Doctoral Dissertation

Blind Estimation of Room Acoustic Parameters and Speech Transmission Index Based on the Concept of the Modulation Transfer Function

Suradej DUANGPUMMET

Supervisor: Professor Masashi UNOKI

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

December, 2021

Abstract

Assessment of the quality of auditory spaces is essential in room acoustics and speech signal processing. In room acoustics, the sound characteristics of auditoriums are related to the quality of life in different aspects. In emergency circumstances, e.g., earthquake or flood, emergency announcements and alarm sounds need to be easily audible and intelligible so that we can follow the safety procedures appropriately. In theaters or concert halls, excellent sound characteristics and superior acoustics are ideal environments for performances. The sounds of live performances should be clear and transparent so that attendees can enjoy the entertainment. Additionally, speech signal processing, such as speech dereverberation and noise suppression, would benefit in which the sound quality and speech intelligibility can be improved based on the room characteristics.

Intelligibility of speech and pleasure to music are subjective descriptions. It is difficult to convey such descriptions from listeners to architects who are responsible for designing auditoriums or diagnosing acoustic problems. Conventionally, speech intelligibility and sound quality can be determined by conducting listening experiments with a group of listeners. Unfortunately, the experiments are expensive, unreliable, and time-consuming. It is also impractical for real-time applications, such as hearing aids, automatic speech recognition, and speaker verification. Thus, the quality of a sound field and subjective aspects are defined through room acoustic parameters and objective indices related to the physical properties of a sound field. Hence, architects, acousticians, and signal processing algorithms, can justify acoustic conditions by measuring acoustical parameters.

Several useful room acoustic parameters and objective indices have been standardized. 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 speech transmission channel. The STI is calculated based on the concept of the modulation transfer function (MTF). The MTFs of seven-octave bands with their weighting values are converted to be a real number from 0 to 1. In addition, ISO 3382: 2009, specifies methods for measurement the reverberation times (T_{60} or T_{30}) and other room-acoustic parameters, including 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). These parameters are derived from measuring the room impulse response (RIR).

In the time domain, an RIR completely describes the characteristics of a sound field. Similarly, a system transfer function in the frequency domain and the MTF in the modulation-frequency domain are the counterparts. In general, the RIR or MTF needs to be measured. However, it is difficult to measure RIR or MTF in daily-life places where people cannot be excluded, e.g., public stations, airports, and department stores. Moreover, by the nature of such public areas, room acoustics are prone to be a time-varying system. Sound absorption, reverberation, or other acoustical parameters are changed by varying occupants and object arrangements. Thus, acoustic parameters that were measured complying with the standards might be different from the current one. Hence, many methods have been proposed to estimate an acoustical parameter without measuring the RIR, known as *blind* estimation methods.

The blind estimation of an acoustical parameter is an ill-posed condition because both sound source and RIR are unknown. The ill-posed or blind inverse problem is challenging since it needs additional assumptions or complementary prior knowledge to formulate the estimation. Furthermore, the robustness of the estimator against various rooms (e.g.,

diffuse/non-diffuse field and connected chamber) and background noise need to be taken into account. To this end, this research presents blind estimation methods for estimating five-room acoustic parameters, STI, and SNR from a speech signal in noisy reverberant environments using a single-channel microphone and the concept of the MTF.

A speech signal can be decomposed into a fine structure and temporal structure. For temporal structure, a power envelope (PE) or temporal amplitude envelope (TAE) is used as a feature. On the basis of the MTF, PE or TAE represents the modulation distortion caused by reverberation and noise of the transmission channel (sound field). The TAE also plays an important role in speech intelligibility. In the proposed scheme, these features are extracted from an observed signal by using Hilbert transform and a low-pass filter. An observed signal in a given room is regarded as the output of the convolution between the RIR and speech signal. Hence, the modulation features (TAE/PE) and the convolution operation using one-dimensional convolutional neural networks (CNNs) were deployed. A more sophisticated deep neural network (DNN), such as a combined network between CNNs and long short-term memory (LSTM) networks, was also utilized. These DNNs were trained from the pairs of TAE/PE and the parameters of RIR models. In addition, data augmentation techniques were used for synthesising the dataset due to limited measured RIRs.

Here, an unknown RIR is modeled by using a stochastic RIR model. Two RIR models were investigated: Schroeder's RIR model and the extended RIR model. The reverberation time is an only parameter in Schroeder's RIR as a simple exponential decay (T_R). The extended RIR model is an extended version of Schroeder's RIR model. It consists of three parameters, including rising parameter (T_h), peak position (T_0), and exponential decay parameter (T_t). Thus, the extended RIR model is much more accurate and flexible. Here, the parameter T_R in Schroeder's RIR and the three parameters of the extended RIR model are blindly estimated. Sub-band analysis is used as the same as the algorithm for calculating the STI. The distortion in seven-octave bands is estimated through the parameters of the RIR model. The approximated RIR for each sub-band can be reconstructed from their envelope modulated with band-limited noise. The wide-band RIR is also approximated from the summation of the sub-band signals based on the superposition principle. Therefore, the estimated acoustical parameters and STI for both sub-band and wide-band can be derived.

The effectiveness and performance of the proposed methods were evaluated. Simulations were performed by estimating the parameters from reverberant and noisy reverberant speech signals. The accuracy of the estimated acoustical parameters was compared with baselines calculated from measured RIRs and existing works. The robustness against various background noise was also evaluated by adding four types of noise with different SNR levels into the reverberant speech signals. The experimental results suggest that the proposed method can correctly, blindly, and simultaneously estimate five-room acoustic parameters, STI, and SNR from a speech signal in reverberant and noisy reverberant environments. The accuracy in terms of standard derivation of the error of the estimator for each parameter, i.e., T_{60}, EDT, C_{80}, D_{50}, T_s, and STI, was 9.4%, 10.5%, 2.7 dB, 14%, 45 ms, and 0.05, respectively. These results of the estimated parameters were close to the standard measurement derived from the RIR.

Keywords: room impulse response, speech transmission index, blind parameter estimation, modulation transfer function, convolutional neural networks.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Prof. Masashi Unoki, my supervision. This research could not succeed without his patient supervision. Pursuit the Ph.D. is such a long journey. Along the journey, many new challenges happen almost every day. There are some days of exhaustion, burnout, and suffering. However, "giving up" has never been in our spirit, students in Unoki-lab. As I was trained to be an engineer before becoming a Ph.D. student, scientific research is, however, no longer an engineering project, so my paradigm could not convey good scientific research. Prof. Unoki persistently motivates and encourages me to morph my view as *a researcher*. He is an expert in broad areas of research. His advice is fruitful when I was struggling. Needless to say, I have learned many things from him and will never forget anything. For example, effectiveness and good results are important; however, the intermediate stages during the process, research strategy, basic principle, and philosophy are much more important to understand/solve scientific issues; even negative results also need thorough consideration, he taught.

Secondly, I appreciate Prof. Masato Akagi, my second supervisor, who always gives brilliant advice and consideration throughout my study. I would also like to express my sincere gratitude to Prof. Huynh, Van Nam, my minor research supervisor. Akagi-sensei and Nam-sensei always give me very kind suggestions. In addition, I also would like to extend my appreciation to Dang-sensei and Yoshitaka-sensei for reading my dissertation, listening to my presentation, and giving constructive questions as well as useful comments.

I wish to express my sincere thanks to my two co-supervisors from Thailand. First, Assoc. Prof. Dr. Waree Kongprawechnon, from Sirindhorn International Institute of Technology (SIIT), Thammasat University. She is similar to my mother during my doctoral degree study. Her generous and kind support helps me to overcome many difficulties. The second co-supervisor is Dr. Jessada Karnjana, who give me this great opportunity to be a Ph.D. student under the supervision of Prof. Unoki. He is not only my bother at NECTEC, but also a mentor in academics. I appreciate tons of his efforts and suggestions.

As a staff of NECTEC/NSTDA, I have the privilege of leaving to study. I would like to take this opportunity to thank Dr. Sarun Sumriddetchkajorn, a former NECTEC's director, Dr. Kanokvate Tungpinmolrut, a research unit director, and Dr. Jasada Kudtongngam. They endorse my leaving with sincere suggestions. I also appreciate the supporting grants in SIIT-JAIST-NSTDA dual doctoral degree program, Monbukagakusho Honors Scholarship, and Thammasat University basic and applied research grant.

Furthermore, many thanks go to my colleagues in the acoustic information science laboratory, especially Takuto Isoyama, Teruki Toya, Dung Kim Tran, Li Kai, Mawalim Olivia Candy, and Kasorn Galajit. They make my JAIST's life to be enjoyable.

Last but not least, I would like to dedicate any of my success to my parents and my beloved wife, Vararat Kongsmai, who always supports me.

Contents

Abstract i

Acknowledgment iii

List of Figures vi

List of Tables ix

Notations xi

Abbreviations xii

1 Introduction 1
- 1.1 Significance and challenges . . . 1
- 1.2 Motivation and research goal . . . 3
- 1.3 Research philosophy . . . 3
- 1.4 Dissertation outline . . . 4

2 Literature review 6
- 2.1 Speech and hearing . . . 7
  - 2.1.1 Some properties of speech . . . 7
  - 2.1.2 Speech intelligibility . . . 7
- 2.2 Room acoustics and its impulse response . . . 9
  - 2.2.1 Measuring room impulse response . . . 10
  - 2.2.2 Schroeder's RIR model . . . 11
  - 2.2.3 The generalized RIR model . . . 11
- 2.3 Modulation transfer function . . . 11
- 2.4 Room acoustic parameters . . . 16
  - 2.4.1 Reverberation time . . . 16
  - 2.4.2 Early decay time . . . 18
  - 2.4.3 Clarity . . . 18
  - 2.4.4 Deulitchkeit . . . 18
  - 2.4.5 Center time . . . 18
  - 2.4.6 Spatial parameters . . . 18
- 2.5 Speech transmission index . . . 19
- 2.6 Blind estimation techniques: state-of-the-art . . . 21
  - 2.6.1 Methods based on the MTF concept . . . 21
  - 2.6.2 Maximum likelihood estimator . . . 22
  - 2.6.3 Multi-channel blind estimation . . . 23

2.6.4 Machine learning and artificial neural networks . . . 25
2.6.5 SNR estimation . . . 26
2.7 Summary . . . 26

3 Blind estimation of speech transmission index 29
3.1 Temporal amplitude envelope of speech and the MTF . . . 29
3.2 Core structure of blind STI estimation . . . 32
3.3 Implementation and evaluation . . . 32
3.3.1 Experimental setup . . . 34
3.3.2 Evaluation matrices . . . 35
3.4 Results and discussion . . . 36
3.5 Summary . . . 39

4 Blindly estimating parameter of RIR model 40
4.1 Sub-band analysis . . . 40
4.2 MTF-based CNN framework . . . 42
4.3 RIR approximation . . . 42
4.4 Evaluations . . . 43
4.5 Results and discussion . . . 43
4.6 Summary . . . 47

5 Blindly estimating room acoustic parameters and STI based on the extended RIR model 49
5.1 The extended RIR model . . . 49
5.2 Core structure of the estimation . . . 50
5.2.1 Objective function . . . 54
5.3 Implementations and evaluations . . . 54
5.3.1 Data augmentation . . . 55
5.3.2 Evaluating estimated parameters of RIR models . . . 56
5.3.3 Evaluating estimated MTFs . . . 56
5.3.4 Evaluating estimated room-acoustic parameters and STI . . . 56
5.4 Estimation in noise environments . . . 65
5.5 Discussion . . . 65
5.6 Summary . . . 67

6 Conclusion 69
6.1 Summary . . . 69
6.2 Contributions . . . 70
6.3 Recommendation and future works . . . 71

Appendices 71

A Controlling estimated STI for protecting privacy of conversation 72
A.1 Feedback controller . . . 72
A.2 Differential evolution optimization . . . 74
A.3 Evaluation and discussion . . . 75
A.4 Summary . . . 78

B Room impulse responses: SMILEdataset 79

C Variation of CNNs for blind parameter estimation **83**
C.1 Experiments and evaluations 83
C.2 Discussion 85

D Supplementary materials **86**

Bibliography **87**

Publications **97**

This dissertation was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Sirindhron International Institute of Science and Technology, Thammasat University.

List of Figures

1.1 Organization of this dissertation. 5

2.1 Block diagram of speech communication in an enclosure. 6
2.2 Example of a clean speech signal with its temporal envelope and spectrum. 8
2.3 A conceptual reflection diagram of a reverberant environment, containing the direct sound, early reflections, and late reverberation components. . . . 12
2.4 Measurement methods of room impulse response: (a) maximum length sequence with $n = 3$ and (b) sine sweep signal. 12
2.5 Example of room impulse response (RIR): (a) RIR signal and (b) its power envelope. 13
2.6 Comparison between Schroeder's RIR and the generalized RIR model in fitting to the envelope of the measured RIR [1]. 13
2.7 Concept of the modulation transfer function (MTF) as a characteristic of the envelope spectrum of any signal that passes through a communication channel (a reverberant and noisy environment). 14
2.8 The MTFs, $\mathbf{m}(f_m)$, of different reverberant environments. 15
2.9 The MTFs, $m(f_m)$, of different noisy environments. 15
2.10 Deriving the reverberation time and early decay time from the energy decay curve of the RIR. 17
2.11 Block diagram of the set-up for measuring and calculating the STI. 20
2.12 Diagram of the previous method based on the MTF concept. 22
2.13 Example of likelihood function of unknown parameter θ from observed data. 24
2.14 Diagram of a single-input multi-channel system: (a) a multichannel model and (b) a sub-channel matching algorithm. 24
2.15 Estimating room acoustic parameter based on artificial neural networks. . . 28
2.16 Framework of estimating SNR using sub-band VAD (Morita *et al.*). 28

3.1 Example signals of (a) clean speech, (b) reverberant speech ($T_{60} = 0.43$ s), and (c) noisy reverberant speech (babble noise, 5 dB SNR) and $T_{60} = 0.43$ s). Dashed lines are power envelopes of each signal 30
3.2 Comparison between TAE and power envelope extracted from the same reverberated speech signal. 31
3.3 Block diagram of the proposed method for estimating STI in noisy reverberant environments using TAE and CNN. 33
3.4 Estimated STIs from observed speech signals in reverberant environment with white noise. 37
3.5 Estimated STIs from observed speech signals in reverberant environment with pink noise. 37

3.6 Estimated STIs from observed speech signals in reverberant environment with babble noise. 38
3.7 Estimated STIs from observed speech signals in reverberant environment with factory noise. 38

4.1 A conceptual diagram of estimating parameter of the RIR model and deriving the room acoustic parameters. 41
4.2 Estimated speech transmission index, STI, from reverberant speech signals. The symbol "o" corresponds to the estimated value from the simulated RIR, "square" indicates the estimated value from the measured RIR, "*" indicates the estimated result using the method proposed by Unoki *et al.* [1], and the dashed line represents the ground-truth calculated from the RIRs. 44
4.3 Estimated reverberation time (T_{60}: Previous method proposed by Unoki *et al.* [2]) and early decay time (EDT) from reverberant speech signals. 45
4.4 Estimated clarity index, C_{80}, from reverberant speech signals. 46
4.5 Estimated Deutlichkeit, D_{50}, from reverberant speech signals. 46
4.6 Estimated center time, T_s, from reverberant speech signals. 48

5.1 Example of complex space and its impulse response. 51
5.2 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. 51
5.3 Fitting results of two RIR models with temporal amplitude envelope of measured RIR: envelopes in time domain (a) and in modulation-frequency domain (b). 52
5.4 Block diagram of proposed method. 53
5.5 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. 57
5.6 Results of estimated parameters of extended RIRs: (a) raising parameter T_h, (b) peak position parameter T_0, and (c) decay parameter T_t. 58
5.7 Actual and reconstructed RIR using proposed method. 59
5.8 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. 59
5.9 Estimated reverberation time (T_{60}) and early decay time (EDT) from reverberant speech signals. 60
5.10 Estimated clarity index, C_{80}, from reverberant speech signals. 61
5.11 Estimated Deutlichkeit, D_{50}, from reverberant speech signals. 61
5.12 Estimated center time, T_s, from reverberant speech signals. 63
5.13 Estimated speech transmission index, STI, from reverberant speech signals. 63
5.14 Block diagram of the proposed method (Scheme III). 64
5.15 Results of the estimated T_{60} in noisy reverberant environments. 66
5.16 Results of the estimated STI in noisy reverberant environments. 66
5.17 Results of the estimated SNRs for sub-bands from reverberant speech signals. 68

A.1 Block diagram of the privacy control based on optimally controlling estimated STI. 73
A.2 Differential evolution algorithm. 74
A.3 The proposed method under two conditions of background noise: (a) the estimated STI at each iteration and (b) the average error. 76
A.4 The proposed method under variations room conditions: (a) the output of the estimated STIs and (b) the controlled parameter of the extended RIR model, T_t. 77

