3)	1.96
18	316	Classic concert hall 4 (with absorption curtain)	1.92
19	317	Classic concert hall 4 (without absorption curtain)	2.55
20	323	Classic concert hall 5 (17,000 m^3)	2.32
21	324	Classic concert hall 6 (1F front)	1.77
22	325	Classic concert hall 6 (2F side)	1.74
23	326	Classic concert hall 6 (3F)	1.69
24	201	Lecture room with flatter echoes	1.36
25	318	Theatre hall (3,900 m^3)	0.85
26	401	Meeting room (130 m^3)	0.62
27	402	Lecture room (400 m^3)	1.12
28	403	Lecture room (2,400 m^3)	1.09
29	404	General speech hall (11,000 m^3)	1.54
30	405	Church 1 (1,200 m^3)	0.71
31	406	Church 2 (3,200 m^3)	1.30
32	407	Event hall 1 (28,000 m^3)	3.03
33	408	Event hall 2 (41,000 m^3)	3.62
34	409	Gym 1 (12,000 m^3)	2.82
35	410	Gym 2 (29,000 m^3)	1.70
36	411	Living room (110 m^3)	0.36
37	412	Movie theatre (560 m^3)	0.38
38	413	Atrium (4,000 m^3)	1.57
39	414	Tunnel (5,900 m^3)	2.72
40	415	Concourse in train station	1.95
41	416	General speech hall 2 (1F front)	1.52
42	417	General speech hall 2 (1F center)	1.57
43	418	General speech hall 2 (1F balcony)	1.40

Bibliography

[1] H. Kuttruff, *Room Acoustics (6th ed.)*. Taylor & Francis, 2016.

[2] C. J. Plack, *The Sense of Hearing*. Routledge, 2018.

[3] B. Moore, *An Introduction to the Psychology of Hearing: Sixth Edition*. Leiden, The Netherlands: Brill, 2013.

[4] V. Gómez Escobar and J. Barrigón Morillas, "Analysis of intelligibility and reverberation time recommendations in educational rooms," *Applied Acoustics*, vol. 96, pp. 1–10, 2015.

[5] Y.-J. Choi, "The intelligibility of speech in university classrooms during lectures," *Applied Acoustics*, vol. 162, p. 107211, 2020.

[6] M. Barron, *Auditorium Acoustics and Architectural Design (2nd ed.)*. London: Routledge, 2009.

[7] F. Everest, *Master Handbook of Acoustics*. Master Handbook of Acoustics, Mcgraw-hill, 2000.

[8] S. K. S. K. Mitra, *Digital signal processing a computer-based approach / Sanjit K. Mitra.* New York, NY: McGraw-Hill, 4th ed. ed., 2011.

[9] T. Houtgast, "Predicting speech intelligibility in rooms from the modulation transfer function, i. general room acoustics," *Acustica*, vol. 46, pp. 60–72, 1980.

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

[11] F. Toole, *Sound Reproduction: Loudspeakers and Rooms*. Audio Engineering Society Presents Series, Elsevier, 2008.

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

[13] I. 60268-16:2020, *Sound system equipment - part 16: Objective rating of speech intelligibility by speech transmission index.*

[14] I. 3382:2009, *Acoustics - measurements of room acoustics parameters - part 1: Performance spaces.*

[15] J. S. Bradley, "Using iso 3382 measures, and their extensions, to evaluate acoustical conditions in concert halls," *Acoustical Science and Technology*, vol. 26, no. 2, pp. 170–178, 2005.

[16] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind system identification," *Proceedings of the IEEE*, vol. 85, no. 8, pp. 1310–1322, 1997.

[17] S. Chaudhuri, R. Velmurugan, and R. Rameshan, *Blind Image Deconvolution: Methods and Convergence*. Springer Publishing Company, Incorporated, 1st ed., 2016.

[18] P. S. R. Diniz, *Adaptive Filtering: Algorithms and Practical Implementation*. Cham: Springer International Publishing, 2020.

[19] V. Ingle, S. Kogon, and D. Manolakis, *Statisical and Adaptive Signal Processing*. 2005.

[20] M. R. Schroeder, "Integrated-impulse method measuring sound decay without using impulses," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 497–500, 1979.

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

[22] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 54, no. 2, pp. 557–557, 1973.

[23] M. R. Schroeder, "Frequency-correlation functions of frequency responses in rooms," *The Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1819–1823, 1962.