B.1 Examples (I): room impulse responses (RIRs), its envelopes (black dotted line), and RIR models. Solid line is the extended RIR model with model's parameters and dashed line is Schroeder's RIR model. Cross symbol is a position of the Center time, Ts. 81
B.2 Examples (II): room impulse responses (RIRs), its envelopes (black dotted line), and RIR models. Solid line is the extended RIR model with model's parameters and dashed line is Schroeder's RIR model. Cross symbol is a position of the Center time, Ts. 82

C.1 CNN models with a variation of scaling: (a) baseline, (b) width scaling, (c) depth scaling (d) resolution scaling. 84

List of Tables

2.1 Numerical representation of a relation between speech intelligibility and STIs. 20

3.1 Network architecture of the robust STI estimator. 34
3.2 Estimated STIs in various conditions from SMILE corpus in the metrics of RMSE and correlation (ρ) from SMILEdataset. 36
3.3 Estimated STIs of speech signals in RIR and background noise from dataset in the Acoustic Characteristic Environment challenge. 39

4.1 Network architecture of the MTF-based CNN model. 42
4.2 Correlation coefficients between the estimated and calculated parameters. . 47

5.1 Network architecture of MTF-based CNN model. 54
5.2 Correlation coefficients between estimated and calculated parameters. . . . 62
5.3 Comparison between standard derivation (SD) of estimated error and just noticeable difference (JND) of acoustical parameters [3, 4]. 62

B.1 Dataset of room impulse responses (RIRs). 80

C.1 Comparison results of variation of CNN Scaling in the MTF-based parameter of the extended RIR model estimation. 85

Notations

C_{50}	clarity index in speech
C_{80}	clarity index in music
D_{50}	Deulitchkeit
$J(\theta)$	cost function
T_0	peak parameter of the extended RIR model
T_R	reverberation time parameter of RIR models
T_h	raising parameter of the extended RIR model
T_s	center time
T_t	decay parameter of the extended RIR model
T_{60}	reverberation time
δ	Dirac delta function
$\hat{\mathbf{h}}(t)$	estimate of room impulse response
$\mathbf{h}(t)$	room impulse response
$\mathbf{m}(f_m)$	modulation transfer function at a modulation frequency (f_m)
ρ	Pearson correlation coefficient
$\mathrm{H}(\cdot)$	Hilbert transform
$e(t)$	temporal amplitude envelope
$e^2(t)$	power envelope
$h_{ext}(t)$	the extended RIR model
m	modulation index
$n(t)$	noise signal
$x(t)$	input signal
$y(t)$	observed signal

Abbreviations

AM	amplitude modulation
ANN	artificial neural networks
CNN	convolutional neural networks
DNN	deep neural network
DRR	direct-to-reverberation ratio
EDC	energy decay curve
EDT	early decay time
IACC	interaural cross correlation
LEF	lateral energy fraction
LSTM	long short-term memory networks
MTF	modulation transfer function
RIR	room impulse response
RMSE	root-mean-square error
SNR	signal-to-noise ratio
SPL	sound pressure level
STI	speech transmission index
WGN	white Gaussian noise

Chapter 1

Introduction

1.1 Significance and challenges

Speech communication and sound perception are the human basis. Humans take hearing for granted. We gain information by hearing, enjoy music, and convey our thought to the world. The world may be communication between people as well as people to machine or machine to people, e.g., smart assistant, speech-to-speech translation, and car navigation. The world also covers the physical spaces or sound fields. The quality of sound fields affects our lives in many aspects since we live mainly in enclosures surrounded by walls and ceilings. Such surfaces reflect or diffuse a sound wave across the area, known as reflection and *reverberation*. Additionally, most areas contain unwanted sounds or *noise*. Inappropriate reverberation and background noise degrade our intelligibility of speech, as well as enjoyment in music [5,6]. Each building or auditory space is designed for different purposes, so the acoustic quality of each place is considered differently.

In lecture halls and meeting rooms, high quality and intelligibility of speech are expected. Hence, many works regarding acoustics of such auditoriums and hearing were studied [7, 7–10]. A well-designed concert hall or theater can give enjoyment in music, or live performance [11]. The reputation of such halls related to income and benefits primarily depends on the acoustic quality as one of the justifications. Although different areas have different expectations, one of the mutual purposes of every area is safe from unintelligible or misunderstanding in conversation. In common spaces, such as airports, department stores, and stations, acoustic properties are much more crucial because hearing is related to safety, e.g., emergency announcements or fire alarm sound [12]. Hence, architects or acousticians must carefully design and evaluate the acoustic properties of the sound fields corresponding to the purposes of such areas.

Furthermore, the characteristics of a sound field are of interest to many research issues and applications of signal processing. In terms of research issues, the related topics are echo cancellation, noise suppression, and source separation. For speech processing applications, automatic speech recognition (ASR), speaker verification systems, hearing-aided devices, and sound reproduction can make use of the information of the sound field from its acoustical indices. Thus, researchers and engineers give attention to deal with undesired sound environments. Measuring acoustic properties and predicting of user's experience can provide more advantages.

User's experience or subjective aspects, i.e., speech intelligibility and transparency as well as clarity in music, can be evaluated by conducting subjective experiments. A group of listeners is asked to give a score to each sound they heard in a particularly tested

auditorium. However, listening test is expensive, time-consuming, and unreliable. It is still difficult to exchange such subjective descriptions, scores, or feelings from a group of listeners to architects responsible for designing auditoriums or diagnosing acoustic problems. Moreover, it is impractical for real-time applications. Therefore, such subjective descriptions of an acoustic environment are represented by physical properties of room acoustics [13].

In signal and system theory, the relationship of sounds and physical properties of a room, such as absorption coefficient, size, and volume of a room, can be described by room impulse response (RIR) [14]. The RIR can fully represent an enclosure from the source position to the receiver position in the time domain. In the frequency domain, the system transfer function of an acoustic environment can be described by the modulation transfer function (MTF) [15, 16].

The RIR is used to derive room acoustic parameters that represent those subjective perceptions. Likewise, the MTF is used for calculating speech transmission index (STI) for predicting speech intelligibility [17]. Hence, RIR and MTF need to measure in general. Room acoustic parameters or objective indices are derived from the RIR or MTF into well-defined values. Those values represented room acoustic characteristics help architects understand listeners, musicians, and audiences explicitly. Architects, musicians, and sound engineers always utilize room acoustics parameters and objective indices in analyzing a sound field [18].

Many room acoustic parameters and objective indices have been studied [13, 19, 20]. Five useful acoustic parameters and one objective index that are often used in an architectural and acoustic field are focused on in this work. For example, reverberation time (T_{60}) is one of the most influential parameters related to the absorption and reflection of a sound wave. The details of the parameters and indices of interest are described in the next chapter. Those parameters and indices are standardized by ISO and IEC [3, 21].

However, it is difficult to measure RIR or MTF in common spaces where people cannot be excluded. Estimating room acoustic parameters and objective indices from an observed signal are therefore necessary. A running speech signal recorded from a general activity in a room can be used rather than well-designed experiments [22]. Estimating room-acoustic parameters and indices without measuring RIR or MTF is the so-called *blind estimation*.

Blind parameter estimation is challenging not only for room acoustics but also in many fields [23–25]. Because blindly estimating an unknown system is regarded as an ill-posed inverse problem since only an observed signal is measurable. This blind estimation problem is similar to blind deconvolution or blind identification that is an active topic in many fields [23, 24, 26]. Once we can accurately estimate a system transfer function or impulse response, the original data (speech or music) can be restored [2, 27–29].

In addition, evaluating background noise level is also important since both our hearing and speech processing, such as speech enhancement, emergency announcement, speaker localization, and privacy protection, are all suffering at a low signal-to-noise ratio (SNR). Interestingly, there is no existing method could simultaneously estimate room acoustic parameter with the noise level. Hence, an interesting research issue is blindly and simultaneously accurate estimation of various room acoustic parameters, STI, and SNR from an observed signal in noisy and reverberant environments. In particular, the estimated STI is deployed in speech privacy protection for a semi-open space [30, 31]. Lastly, this study might be extended and be beneficial for other applications in many fields, such as blind equalization of transmission channels, seismic analysis, forensics, sound reproduction, virtual reality, and vital medical signs [23, 25, 32–35].

1.2 Motivation and research goal

The motivation for this research has started from my interest in the relationship between hearing perception and characteristics of a sound field. Then, more curiosity followed: what can we do once we are in unknown and difficult to control auditory spaces. From a control engineering point of view, we cannot do anything if we cannot observe them. The questions are how to observe and what parameters of a sound field need to measure. The theory of system identification clearly explains that we need to stimulate the system with well-designed signals [36]. The aforementioned challenges are presented. Is it possible to bring a *de facto* concept of the MTF incorporating into emerging deep learning to overcome these challenging issues? It is anticipated that the critical problem could be overcome. Therefore, this motivation becomes the research goal that is to propose blind methods for estimating room acoustic parameters and speech transmission index from an observed signal in noisy reverberant environments. Estimating room acoustic parameters from degraded speech signals is an active research question. Contribution to the field of research is needed for many current and future applications. Blind parameter estimation is also a challenging issue. Since many existing pieces of research, which are explained in Chapter 2, could provide one or two parameters, it is inadequate to describe room acoustics completely. Estimating multiple parameters is, therefore, much more fruitful. Thus, this estimation is similar to standard methods that can derive multiple parameters from measured RIR. Moreover, the robustness to background noise and a level of noise has not been studied explicitly.

1.3 Research philosophy

The philosophy of this work is one of the truths of the universe that is *Causality* and the *Evidence* of that truth. It is the so-called **"evidence-based estimation from cause and effect of such a thing."** As a noisy and reverberated speech signal is a result of characteristics of a sound field, and such an observed signal is then the effect of reverberation and noise. This phenomenon of caused and effect between speech signal and a sound field can be described by the MTF.

The MTF is used to describe the physical properties of a sound wave in a transmission channel. The signal from a speaker to a listener is smeared by background noise and reverberation. Hence, speech quality and speech intelligibility are reduced according to a reduction of the modulation depth [16, 37, 38]. From psychoacoustics studies, temporal amplitude envelope (TAE) is an important cue to predict speech intelligibility [39]. The TAE and power envelope of the observed signal are considered as physical and psychoacoustical meaningfully features. Furthermore, the STI, one of the parameters of interest used for predicting speech intelligibility, is based on the MTF.

However, the knowledge of the cause and effect based on the MTF cannot fully apply since the reverberated signal is the only information we can measure. Lack of the well-defined modulated input signal causes blind estimation of room acoustic parameters as solving an ill-posed problem. The relevant assumption needs to formulate the appropriate estimator. We assume the reverberation channel as a random process so that in the time domain, such a system can be approximated by using a stochastic RIR model.

A stochastic RIR model for representing an unknown RIR is the second pillar of belief. With regard to an RIR model, many models have been proposed. Each model might be suitable for a particular assumption and the following algorithm. Some of them might

be suitable for a specific case. In this study, we also believe in a principle of *garbage in, garbage out*, so an inaccurate RIR model is prone to provide inaccurate estimated results because of model mismatch with real auditory conditions. After RIR models have been investigated, a new RIR model is introduced to approximate the unknown RIR from noisy reverberant speech signals.

The third pillar of belief is the evidence or data-driven. This is also the basis of statistics and probability theorem. Once we have gathered enough evidence from the cause and effect, we are interested in that is room acoustics. The adequate data that cover as much as the possibility of the relation can help construct a model. The model that has been trained from the appropriate amount of data can be used to predict the relationship of the MTF.

From this philosophy, this research can yield the original and significant contributions to knowledge as the following. For the field of research, this work does not focus on estimating only a specific room acoustic parameter but also expands to five room-acoustics parameters and STI. The extended RIR model and estimating techniques of its parameters have been proposed so as to improve the accuracy of the approximate unknown RIR. The proposed method also correctly estimates those acoustical parameters even if the reverberation consists of background noise. Consequently, the SNR of such conditions is also estimated along with those acoustical parameters.

These contributions make this work different from existing works and bridge the gaps in the field. The estimated acoustical parameters can be used in various speech processing applications, such as dereverberation and speech enhancement, or for improving the quality of emergency announcements. Furthermore, this blind estimation method for parameters of a transmission channel might be applied to other fields. For example, blind kernel estimation for deblurring images, blind equalization of transmission channels, seismic data analysis, and crime scenes identification.

1.4 Dissertation outline

The organization of this dissertation is shown in Figure. 1.1. The rest of this dissertation has been organized as follows.

Chapter 2 introduces background knowledge regarding speech signal in a sound field, characteristics of room acoustics, and some important room-acoustic parameters for describing characteristics of such a giving space. Also, how to obtain these parameters from the direct measuring method and its limitation is presented. The current methods of blind estimation are discussed so that the research gaps are addressed.

Chapter 3 reports the preliminary study for blindly estimating the STI in realistic environments. A method incorporating temporal amplitude envelope into convolutional neural networks is investigated whether or not it can blindly estimate the STI.

Chapter 4 introduces a scheme for simultaneously estimating five-room acoustic parameters and STI rather than formulating a method for estimating a particular parameter.

in Chapter 5 describes a more accurate estimation method by repressing a conventional RIR model with the extended RIR model.

Finally, Chapter 6 summarizes this research. The contributions and recommendations for further research are also pointed out.

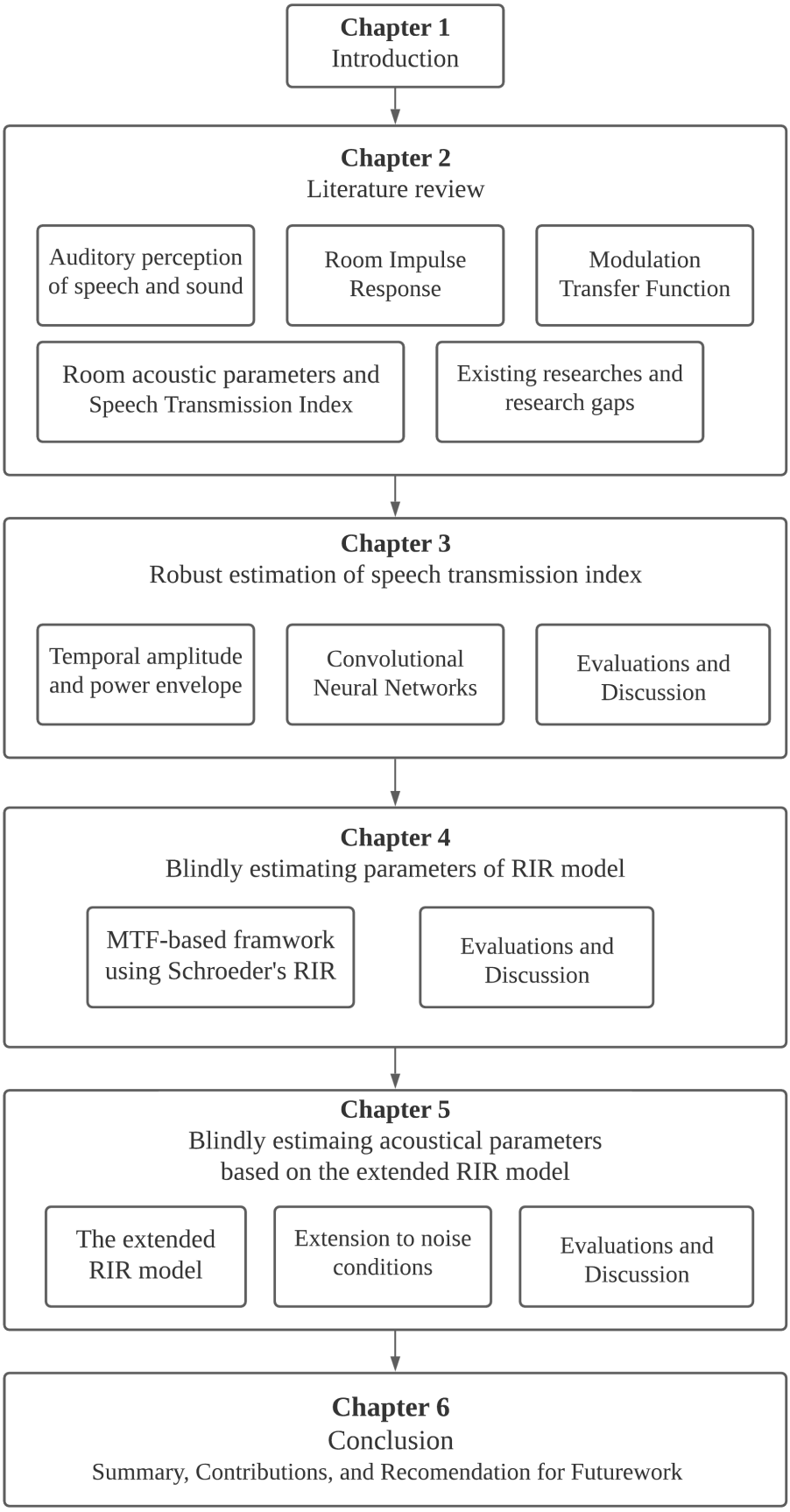

Figure 1.1: Organization of this dissertation.

Chapter 2

Literature review

This chapter introduces theories and concepts of room acoustics related to speech and sound. Some properties of the speech signal and auditory perception are briefly presented so as to understand the effects of room acoustics with speech intelligibility. Then, some of the essential room-acoustic parameters of interest are described. In the last section, the literature on existing techniques for blindly estimating these parameters are reviewed. The advantages and disadvantages of them are thoroughly discussed.

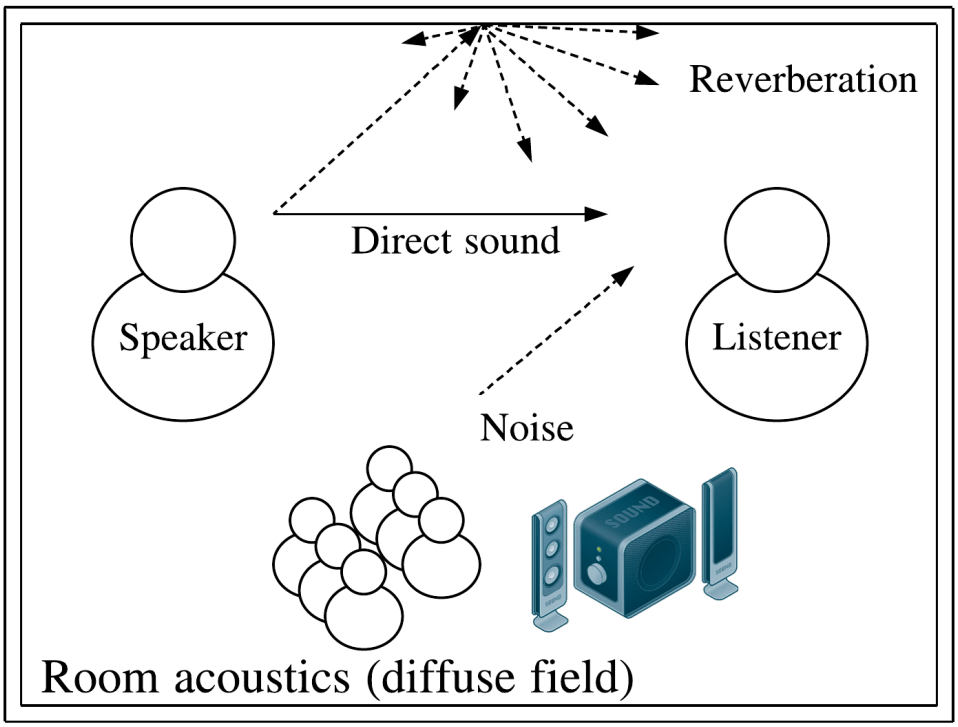

Figure 2.1: Block diagram of speech communication in an enclosure.

2.1 Speech and hearing

Since this research focuses on characteristics of room acoustics related to hearing perceptions, some backgrounds of auditory perception and properties of the speech signal are firstly reviewed. It is well-known that speech is the most natural and effective communication method of humans and humans with machines (human-machine interface), e.g., speech recognition and speaker verification systems. Figure. 2.1 shows primary factors for conversation in bounded spaces, including speakers (sound source), listeners (receiver), and transmission channels (sound field). Knowledge of speech and hearing helps understand the importance of obtaining room-acoustic parameters. Many studies have revealed that the human perceptual system has remarkable ability and more extraordinary than any machine hearing [5, 6, 40]. Hence, acoustic characteristics affect our hearing ability and performance of devices differently, as presented in the following.

2.1.1 Some properties of speech

Auditory systems and speech production are related to speech intelligibility in any situation no matter in room acoustics or in any communication channels. Basic properties of speech are therefore useful for understanding the reasons that why room acoustic parameters and objective indices can be used to predict speech intelligibility as well as speech quality.

Characteristics of speech are dynamic and non-stationary signals. Spectral information is important in speech perception as well as the perception of musical instruments. The smallest unit of speech is called *phoneme*. Each phoneme has a specific spectrum. The peaks in the spectra are called *formants*. The first peak is called the *fundamental frequency* (F_0). In a complex sound as speech, F_0 is varied along with an utterance. F_0 also represents the information of the speech production (i.e., glottal or vocal-fold) [5]. Hence, many speech applications use F_0 as an important cue. Besides F_0, *level* is varied to emphasize individual syllables, words, or entire sentences. The level can express non-linguistic information (e.g., vocal-emotion recognition). Also, different speeds can convey some meaning such as a feeling of urgency. The speed variation is called *tempo*. A combination of these variations in speech is called *prosody*.

Speech signal can be regarded as a combination of fine structure and temporal component. In particular, the modulation frequencies between 1 Hz and 16 Hz contribute to intelligibility, and the dominant frequency is at 4 Hz [39, 41, 42].

2.1.2 Speech intelligibility

There might be some confusion between two terminologies: *speech intelligibility* and *speech quality*. Speech intelligibility is a subjective quantity while speech quality can be measured by using multi-dimensional manners [43]. Yet, speech intelligibility depends on sound quality. In other words, speech intelligibility is regarded as one of aspects of speech quality [43]. Speech intelligibility also has a high correlation to *clarity* of speech.

On the other hand, speech quality covers more dimensions than speech intelligibility [44]. The speech quality can be evaluated by using a method, i.e., the perceptual evaluation of speech quality (PESQ) [45]. To understand speech intelligibility, psychoacoustic studies have revealed the important dimensions of auditory perception as follows.

- Sound intensity or sound pressure level (SPL), as defined as perceived loudness

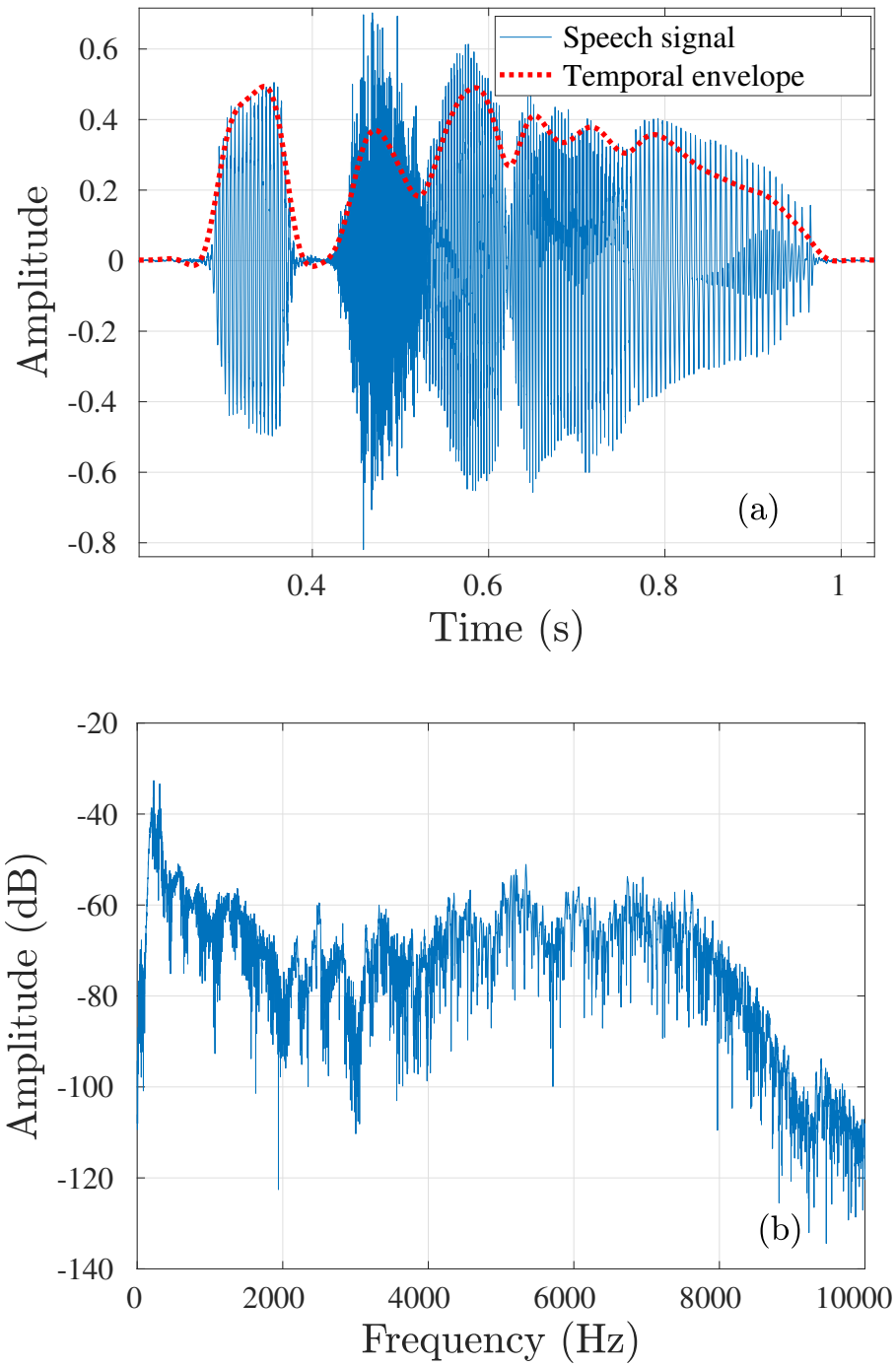

Figure 2.2: Example of a clean speech signal with its temporal envelope and spectrum.

- Major frequency components, as defined as a perceived pitch
- Temporal pattern and rhythm, as defined as fluctuations
- Interaural disparities (i.e., differences across the ears) related to spatial location

By definition, speech intelligibility is comprehensible or recognizable in a message of a speaker [5, 46]. It can be assessed by a proportion of speech (i.e., syllables, words, or sentences) that is correctly repeated by a group of listeners (normal hearing) in a given experiment. In the experiments, a set of words should follow the word familiarity-controlled word-lists [47, 48]. Each word is played only one time, and listeners write the words they heard. The number of corrected words that subjects can repeat is used to calculate *articulation index* (AI) [49]. Later, speech intelligibility index (SII), which is the extended index from the AI, has been standardized [50, 51].

Even though subjective tests by a group of listeners can be conducted, it is imperfect, unreliable, time-consuming, and expensive. Therefore, many objective methods have been proposed such as hearing aid speech intelligibility index (HASPI) [52, 53] and short-time objective intelligibility (STOI) [54]. The underline concept of STOI is based on a correlation between the temporal envelope of clean speech and noisy speech signals. An example of the temporal envelope of a speech signal under noisy and noise-free condition is shown in Fig 2.2 (a). In [55], the STOI was compared with the STI for predicting intelligibility. Nevertheless, it was found that the STI, which is described later, is more suitable for applying in acoustic spaces whereas the STOI is recommended for general speech communication systems.

In the early psychoacoustic study, a transmission channel referred to telephone communication [56]. However, a transmission channel in this study means room acoustics which is of great importance regardless of the quality of sound sources or receivers since noise and reverberation degrade speech quality and intelligibility [16, 38, 57, 58]. A condition of background noise and speech can be expressed by the signal-to-noise ratio (SNR). SNR is a ratio between the energy of signal and noise. The SNR greater than 30 dB is regarded as a clean speech signal, and 0 dB SNR is the equal level of the signal and noise. To be intelligible, the threshold of speech intelligibility in terms of the SNR, it should be greater than -30 dB [59]. More details of noise are presented in Chapter 4. Secondly, the characteristics of room acoustics in an assumption of the noise-free condition can be characterized by using measuring room impulse response (RIR).

2.2 Room acoustics and its impulse response

In a room, a sound wave propagates in all directions from the source position before reaching the listener. Let's assume an enclosure as a LTI system, an energy of a sound wave (speech or music) that travel in a sound field can be expressed a convolution of an sound source and room impulse response (RIR), that is

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

where $y(t)$ is the observed signal, $x(t)$ is the original signal, $h(t)$ is the RIR, and the asterisk ('$*$') indicates convolution operation. Also, an energy of a signal, $g(t)$, is a signal with finite energy, that satisfies

$$\int_{-\infty}^{\infty} |g(t)|^2 dt < \infty. \tag{2.2}$$

The listener perceives the sound not only a single path from sound sound but also the multiple reflections. Those reflections are caused by the walls, floors, ceiling, and furniture in a room and reach the listener at various times and energy. The direct sound wave from the source to the listener calls *direct sound*. In room temperature of 20 °C, the speech of sound waves through the air is about 343 m/s. The time delay between the source to listener positions depends on the distance. In the standard procedure, the minimum distance between source and microphone is 1.5m [3]. The energy of the reflect components that close to the direct sound is called *early reflection*. The last component of the reflection is referred to as *the reverberation*. Figure 2.3 shows an ideal reflection component of an RIR.

In a rectangular room, the amplitude of the reflect component decays exponentially with time. Figure 2.3 depicts a simple reflection diagram of a reverberant environment. It can see that the first component might take a few milliseconds from the sound source position. This might be caused by the reflection of a sound with a surface before reach to the receiver. The later components are the sound energy that reflects from any surfaces of a room, e.g, walls, ceiling, partitions, floor, and furniture as well as occupants.

2.2.1 Measuring room impulse response

Obtaining RIR or transfer function is a complete description of a linear time-invariant (LTI) system. This kind of system identification needs a wide frequency spectrum as much as possible. In theory, Dirac delta function/impulse, δ, which is infinity power at time zero, is defined. In practice, a short unit impulse signal is used, such as a gunshot, bursting balloons, or any stimulus covering the full spectrum of interest [13]. The stimulus signal (sound source) must have enough energy in that range and reliable for many points of measurements. As a result, white noise, which is a flat spectrum, or pink noise is often used. Besides the stimulus signal, equipment needs to be well prepared, as in ISO $3382-1$ source requirements [3]. A loudspeaker is used to excite the signal. It is preferred to generate the excitation signals for all directions. Thus, an omnidirectional loudspeaker that bundles with 10 to 20 loudspeakers, , e.g., a so-called Dodecahedron loudspeaker, is recommended. A high-quality microphone also needs omnidirectional receiving characteristics. The output of this method is the measured RIR.

Furthermore, there are two more complicated methods that need post-processing: maximum length sequences (MLS) and exponential swept sine (ESS) method [13, 60, 61]. First, the MLS are periodic binary sequences, e.g. , -1 and $+1$. They are generated by the period length of L, and $L = 2^n - 1$. Figure 2.4 (a) shows an example of a binary sequence. The 3th order of the shift register generate the sequence while the outputs of a certain stage is fed back to the input. The autocorrelation function of a discrete signal is applied. Second, exponential swept sine method is varying frequency of sinusoidal signal with constant magnitude, as shown in Fig.2.4 (b). Then, the measured RIR is calculated from *matched filtering* that is the inverse Fourier transform of the system. See in [13] for more detail. It was found that ESS has some advantages over MLS in terms of robust to background noise and non-linearity of a system [62, 63]. Lastly, the impulse response is needed to be passed through octave band filtering. The following room acoustic parameters are then calculated for each sub-band. The octave bands are

range from 125 Hz to 4000 Hz [64].

As room temperature and pressure affect the speed of sound, a thermometer is required. A sound field is regarded as a linear system. In the time domain, room impulse response (RIR) fully represents the acoustic characteristics. In the frequency-domain, system transfer function using Fourier transform of an RIR is also preferred [65–67]. The realistic impulse response and its power envelope are shown in Fig. 2.3.

2.2.2 Schroeder's RIR model

According to the definition of the reverberation time introduced by Sabine, the relation between delta function excitation in a rectangular room can be expressed as

$$-60 = 10 \log 10 \exp(-2\delta T_{60}). \tag{2.3}$$

Hence, the reverberation time is follows

$$T_{60} = 3 \ln(10)/\delta \approx 6.91/\delta. \tag{2.4}$$

Manfred R. Schroeder proposed a stochastic model based on the above assumption, i.e., exponential decaying function, namely Schroeder's RIR model [15]. Schroeder's RIR model is defined as

$$\mathbf{h}(t) = e(t)n(t) = a \exp(-6.9t/T_{60})c_h(t), \tag{2.5}$$

where $\mathbf{h}(t)$ represents room impulse response, $e(t)$ is envelope of the RIR, a is a gain factor of RIR, T_R is reverberation time, and $n(t)$ is a stationary white-noise process [68].

2.2.3 The generalized RIR model

Later, Unoki *et al.* introduced a more flexible RIR model by modifying from Schroeder's RIR, namely the generalized RIR model [1]. The generalized RIR model has one more parameter, so-called parameter b. Hence, it is more flexible and well fit to measure RIRs than Schroeder's RIR. The generalized RIR model is defined as

$$\mathbf{h(t)} = e(t)n(t) = at^{b-1} \exp(-6.9t/T_R), \tag{2.6}$$

where b is the order of the RIR. This equation become the same as Schroeder's RIR at $a = 1$. The comparison between Schroeder's RIR and the generalized RIR is shown in Fig. 2.6. Unfortunately, the parameter b of the generalized RIR model has no physical meaning in representing the shape of the envelope of RIR. Therefore, this model is not included in the next investigation.

2.3 Modulation transfer function

The modulation transfer function (MTF) is a concept that is widely used in fields of physic, optic, and acoustics [69]. In room acoustics, M.R. Schroeder first proposed the definition of the MTF and its measurement method for auditory system [15]. In the meantime, Houstgast and Steeneken proposed an objective index to predict speech intelligibility from the quality of a transmission channel based on the MTF [16]. Figure 2.7 shows the concept of the MTF.