[24] M. Unoki and S. Hiramatsu, "MTF-based method of blind estimation of reverberation time in room acoustics," in *2008 16th European Signal Processing Conference*, pp. 1–5, 2008.

[25] J.-D. Polack, *Are Impulse Responses Gaussian Noises?*, pp. 77–91. Cham: Springer International Publishing, 2015.

[26] J. Li and Y. Liu, "Modulation transfer function measurements using a learning approach from multiple diffractive grids for optical cameras," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–8, 2021.

[27] A. D. A. Maidment and M. Albert, "Conditioning data for calculation of the modulation transfer function," *Medical Physics*, vol. 30, no. 2, pp. 248–253, 2003.

[28] G. Leembruggen, "Is sii better than sti at recognising the effects of poor tonal balance on intelligibility?," vol. 28, pp. 24–34, 01 2006.

[29] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[30] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz, "Better speech recognition with cochlear implants," *Nature*, vol. 352, no. 6332, pp. 236–238, 1991.

[31] P. J. Boyle, T. B. Nunn, A. F. O'connor, and B. C. J. Moore, "Starr: A speech test for evaluation of the effectiveness of auditory prostheses under realistic conditions," *Ear and Hearing*, vol. 34, p. 203–212, 2013.

[32] B. C. J. Moore, "The roles of temporal envelope and fine structure information in auditory perception," *Acoustical Science and Technology*, vol. 40, no. 2, pp. 61–83, 2019.

[33] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 54, no. 2, pp. 557–557, 1973.

[34] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, 1980.

[35] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[36] M. Unoki, A. Miyazaki, S. Morita, and M. Akagi, "Method of blindly estimating speech transmission index in noisy reverberant environments," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 8, pp. 1430–1445, 11 2017.

[37] 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 Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1884–1893, 2012.

[38] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[39] 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, 09 2007.

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

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

[42] P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang, and J. A. Chambers, "Monaural room acoustic parameters from music and speech," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 278–287, 2008.

[43] J. F. Santos and T. H. Falk, "Blind room acoustics characterization using recurrent neural networks and modulation spectrum dynamics," *Journal of the audio engineering society*, January 2016.

[44] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 591–595, 2018.

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

[46] 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," *Applied Acoustics*, vol. 185, p. 108372, 2022.

[47] C. J. Steinmetz, V. K. Ithapu, and P. Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 221–225, 2021.

[48] M. Unoki, D. Ishikawa, Y. Kashihara, M. Kobayashi, and M. Akagi, "Study on modeling of room impulse response and its room acoustic parameters," *IEICE Technical Report; IEICE Tech. Rep.*, vol. 116, no. 302, pp. 79–84, 2016.

[49] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes with Errata Sheet*. International student edition, McGraw-Hill, 2002.

[50] P. P. Vaidyanathan, *The Theory of Linear Prediction*. Synthesis Lectures on Engineering Series, Morgan & Claypool, 2008.

[51] T. Kailath, *Linear Systems*. Information and System Sciences Series, Prentice-Hall, 1980.

[52] A. H. Sayed, *Fundamentals of Adaptive Filtering*. IEEE Press, Wiley, 2003.

[53] N. Yoder, "peakfinder(x0, sel, thresh, extrema, includeendpoints, interpolate)." MATLAB Central File Exchange, 2022. [Online; accessed Aug 05, 2022].

[54] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust Statistics for Signal Processing*. Cambridge University Press, 2018.

[55] S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang, "Temporal properties of spontaneous speech—a syllable-centric perspective," *Journal of Phonetics*, vol. 31, no. 3, pp. 465–485, 2003. Temporal Integration in the Perception of Speech.

[56] 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," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 233–244, 2020.

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

[58] A. I. of Japan, *Sound library of architecture and environment*. Gihodo Shuppan Co., Ltd., Tokyo, 2004.

Publications

1. L. Wang and M. Unoki, "Study on method for blindly estimating parameters of extended model of room impulse," IEICE Technical Report, vol. 122, no. 266, EA2022-43, pp. 7-12, 2022.

2. L. Wang, S. Duangpummet and M. Unoki, "Blind estimation of room acoustic parameter from speech signals based on extended model of room impulse response," arXiv:2212.13009, http://arxiv.org/abs/2212.13009.

3. L. Wang, S. Duangpummet and M. Unoki, "Study on Blind estimation of Room Acoustic Parameters from Speech Signals Based on Extended Model of Room Impulse Response," in The 2023 Spring Meeting of the Acoustical Society of Japan, 2023.