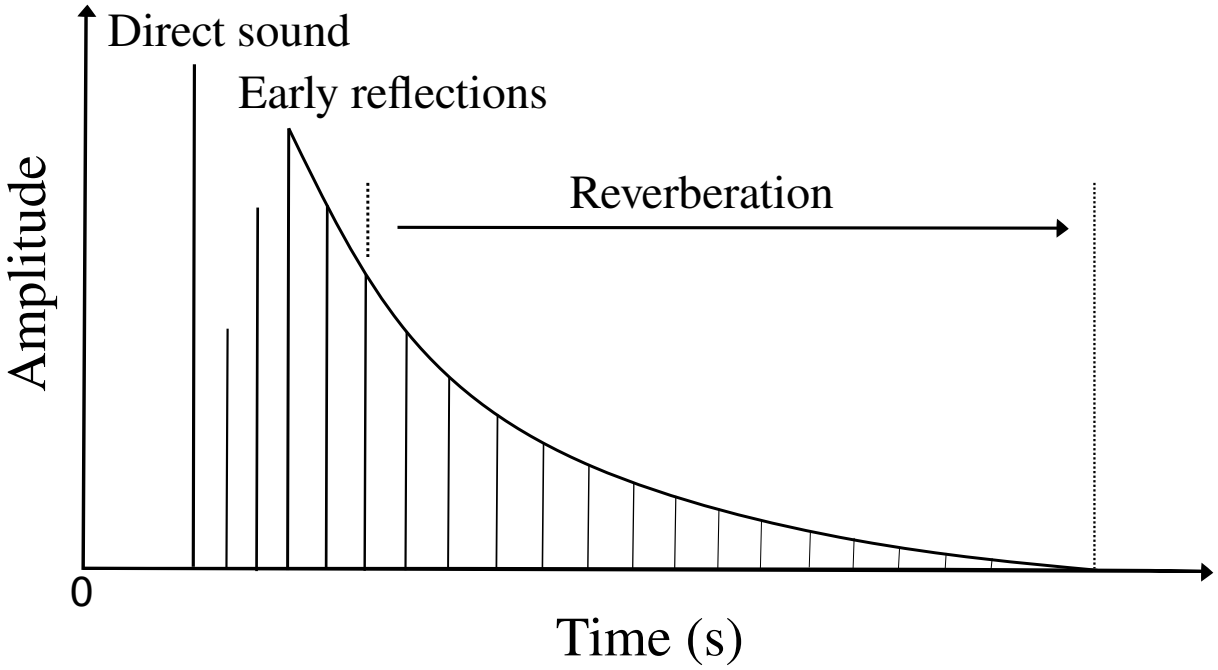

Figure 2.3: A conceptual reflection diagram of a reverberant environment, containing the direct sound, early reflections, and late reverberation components.

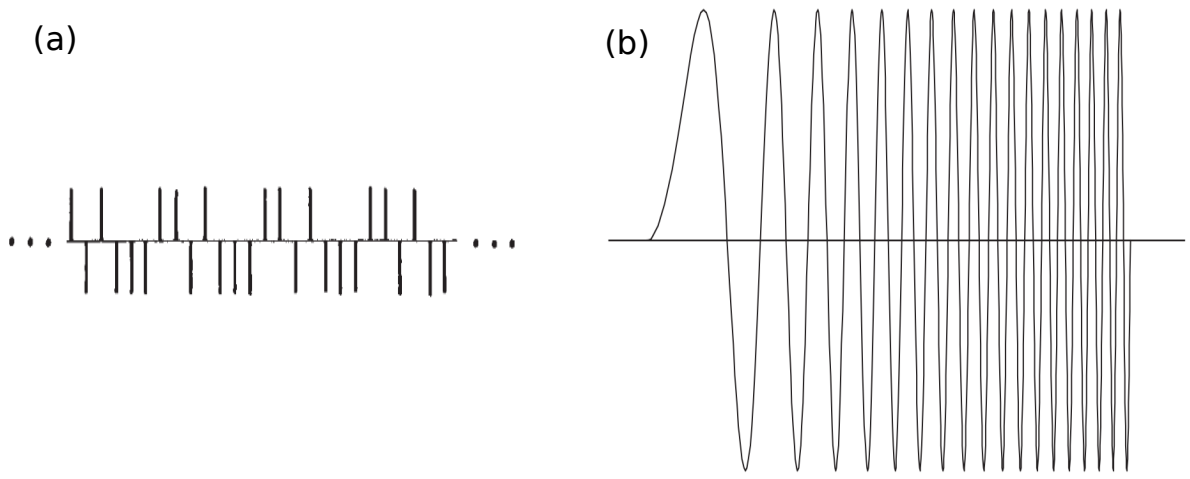

Figure 2.4: Measurement methods of room impulse response: (a) maximum length sequence with $n = 3$ and (b) sine sweep signal.

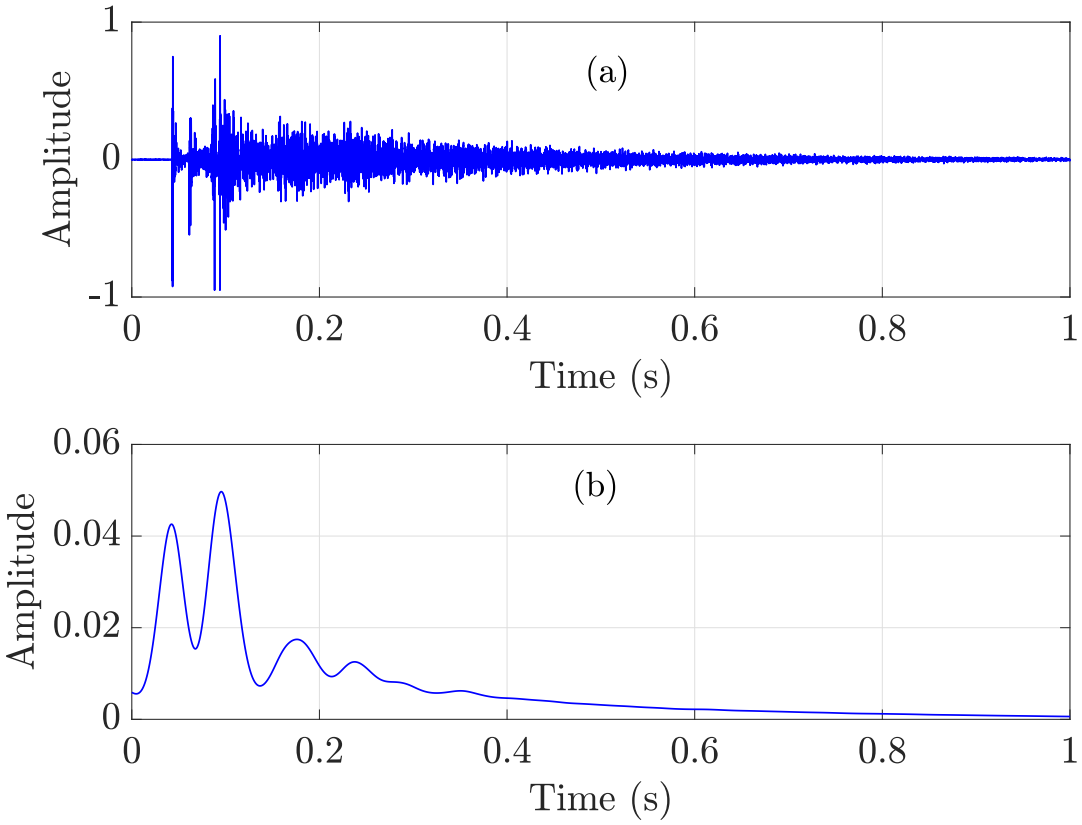

Figure 2.5: Example of room impulse response (RIR): (a) RIR signal and (b) its power envelope.

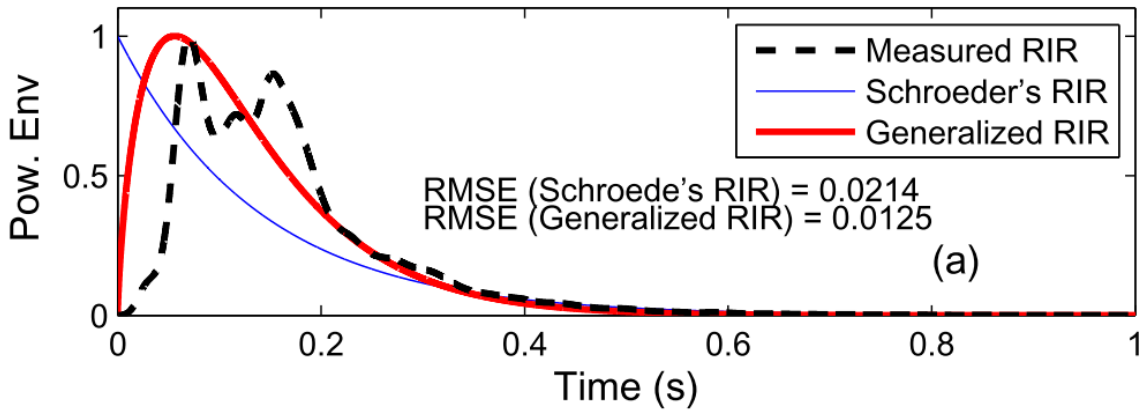

Figure 2.6: Comparison between Schroeder's RIR and the generalized RIR model in fitting to the envelope of the measured RIR [1].

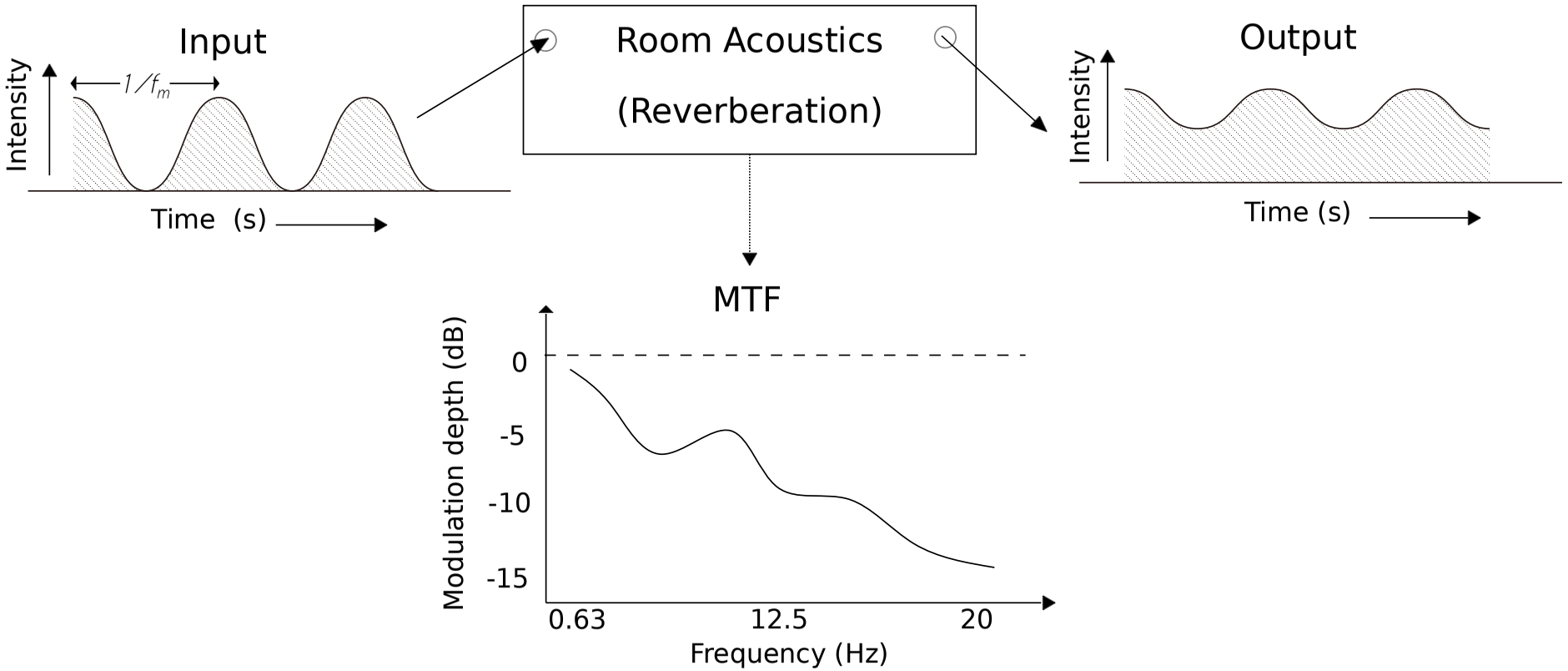

Figure 2.7: Concept of the modulation transfer function (MTF) as a characteristic of the envelope spectrum of any signal that passes through a communication channel (a reverberant and noisy environment).

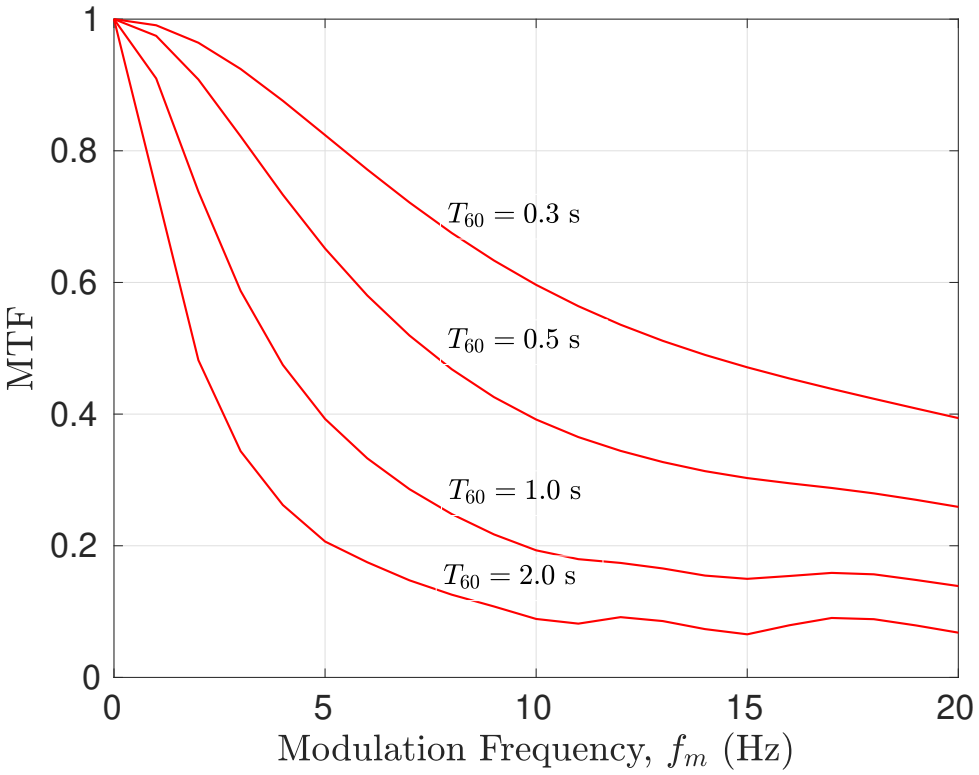

Figure 2.8: The MTFs, $\mathbf{m}(f_m)$, of different reverberant environments.

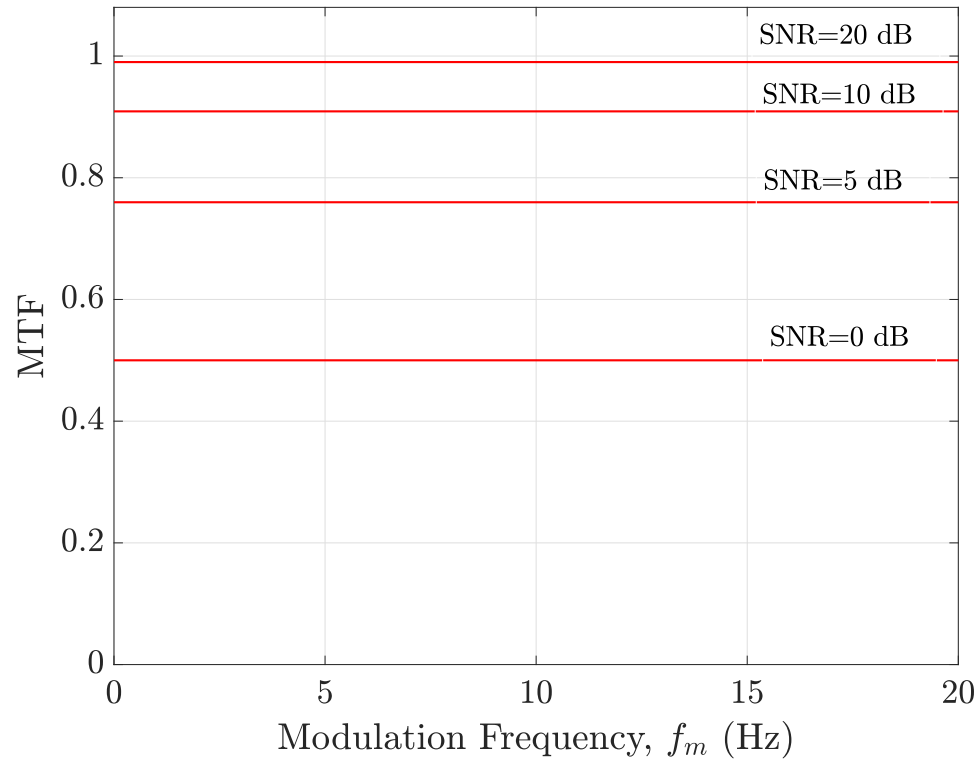

Figure 2.9: The MTFs, $m(f_m)$, of different noisy environments.

The characteristics of an auditory space that consist of reverberation and/or noise can be represented by the MTF. Figure 2.11 shows the relationship between reverberation and/or noise and MTF.

RIR is used to characterize a sound field in the time domain. In the frequency domain, a system transfer function in terms of the MTF describes the same system by the modulation distortion between input and output modulated signals. The MTF is derived by a fraction between the Fourier transform of squared RIR and its total energy. The complex MTF is defined as

$$m(f_m) = \frac{\int_0^\infty h^2(t)\, \exp(-j2\pi f_m t)dt}{\int_0^\infty h^2(t)dt} \left(\frac{1}{1 + 10^{(-\mathrm{SNR}/10)}} \right), \tag{2.7}$$

where $m(f_m)$ is the MTF at modulation frequency f_m and $h(t)$ is an room impulse response. The term in parenthesis is defined later in order to take SNR into account. Figure 2.11 shows the MTF concept.

2.4 Room acoustic parameters

Subjective aspects in speech and music assessments, such as speech intelligibility and music clarity, can be objectively expressed by using room acoustic parameters and objective indices [4, 13]. Many useful room acoustic parameters have been studied to describe the physical properties of an acoustic environment. The physical properties of a given room are related to architectural acoustics such as reverberation.

2.4.1 Reverberation time

In the early 19th century, P.E. Sabine was the first researcher, who formulated a measurement of reverberation in a room [19]. With only his ears and a stopwatch, he could measure the time after the last sound source has been emitted. This time of sound energy reflected by room acoustic is reverberation time. Reverberation time represents the physical property of a source field related to energy reflecting of source wave and a room. The reverberation time (T_{60}) is the duration of sound decay in seconds. T_{60} is derived from an energy decay curve (EDC) of the RIR. The EDC is calculated from the energy of the RIR in dB as follows.

$$\mathrm{EDC} = 10 \log_{10} h^2(t). \tag{2.8}$$

Moreover, reverberation time is frequency-dependence. Thus, it is usually considered in octave bands [64]. The energy decay curve is fitted by using linear regression. T_{60} is the time period that the fitted line intersects with -60 dB. In practice, the decay curve between -5 dB and -35 dB below the maximum initial level is recommended to avoid interference of noise [3]. Figure 2.10 shows an example of the energy decay of an RIR and its fitting lines. T_{60} is calculated from the slope of the fitting line between -5 dB and -35 dB as

$$T_{60} = \frac{60}{\mathrm{slope}}. \tag{2.9}$$

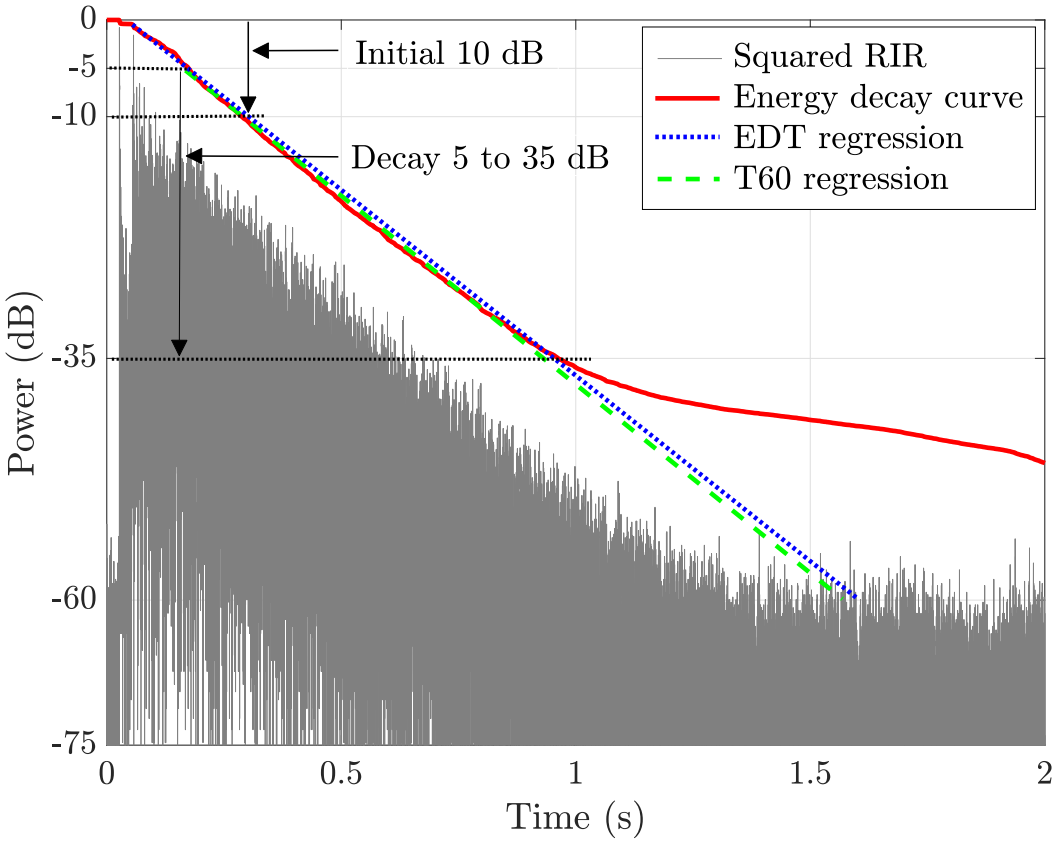

Figure 2.10: Deriving the reverberation time and early decay time from the energy decay curve of the RIR.

Note that ISO 3382 specifies a minimum distance between source and receiver at least 1.5m [3]. This distance is known as the critical distance.

2.4.2 Early decay time

Some room acoustic parameters have mutual relation [70]. The energy decay curve was studied in terms of perception with respect to reverberation property [20]. It was found that early of the energy decay curve between initial to 10 dB is strongly related to the perception of reverberation than T_60 [71]. This regression using the early energy decay curve is defined as the early decay time (EDT).

2.4.3 Clarity

Clarity index or (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

$$C_{80} = 10 \log_{10} \frac{\int_0^{80\text{ms}} h^2(t)dt}{\int_{80\text{ms}}^{\infty} h^2(t)dt}. \tag{2.10}$$

2.4.4 Deulitchkeit

Deulitchkeit, D_{50}, is also related to the energy ratios between the early and late reflection of the RIR. Note that, in some documents, D_{50} might be named as *Definition* [13]. D_{50} is used to evaluate the speech intelligibility of lecture halls or classrooms (in percentage). D_{50} is defined as

$$D_{50} = \frac{\int_0^{50\text{ms}} h^2(t)dt}{\int_0^{\infty} h^2(t)dt} \times 100. \tag{2.11}$$

2.4.5 Center time

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

$$T_s = \frac{\int_0^{\infty} h^2(t)t\,dt}{\int_0^{\infty} h^2(t)dt}. \tag{2.12}$$

For speech, a low value of T_s indicates the high speech intelligibility. It was recommended that should not exceed about 80 ms [13].

2.4.6 Spatial parameters

Even though this research focuses on estimating room acoustic parameters from a single-channel input, subjective descriptions related to the sense of space and direction of sound

are useful in speech applications. Hence, some interesting acoustic parameters to express spatial perception or *sensation of space* are briefly provided. The following parameters need at least two measuring points of receivers (i.e., binaural impulse response) or more.

Lateral energy fraction (LEF) and the late lateral energy (LG_{80}):

$$\mathrm{LEF} = \frac{\int_{5\mathrm{ms}}^{80\mathrm{ms}} \left(h(t)\cos\theta\right)^2 dt}{\int_{5\mathrm{ms}}^{80\mathrm{ms}} h^2(t)dt}, \tag{2.13}$$

where θ is the angle between the direction of sound source and the ears of a listener.

Inter-aural cross-correlation coefficient (IACC): It is the maximum delay time interval of the cross-correlation function between the left and right ear. IACC is defined as

$$\Psi_{rl} = \left(\int_0^\infty h_1(t)h_2(t+\tau)dt\right) / \left(\int_0^\infty h_1^2(t)dt \int_0^\infty h_2^2(t)dt\right)^{\frac{1}{2}}, \tag{2.14}$$

$$\mathrm{IACC} = \max|\Psi_{rl}|, \quad -1\mathrm{ms} < \tau < 1\mathrm{ms}, \tag{2.15}$$

where Ψ_{rl} is the normalised inter-aural cross correlation function, $h_1(t)$ and $h_2()$ are the impulse response at the left and right ear.

Lastly, there are some remaining parameters that are not defined as architectural acoustic parameters as the aforementioned parameters. However, they are interested in the engineering field and could be used to characterize auditory spaces. In 2015, the acoustic characterization of environments (ACE) challenge was held to determine the state-of-the-art in blind acoustic parameter estimation [72]. Besides reverberation time (T_{60}), the direct-to-reverberation ratio DRR is another parameter that might need to estimate blindly [29, 73]. The DRR is similar to those energy ratio parameters with a few different calculations [74]. The DRR is defined as

$$\mathrm{DRR} \;=\; 10\log 10\left(\frac{\int_{n_d-n_0}^{n_d+n_0} h^2(t)dt}{\int_0^{n_d-n_0} h^2(t)dt + \int_{n_d+n_0}^{\infty} h^2(t)dt}\right), \tag{2.16}$$

where n_d is the time of the direct sound and n_0 is 2.5 ms. Note that n_0 is defined specifically by the ACE to represents the time of an additional path difference in their setup [72].

2.5 Speech transmission index

The speech transmission index (STI), which is an objective index, is used to predict speech intelligibility and listening difficulty. The quality of a transmission channel from a talker to a listener can be indicated by a signal number. Houstgast and Steeneken proposed the STI based on the modulation transfer function (MTF) [17, 21, 75].

The MTF can be regarded as transfer function of a linear system. The MTF represents the characteristics of a transmission channel as a function of the modulation frequency and the decrease of modulation depth [13]. The power envelope of any amplitude modulated AM signal is observed. In room acoustics, the MTF concept is used to quantify the effects of reverberation and noise. The higher reverberation, the lower the modulation

Table 2.1: Numerical representation of a relation between speech intelligibility and STIs.

Quality	Bad	Poor	Fair	Good	Excellent
STI	0.0 - 0.30	0.30 - 0.46	0.46 - 0.60	0.60 - 0.74	0.75 - 1.0

depth of modulated signals that pass through the room. A nonlinear relationship between reverberation time and MTF can be demonstrated in Fig. 2.8. The modulation distortion ratios between the input envelopes and the corresponding outputs are known as modulation indices. The magnitude MTF is defined as

$$m(f_m) = \frac{\left| \int_0^\infty h^2(t)\, \exp(-j2\pi f_m t) dt \right|}{\int_0^\infty h^2(t) dt}, \tag{2.17}$$

where $m(f_m)$ is the MTF at modulation frequency f_m and $h(t)$ is an room impulse response.

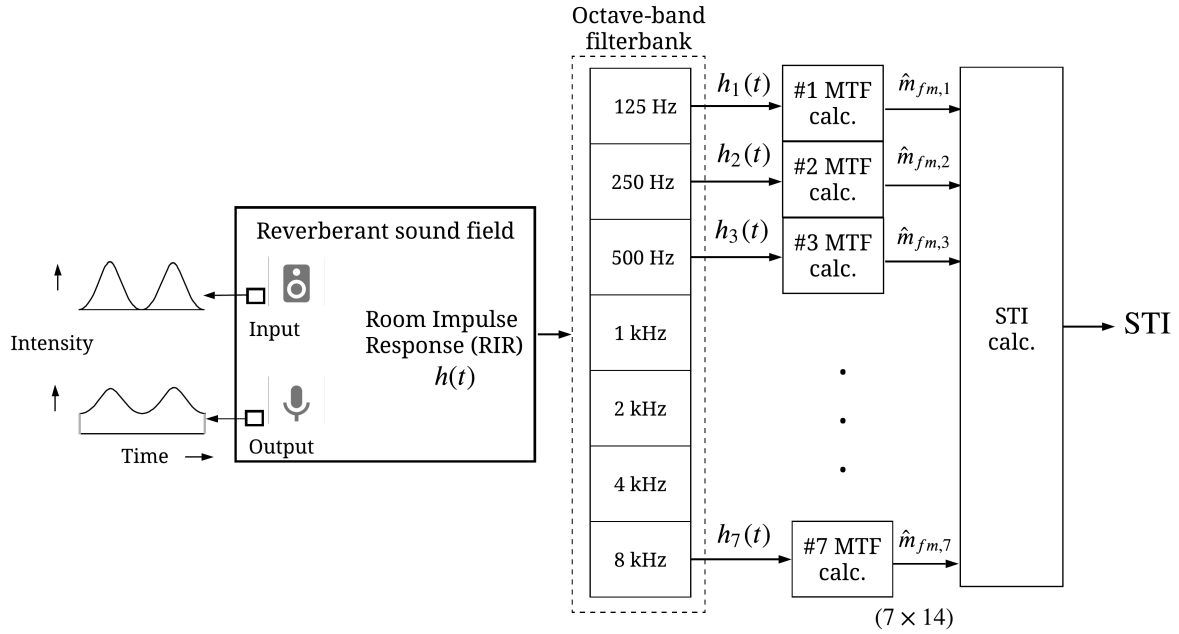

Figure 2.11: Block diagram of the set-up for measuring and calculating the STI.

The STI method has been standardized by IEC 60268 − 16 [21]. Fig. 2.11 shows a diagram of the setup 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

$$N(k,i) = 10 \log_{10} \left(\frac{m_{(f_{m_{i,k}})}}{1 - m_{(f_{m_{i,k}})}} \right), \tag{2.18}$$

where $i = 1$ to 14 and $k = 1$ to 7. Those values are limited to the range of −15 dB to +15 dB and are normalized, called transmission indices (TIs). For each k and i, the TI,

$T(k,i)$, is calculated by normalizing the corresponding modulation distortion, $N(k,i)$.

$$T(k,i) = \begin{cases} 1, & \text{if } 15 < N(k,i), \\ \frac{1}{30}\left(N(k,i)+15\right), & \text{if } -15 \leq N(k,i) \leq 15, \\ 0, & \text{if } N(k,i) < 15. \end{cases} \tag{2.19}$$

Then, the modulation transmission indices (MTIs), $M(k)$, are calculated by averaging $N(k,i)$ as

$$M(k) = \frac{1}{14}\sum_{i=1}^{14} N(k,i). \tag{2.20}$$

Finally, STI is calculated from multiply with weighting factors, as

$$\text{STI} = \sum_{k=1}^{7} W(k)M(k) \tag{2.21}$$

where $W(k)$ represents the octave band weighting factors. The contributions to the index are $W_1 = 0.129$, $W_2 = 0.143$, $W_3 = 0.144$, $W_4 = 0.114$,$W_5 = 0.186$, $W_6 = 0.171$, and $W_7 = 0.143$. 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. (2.17), known as the indirect method [21].

2.6 Blind estimation techniques: state-of-the-art

Blind estimation of the above parameters and indices is interesting. In the past decade, many researchers have proposed techniques to solve this issue. Those methods can be considered in three approaches: (1) methods based on the MTF concept, (2) statistical approach, and (3) machine learning approach.

2.6.1 Methods based on the MTF concept

From the above definition of the MTF, Unoki *et al.* proposed methods based on the MTF concept to estimate T_{60} and STI [1, 76, 77]. The authors also used the similar concept in speech applications, such as dereverberation, envelope of original speech, and voice activity detection [78]. From the observation of a reverberant signal and the MTF, there are there useful characteristics that were used to formulate the constraints of the estimations. First, a DC signal does not affect by reverberation, so the MTF of any signal at zero Hz is equal to one. Second, the MTF of a dominate frequency is close to the MTF of the input signal. Third, the MTF of the input signal decreases when the reverberation increases. Based on these properties, the power envelope of an input signal could be restored by using an inverse MTF method. The method provided coefficients of the IIR filter of the inverse MTF. Figure 2.12 shows a diagram of the method based on the MTF concept.

Later, Unoki *et al.* proposed a blind method for estimating the STI. This method is based on the generalized RIR model and the basic concept of the MTF. According the MTF concept, the power envelope of an observed signal, $y(t)$, is approximated as

$$e_y^2(t) = e_x^2(t) * e_h^2(t) + e_n^2(t). \tag{2.22}$$

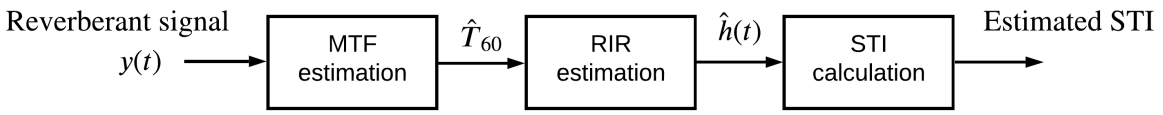

Figure 2.12: Diagram of the previous method based on the MTF concept.

The definition of the MTF of a system, namely, its frequency transmission characteristics, is presented by the fraction of the Fourier transform of the response of the system and its total energy [16]. The MTF at a modulation frequency f_m, $m(f_m)$, is defined as

$$m(f_m) = \frac{\int_0^\infty h^2(t) e^{-j2\pi f_m t} dt}{\int_0^\infty h^2(t) dt}, \tag{2.23}$$

The unknown RIR can be obtained by estimating two parameters, i.e, T_{60} and b in Eq. (2.6). The STI can be then calculated indirectly from this estimated RIR using the definition of the MTF in Eq. (2.17). T_{60} and b are estimated on the three specific conditions and assumptions: the MTF at 0 Hz is 0 dB, the original modulation spectrum at the dominant modulation frequency f_d is the same as that at 0 Hz, and the entire modulation spectrum of the reverberant signal is proportionally reduced by the reverberation time [1]. Thus, these relations can be used for estimating the T_R and b of the RIR model by minimizing the root mean square (RMS), defined as

$$\mathrm{RMS}(T_{60}, b) \; = \; \sqrt{\frac{1}{2}\sum_{l=1}^{2}[|E_y(f_{m_l})| - m(f_d, T_{60}, b)^2]}, \tag{2.24}$$

where $E_y(f_{m_l})$ is the modulation spectrum of the envelope of a reverberant signal $y(t)$ at a specific frequency f_{m_l} and $m(f_d, T_R, b)$ is the derived MTF at the frequency f_d from the RIR model, as in Eq. (2.7). The SNR is estimated from the mean power ratio of speech sections to noise sections using robust voice activity detection. This estimated RIR is then used to calculate MTF and STI.

In their proposed methods, the RIR was first approximated using Schroeder's RIR model and later by the more general model, namely the generalized RIR model. The concept of the MTF was used to restore the modulation spectrum from the observed signal. Then, the optimal parameters of the RIR models were calculated. The estimated STI was derived from the MTF of the generalized RIR model. The generalized RIR model has proposed by modifying Schroeder's RIR model [1]. The generalized RIR model is more accurate and closer to the measured RIRs than Schroeder's RIR model. The model has two parameters. The first parameter (i.e., T_R) represents reverberation time, but the second parameter (i.e., the parameter b) has no physical meaning [1].

2.6.2 Maximum likelihood estimator

A maximum likelihood estimation (MLE) is based on statistics for estimating parameters from observed data. Ratnam *et al.* first proposed a maximum-likelihood procedure for estimating reverberation time. They modeled impulse response as an exponentially

damped Gaussian white noise process [79]. Later, Kendrick *et al.* used a similar method to approximate energy decay curves from reverberant speech and music, but they proposed a new decay model, i.e., multiple decay curves [80–82]. In the meantime, J. Wen deployed the MLE into a number of statistical models for estimating decay rates from those models [83]. As the above mention, the energy decay curve is used for calculating the reverberation time and early decay time.

A maximum likelihood estimator (MLE) is a statistical method to estimate an unknown parameter θ [84]. A statistical process is described by the probability density function (PDF).

Let θ be a parameter of interest (unknown) and $x_1, \ldots, x_n$ be a random sample that are identical and independent distributed (i.i.d). A function that describes a result of a process in terms of sample data $x_1, \ldots, x_n$ related to parameter θ is called likelihood function. A likelihood function of a parameter θ is express as $L(\theta|x_1, \ldots, x_n)$.

Figure 2.13 shows an example of a likelihood function of a parameter θ as a normal distribution. Hence, the peak position of the function is the so-called maximum likelihood.

To obtain such a likelihood function, joint PDF is used to describe the likelihood function of the observed sample. The PDF of samples, $x_1, \ldots, x_n$, denotes as $f(x_1, \ldots, x_n \mid \theta)$.

Then, taking the derivative to get the maximum point of the likelihood function.

$$\frac{\partial L(\theta)}{\partial \theta} = 0. \tag{2.25}$$

The result is the value of the unknown parameter θ. Computing MLE is usually by taking log into the function to make it easier. It is known as the Log-likelihood estimator. In Kendrick's works, quadratic programming is the technique to search the maximum value of the likelihood function. The MLE approach has been used until nowadays research with different additional methods after formulating those likelihood functions from the observed data, e.g., in [85].

2.6.3 Multi-channel blind estimation

Blind channel identification has been studied for speech processing and communication applications in order to avoid a time period for identification with an exciting input. A number of state-of-the-art adaptive blind channel algorithms have been proposed. High order statistics (HOS), which have greater than second order, were proposed. Two conditions are necessary and assumed to formulate a blind acoustic channel based on a single input multiple outputs (SIMO) system. First, all the channels do not share any common zeros. Second, the autocorrelation matrix of the input signal is full rank.

- Least mean square
- Newton algorithm

However, those algorithms have many drawbacks and limitations. For example, since the algorithm requires a greater number of outputs than the inputs, the system needs more receivers (microphones) placed in different positions. The accuracy of HOS also depends on a number of observations. A small number of receivers cannot be accurately computed. Moreover, the algorithms are intensive computational complexity, slow convergence, and local minimum. This approach also requires multiple channels and multiple receivers, so these requirements are far beyond the scope of this study which is taking only a single channel of observed signal.

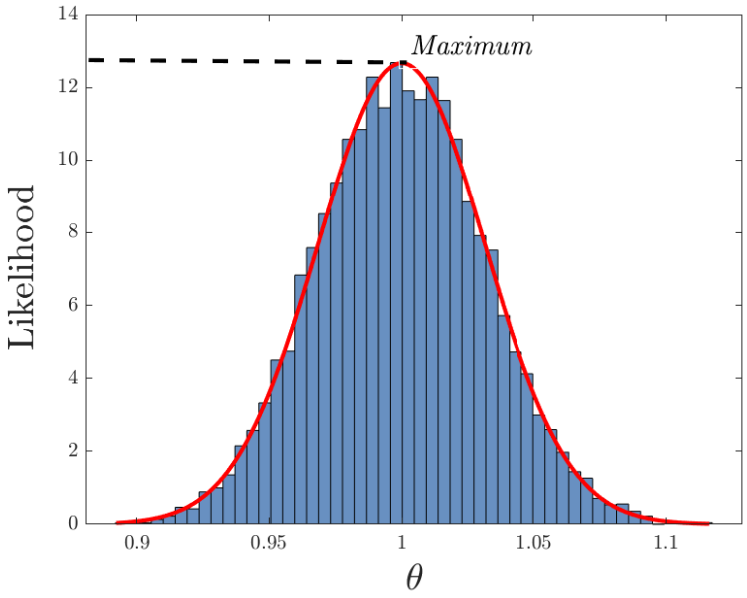

Figure 2.13: Example of likelihood function of unknown parameter θ from observed data.

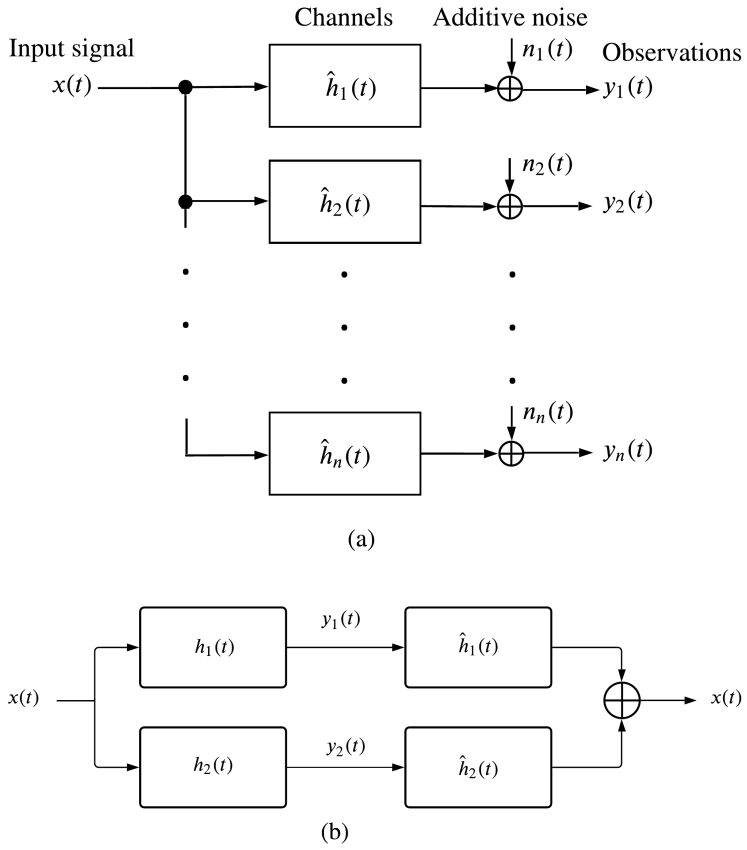

Figure 2.14: Diagram of a single-input multi-channel system: (a) a multichannel model and (b) a sub-channel matching algorithm.

2.6.4 Machine learning and artificial neural networks

In the last two decades, machine learning and artificial neural networks (ANN) are applied in many areas of researches and applications. Particularly, in room acoustic parameters estimations, techniques based on machine learning were successfully employed to estimate T_{60}, EDT, C_{50}, C_{80}, and STI [72, 81, 86–100]. The technique based on machine learning and artificial neural networks for estimating room acoustic parameters can be categorized as depicted in Fig. 2.15. On one hand, some meaningful acoustic features are extracted before training and estimating from neural networks. On the other hand, modern neural networks, such as convolutional neural networks (CNN) and long short-term memory networks (LSTM) directly map a speech signal (either in the time domain or time-frequency domain) to the target parameter. The latter approach is called an end-to-end approach.

In an early state, a multi-layer perceptron was used to estimate the STI [86, 101]. It is a kind of shallow neural network (i.e., one input, one output, and one hidden layer) to avoid a vanishing gradient problem [102]. The vanishing gradient is the situation that deeper layers do not learn anything. The learning means updating parameters or weights of the network. Those weights are calculated by backpropagation algorithm. It calculates the difference or gradient between the targets and its estimated results for each node. A stochastic gradient descent cannot return the gradient back since the value is smaller and vanished before update the weight of the higher layers. Besides the shallow network, the dimensions of the input or a number of features are also limited. Then, a machine learning technique for reducing dimensions, namely principal component analysis (PCA), was employed. A noisy reverberant speech signal is used for calculating its envelope spectrum, as shown in Fig. 2.2 (b). The output of the PCA is only 14 values to represent the whole envelope spectrum. Finally, the multi-layer perceptron that was trained from the convolution of clean speech signals and RIRs with additive noise could estimate the STI from those PCA features [88].

Since the emerging of deep learning and parallel computing, many deep neural networks (DNNs) have been proposed for estimating acoustic parameters. An insight algorithm of DNNs is somewhat different from the original concept that tried to mimic a network of a biological brain. Mathematically, an algorithm, which is updates the parameters of a network, can provide the convergence algorithm. Consequently, the global minimum is exist so that impressive results could be obtained. A convolutional neural network (CNN), which is trained from massive reverberant speech signals, could estimate the STI efficiently without feature extraction, known as end-to-end model [91].

For T_{60} estimation, many approaches have been evaluated in the Acoustic Characterization of Environments (ACE) challenge [72]. For example, Gamper and Tashev proposed a CNN with spectra-temporal features in the time-frequency domain [96]. Recently, a combination of a CNN and long short-term memory (LSTM) network has been proposed [92]. These two works showed that a method using either CNN or CNN with LSTM is computationally efficient, and only a few seconds of the observed speech signal is sufficient for reasonable accuracy. Parada *et al.* proposed bidirectional LSTM to estimate clarity index at 50 ms (C_{50}) using a spectral envelope in the modulation-domain as an input [93, 94]. 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 [99].

Interestingly, only a few of the current methods could estimate more than one parameter. Xiong *et al.* proposed a classification method rather than regression as usual to

estimate T_{60} and early-to-late reverberation ratio [97]. Similarly, in [103], T_{60} and DRR were blindly estimated by using sub-band speech decomposition. Decomposition speech signal into sub-bands before further processing has been proved that it can improve the accuracy of the estimation. A recent work proposed by Looney *et al.* provided three outputs that are T_{60}, DDR, and SNR [104]. Also, in [105], Mel-frequency Cepstral coefficients (MFCC) from a mixture of speech and music data were used as a feature so that a combined architecture of CNN and one of RNNs, namely gated recurrent unit, could map to T_{60}, C_{50}, and DRR.

2.6.5 SNR estimation

Background noise level or signal-to-noise ratio (SNR) is regarded as a significant indicator for more complete evaluation of the acoustical quality of rooms. The SNR is also an important information for speech enhancement algorithms in which they can enhance the speech before presenting it to compensate for degradation.

In measurement of room-acoustic parameters, as in ISO-3382, the requirement for the reverberation time is at least 35 dB SNR. Consequently, the estimated results are prone to be inaccurate due to the high level of background noise. However, none of the current approaches yield SNR along with estimating room acoustic parameters of interest to the best of our knowledge. Even though the CNN framework reported in [104] shows the SNR as one of the outputs, only the result of estimated T_{60} was evaluated. Hence, background noise level is unclear whether or not it can be accurately estimated along with the acoustical parameters [73] [106]. As in common places, background noise is inevitable, the SNR is also estimated in this research.

Noise might be classified into two sub-classes: stationary noise and non-stationary noise. Stationary noise means that its amplitude and spectrum remain almost constant over time, e.g., fan noise and air duct, whereas non-station noise is not, e.g., multiple people talking called babble noise [59]. This study uses four noise types from NOISEX-92 [107]. The NOISEX-92 corpus is used in speech recognition systems and consists of stationary and non-stationary noises.

The SNR is the energy ratio between speech signal and noise. The problem is how to discriminate between speech and non-speech from a mixture signal. Identifying portions of speech and non-speech from mixed signals is a straightforward strategy. This technique is called voice activity detection (VAD). The estimating STI based on the MTF concept, as the above mention [1], is also used VAD in the algorithm [78]. As there is a variety of noise, and it affects the signal in each frequency differently. Detecting voice and none voice from sub-band processing make much more accurate than with the global full-band method. Thus, Morita *et al.* proposed the estimating SNR using sub-band VAD [108], as shown in Fig. 2.16.

2.7 Summary

This chapter introduced the background of room acoustics. The current blind methods for estimating room-acoustic parameters were reviewed. The MTF-based methods, utilizing the relationship of modulation spectrum and the basis of the MTF, provided the correct results. These intuitive methods rely on the physic representation of reverberation and noise affecting the power envelope of the signal. On the other hand, many methods using machine learning and DNNs can obtain impressive results. However, these DNNs

approaches have limitations in estimation only one or two of room-acoustic parameters. They might be re-trained for a new target, but the re-training process with a new dataset is expensive. To this end, a more comprehensive method will be investigated in the following three chapters.

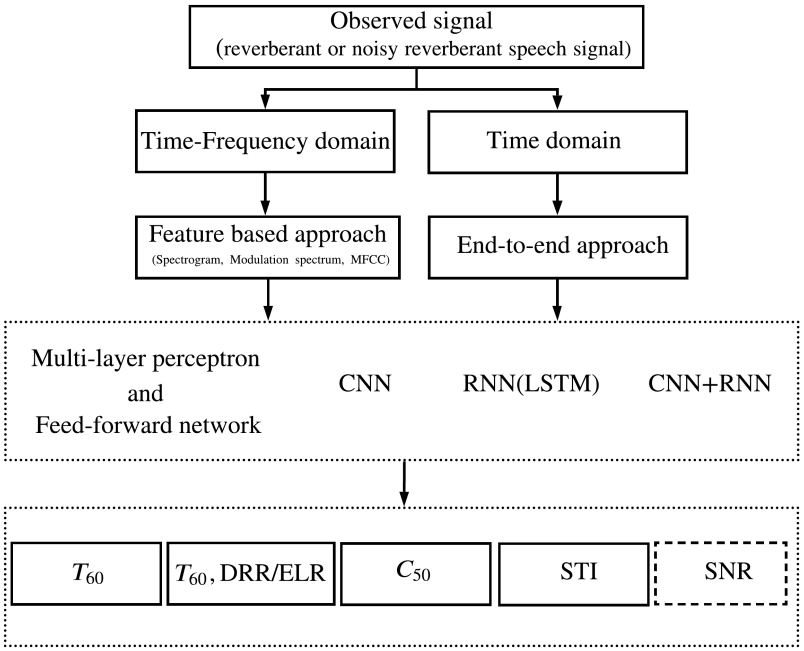

Figure 2.15: Estimating room acoustic parameter based on artificial neural networks.

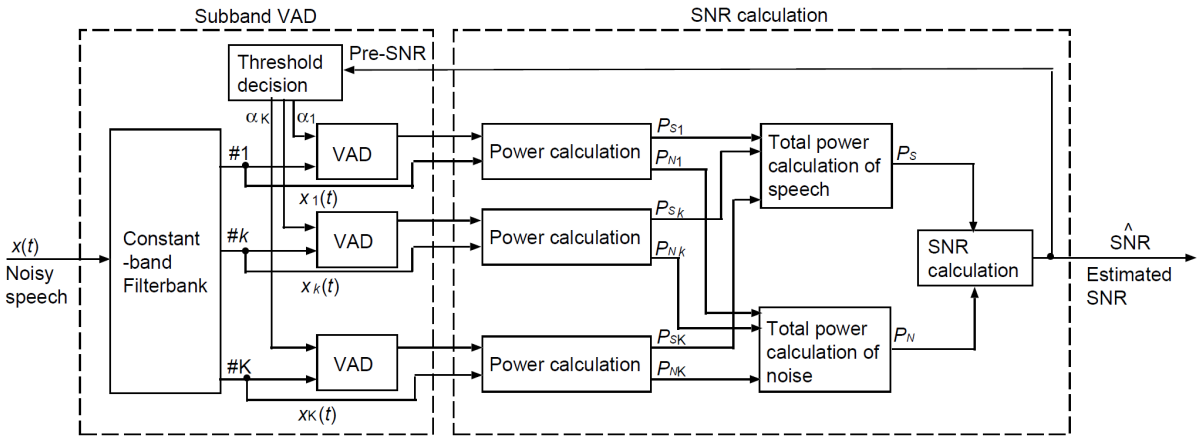

Figure 2.16: Framework of estimating SNR using sub-band VAD (Morita *et al.*).

Chapter 3

Blind estimation of speech transmission index

This chapter introduces a preliminary study for blindly estimating one of the room-acoustic indices, i.e., the speech transmission index (STI). The previous chapter shows that the STI is calculated from the measured MTFs of a given room. Hence, a basis of the MTF is investigated whether or not by using some features based on the MTF with a nonlinear regression technique can be used for estimating the STI. Two features based on the MTF are studied. A temporal amplitude envelope (TAE) of a reverberant signal is firstly used in this chapter, while a power envelope of a signal will be used later.

3.1 Temporal amplitude envelope of speech and the MTF

A speech signal can be regarded as a combination of a temporal amplitude envelope (TAE) with temporal fine structure. From the previous chapter, a TAE of a speech signal has been studied in which it plays an important role in speech intelligibility [39]. Hence, the TAE of a speech signal in noisy reverberant environments is a crucial feature for estimating the STI.

Figure 3.1 shows an example of a speech signal when it passes through a reverberant and noisy reverberant environment. The modulation depth (or modulation index, m, in the modulation frequency domain) of an original speech signal is reduced, corresponding to the degree of reverberation and background noise. This basis can be represented in the characteristics of TAE, e_y, and power envelope, $e^2(t)$.

From an observed signal, the TAE can be extracted by applying Hilbert transform $\mathrm{H}(\cdot)$ and a lowpass filter (LPF). Since, the modulation frequency between 4 to 16 Hz has highly contribution for speech communication both linguistic and non-linguistic, these region is then considered in designed the lowpass filter. The TAE of reverberant signal is obtained by applying the following equation.

$$e_y(t) = \mathrm{LPF}\left[|y(t) + j\mathrm{Hilbert}(y(t))|\right]. \tag{3.1}$$

As both TAE and power envelope contained the modulation distortion information of a given transmission channel, they are considered to be a feature in the core structure of this study. To clear understand how close they are, the TAE and power envelope extracted from the same reverberated speech signal are then shown in Fig. 3.2.

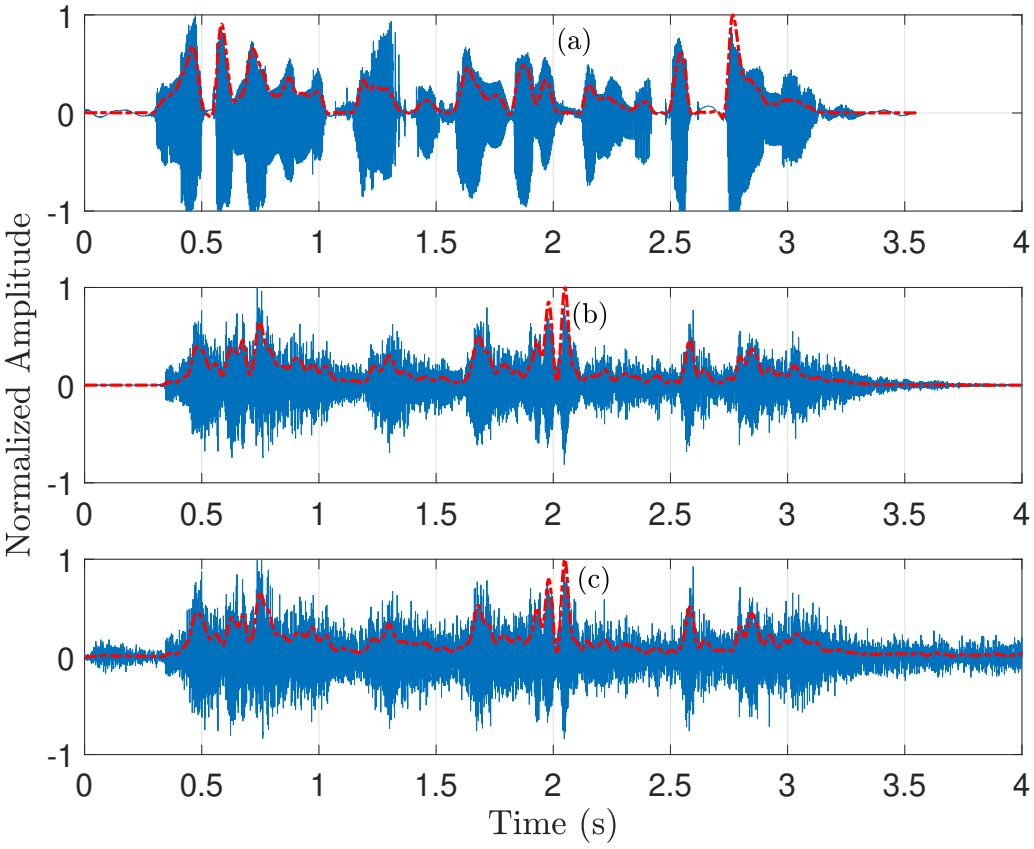

Figure 3.1: Example signals of (a) clean speech, (b) reverberant speech ($T_{60} = 0.43$ s), and (c) noisy reverberant speech (babble noise, 5 dB SNR) and $T_{60} = 0.43$ s). Dashed lines are power envelopes of each signal

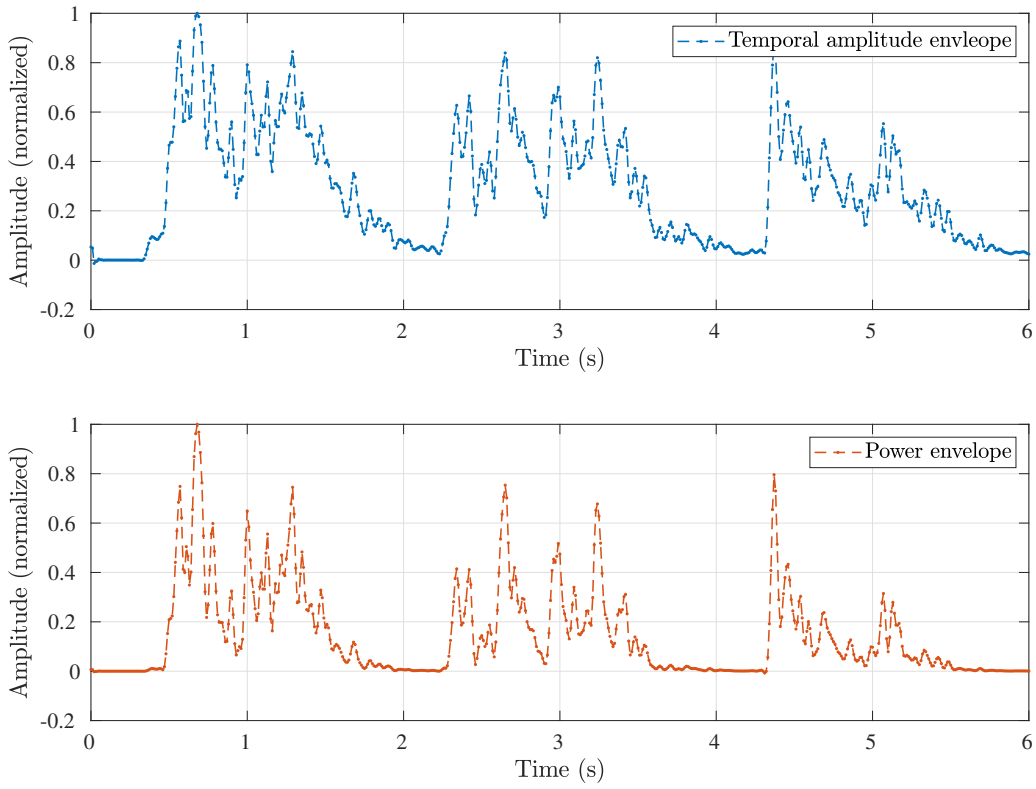

Figure 3.2: Comparison between TAE and power envelope extracted from the same reverberated speech signal.

3.2 Core structure of blind STI estimation

Mapping characteristics of the TAE of a speech signal under noisy reverberant conditions to STI is the core structure of this study. The studies proposed by Unoki *et al.* have revealed that the characteristics of the MTF and its properties are successful for estimating the reverberation time and STI [1, 27, 32, 109, 110] . However, there is some remaining error that might be caused by the mismatch between the stochastic models and real environments. Therefore, the useful concept of the MTF is then incorporated into a nonlinear regression trained from big data. As a result, a neural network trained from data in various reverberation times and noise is introduced to overcome this issue.

The convolution operation in the time domain is the main idea in designing the neural network. Thus, the convolutional neural network (CNN) is employed. A number of RIRs from synthetic and measured RIRs with different reverberation times as well as background noise were exploited. Based on an assumption of solving a regression problem, the estimated STI is mapped from the distortion of the TAE. The CNN is deployed to map the characteristics of the TAE to the estimated STI. The robust STI estimation scheme is shown in Fig. 3.3.

3.3 Implementation and evaluation

The CNN trained with observed reverberant envelopes at various noise types and levels can determine the associated STIs. We assume a blind STI estimation as a blind deconvolution with a regression problem. As a blind deconvolution problem, this CNN performs the deconvolution operation of the observed envelope and solves the regression problem. The CNN consists of three convolutional layers and complementary layers.

In design a reasonable CNN, a convolution operation in the time domain of an envelope signal is represented by one-dimensional convolution in the first layer. Another one-dimensional convolution is applied again to construct a new two-dimensional data inspired by the deep CNN [91]. The last two convolutional layers apply the two-dimensional convolutional filters to perform a regression task. In a mathematical viewpoint, high-dimensional spaces expand a possibility for problem-solving. Similarly, in neuroscience, the middle layer of the perceptron model has more number of neurons than the other layers. Thus, the middle layer is assigned with more number of filters. The numbers of filters in these layers are assigned to be 32 and 16, respectively For complementary layers, a pooling unit accompanies a convolution layer for down-sampling and reducing variation of the input. Here, max pooling, which is non-linear operation, corrects the highest value from their neighbors. The outputs are then passed through an activation function, which is a rectified linear unit (ReLU). The ReLU function has been proposed to deal with a vanishing gradient problem, which behaves as a half-wave rectifier according to $f(x) = \max(x, 0)$. The ReLU output is 0 when input $x < 0$, and is a linear function when $x \geq 0$. We also employ a batch-normalization to scale the values to be a unit norm. A regularization technique called dropout is set with a probability of 0.2 to avoid an over-fitting and memorizing problem. A flattening layer or fully connected layer is an operator that converts a two-dimensional array into a vector. The last layer, named dense layer, estimates output by a sum of products between the vectors and its weights, so that

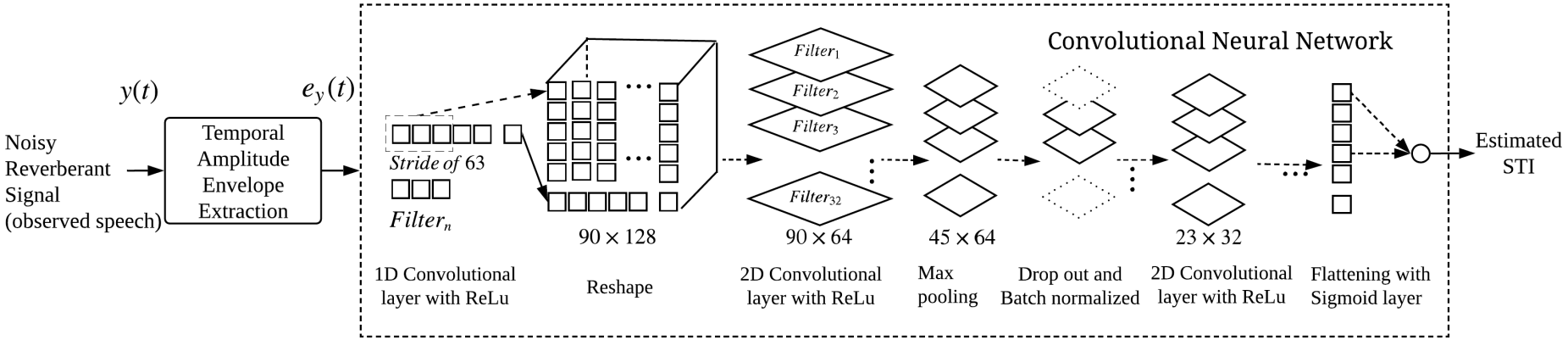

Figure 3.3: Block diagram of the proposed method for estimating STI in noisy reverberant environments using TAE and CNN.

Table 3.1: Network architecture of the robust STI estimator.

No.	Layer Type	Parameters
1	Input	Input shape = 374×1
2	Convolution^{1st}	128 filters, filter size = 128×1, ReLU
3	Pooling	Max pooling, size $= 2$, stride $= 1$
4	Convolution^{2nd}	128 filters, filter size = 5×1, ReLU
5	Reshape	filter size = 128×21
6	Convolution^{3rd}	32 filters, filter size = 90×64, ReLU
7	Pooling	Max pool, size $= 2$, stride $= 1$
8	Batch Normalization	-
9	Dropout	0.2
10	Convolution^{4th}	16 filters, filter size = 23×32, ReLU
11	Fully Connected	Sigmoid
12	Regression Output	Mean-square-error

the estimated STI as the output can be presented as

$$\hat{\text{STI}} = \text{SIGM}\left(\sum_{i=1}^{j} W \otimes a_i + b\right), \tag{3.2}$$

where $\hat{\text{STI}}$ is an estimated STI, SIGM is a sigmoid function, W is weight matrix, a_i is an input from a previous layer for i to the total elements j, "$\otimes$" is element-wise operation, and b is bias. The RMSprop is an optimization algorithm to minimize the cost function, which is mean square error (MSE), and the optimizer is set a learning rate of 0.001. These tunable filters are updated along with the training process. The CNN architecture is detailed in Table. 3.1.

3.3.1 Experimental setup

There are three groups of data used along with this study: (1) room impulse response, (2) speech signals, and (3) noise types and conditions. First, the RIR signals can be separated into two groups that are measured RIRs and synthesized RIRs. The main dataset is the 43 measured RIRs from SMILEdataset [111]. There are also some sharing dataset, such as seven RIRs in ACEdataset [72] and three RIRs in the REVERB challenge [112], and two RIRs from the Architectural Institute of Japan [64]. Second, speech signals are from two databases. One is Japanese utterances from the ATR database [113]. They consist of long sentences, uttered by ten speakers (five male and five female). The other is English utterances from the VCTK corpus [114]. They consist of one hundred English speakers that are randomly selected from various ages, genders, accents, and regions. The third group is the noise data. Noise data are from the NOISEX-92 [107]. Four noise types were used, including white, pink, factory, and bubble noise. The levels of noise conditions were SNRs of 20, 15, 10, and 5 dB.

As measuring RIR needs specific procedures and equipment, as mentioned in Chapter 2, the measured RIRs from real environments are limited, in particular for DNN approach as the proposed methods. Therefore, a data augmentation technique was used to generate a sufficient amount of training set. There are a few data augmentation methods to generate simulated RIRs [115, 116]. In recent work, some room acoustic parameters are

used to synthesis RIR. The parameters are T_{60}, EDT, and DRR as well as one parameter of the RIR called the initial time delay gap, which is similar to parameter T_0 of the extended RIR model [116]. However, in the work, one conventional method based on geometrical acoustics, known as an image-source method, and the two RIR models were used [117, 118].

For the first set of RIR, an image-source method was used. The image-source method can generate different RIR from its setup. The setup parameters include positions of a sound source and a receiver, reverberation time, and absorption coefficient. One hundred RIRs were then generated by varying such parameters of the image-source method. The second and the third dataset of simulated RIRs were synthesized based on Schroeder's RIR model and the extended RIR model. According to Schroeder's RIR model, the reverberation time, T_{60}, as in Eq. (2.5), 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. Those RIRs were then convoluted with speech signals. A total of $29,000$ reverberant speech signals were generated.

The reverberant and noisy reverberant speech signals were the results of the convolution between the RIR and speech signal and with the additive noise for the latter.

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 be fit well with a simple exponential decay. This means that such RIRs can be represented by Schroeder's model. Nevertheless, the mismatch issue was found for 14 RIRs, as shown in Fig. 5.3. 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 can be used for the proposed method based on the extended RIR. All signals had a five-second period, a sampling rate of 16 kHz, 32-bit quantization, and one channel.

3.3.2 Evaluation matrices

As estimating STI is a regression problem, the two metrics are used to evaluate the performance of the proposed method: root-mean-square error RMSE and Pearson's correlation coefficient . The low RMSE and high correlation ρ indicate the high performance of the STI estimator. Note that RMSE is the square root of MSE, as used in the optimization of the filters of the CNN, to make the scale of the estimation error to be the same as the scale of STI. RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\hat{\mathrm{STI}}_n - \mathrm{STI}_n\right)^2}, \tag{3.3}$$

where $\hat{\mathrm{STI}}_n$ is the estimated STI, $\mathrm{STI_n}$ is the ground truth calculated from RIR and SNR as in Eq. (2.17), n is an index of observed signal, and N is a total number of the signals. The second evaluation metric, i.e., correlation (ρ), is defined as

$$\rho = \frac{\sum_{n=1}^{N}(\hat{\mathrm{STI}}_n - \overline{\hat{\mathrm{STI}}_n})^2(\mathrm{STI}_n - \overline{\mathrm{STI}}_n)^2}{\sqrt{\sum_{n=1}^{N}(\hat{\mathrm{STI}}_n - \overline{\hat{\mathrm{STI}}_n})^2\sum_{n=1}^{N}(\mathrm{STI}_n - \overline{\mathrm{STI}}_n)^2}}, \tag{3.4}$$

Table 3.2: Estimated STIs in various conditions from SMILE corpus in the metrics of RMSE and correlation (ρ) from SMILEdataset.

Noise	Method	RMSE 20 dB	RMSE 5 dB	ρ
White	MTF-based	0.25	0.33	0.72
	deep CNN	0.09	0.10	0.89
	Proposed	0.07	0.09	0.90
Pink	MTF-based	0.20	0.23	0.71
	deep CNN	0.13	0.14	0.89
	Proposed	0.08	0.14	0.85
Babble	MTF-based	0.29	0.12	0.64
	deep CNN	0.12	0.11	0.90
	Proposed	0.11	0.12	0.92
Factory	MTF-based	0.37	0.11	0.74
	deep CNN	0.11	0.15	0.89
	Proposed	0.13	0.18	0.82

where $\overline{\hat{\mathrm{STI}}_n}$ is the average of $\hat{\mathrm{STI}}_n$, and $\overline{\mathrm{STI}_n}$ is the average of $\mathrm{STI_n}$.

3.4 Results and discussion

Figures 3.4–3.7 show the estimated STIs from observed speech signals in reverberant environments with four types of background noise.

As the MTF concept assumes noise as white Gaussian noise (WGN), noise in real environment is, however, different from the model. For example, babble noise and factory noise that are non-stationary yield some mismatch and error in the existing MTF-based estimation. On the other hand, the proposed method that uses the CNN learned from various noise types can overcome this problem, and the accuracy of the estimation in such background noise and reverberation environments can be maintained. However, estimating STIs from observed speech signals with factory noise is still challenging because some inconsistencies of the estimated results remain.

The proposed not only satisfies the accuracy and robustness, but also has other advantages in additional aspects. First, the proposed model can reduce the operation time from the conventional STI measurement time of 15 minutes [16]. The proposed method that takes only a short four-second speech segment can provide accuracy comparable to that of the conventional method [21]. Hence, the operation time is reduced by 180 times. Second, the proposed model significantly reduces the computational time: it is $4,666$ times faster than the MTF-based because it does not need to search for the optimal parameters. The optimal filters of the CNN can calculate STIs promptly.

Although the proposed method can successfully estimate the STI, there are two concerning issues from this work that should be pointed out. First, since the machine learning methods are based on several hyper-parameters, there are enormous possibilities for designing network architecture, and this makes it difficult to reach an optimal solution. A robust estimation model should be generalized enough for dealing with random data by a large margin. The generalized model needs to compensate for the trade-off between

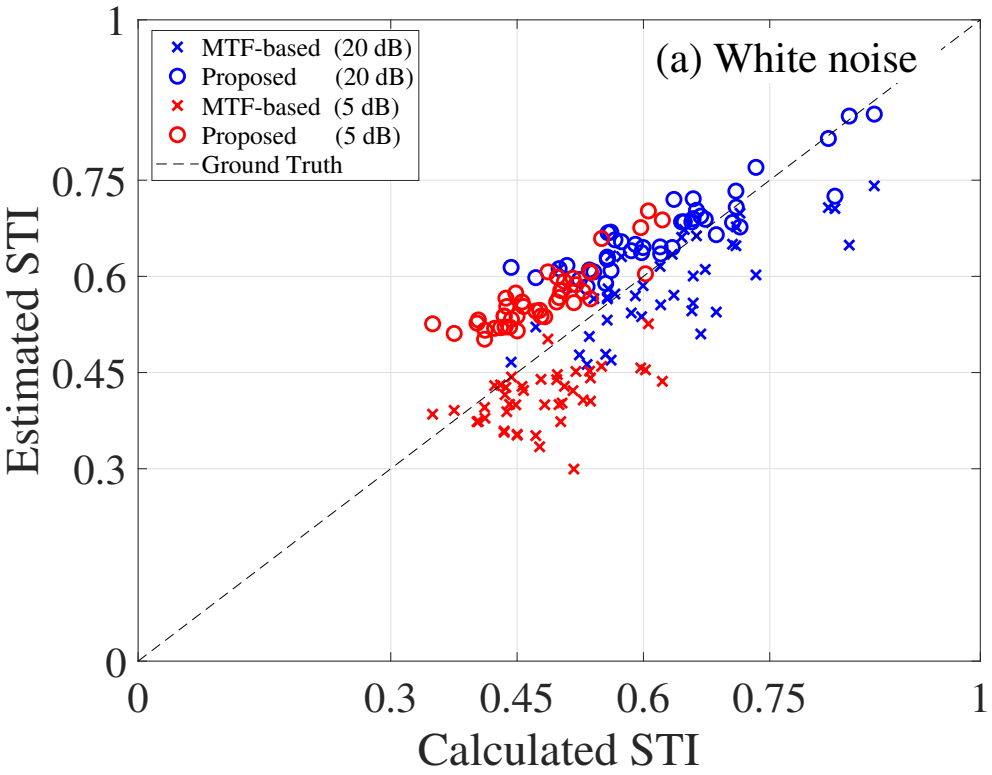

Figure 3.4: Estimated STIs from observed speech signals in reverberant environment with white noise.

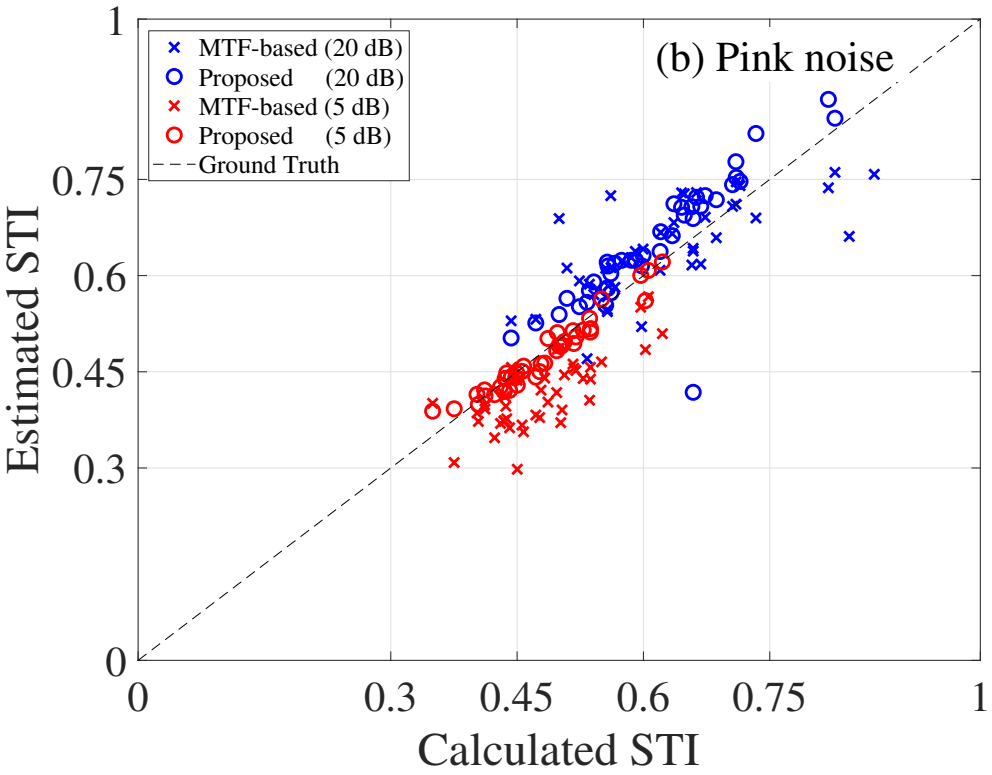

Figure 3.5: Estimated STIs from observed speech signals in reverberant environment with pink noise.

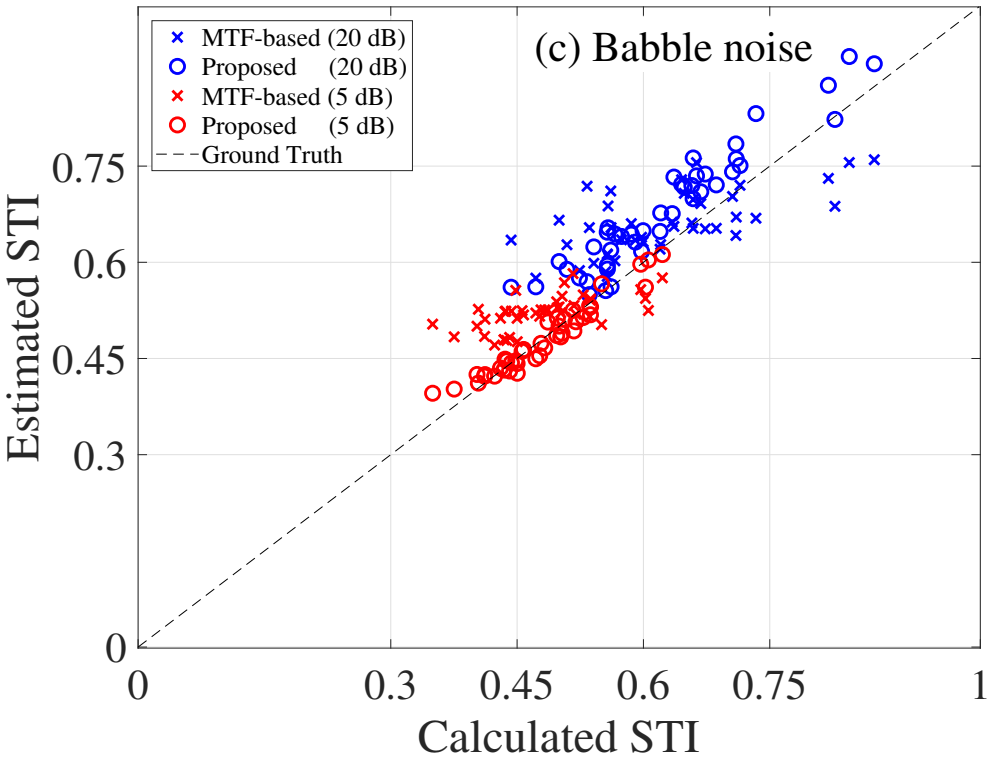

Figure 3.6: Estimated STIs from observed speech signals in reverberant environment with babble noise.

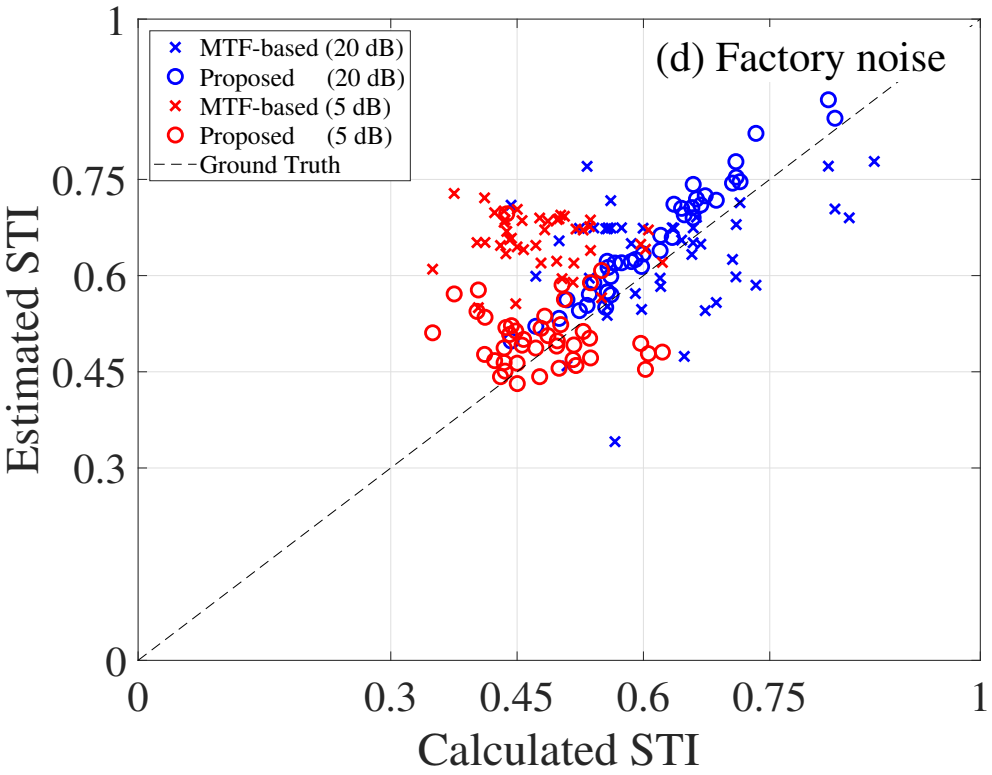

Figure 3.7: Estimated STIs from observed speech signals in reverberant environment with factory noise.

Table 3.3: Estimated STIs of speech signals in RIR and background noise from dataset in the Acoustic Characteristic Environment challenge.

Noise	Method	RMSE		ρ
		20 dB	5 dB	
Ambient	MTF-based	0.14	0.35	0.64
	deep CNN	0.07	0.09	0.87
	Proposed	0.07	0.11	0.86
Fan	MTF-based	0.17	0.18	0.73
	deep CNN	0.07	0.08	0.76
	Proposed	0.08	0.09	0.83
Babble	MTF-based	0.18	0.26	0.63
	deep CNN	0.09	0.08	0.89
	Proposed	0.13	0.15	0.86

high accuracy and model complexity. For example, we found that the more extended envelope input the CNN takes (i.e., from one second to four seconds), the more accurate the performance of the CNN. Thus, in this study, while we empirically propose the CNN architecture to maintain the acceptable performance, it should be fine-tuned so as to deliver an even better performance.

3.5 Summary

This chapter introduced a preliminary study that the proposed method can estimate the STI in noisy reverberant environments. The results suggest that the proposed method using CNN with the TAE extracted from an observed signal on the basis of the MTF provided a robust estimator. Unfortunately, there are many room acoustic parameters that can be beneficial and needed for difference purposes. This limitation of training the model for only a single parameter will be considered and improved in the next chapter.

Chapter 4

Blindly estimating parameter of RIR model

This chapter introduces a more worthwhile estimation method that the unknown RIR is estimated. Instead of modeling, training, and estimating a single room acoustic parameter one by one, as the previous STI estimation, the unknown RIR is modeled and estimated. Schroeder's RIR is used for representing the unknown RIR. The proposed method intends to estimate a parameter of the RIR model. The basis of the MTF concept is exploited for sub-octave bands.

The main concept using Schroeder's RIR is introduced. The nonlinear function approximation using CNN is developed. The TAE feature with CNN is so-called the MTF-based CNN framework. Later, the method based on the extended RIR model is studied. Thus, five-room acoustic parameters and STI can be derived. The five room-acoustic parameters include T_{60}, EDT, C_{80}, D_{50}, and T_s.

We propose a scheme for estimating five room-acoustic parameters and an STI, namely MTF-based CNNs, as shown in Fig. 4.1. The scheme incorporates the MTF concept into a nonlinear regression using CNNs. The T_{60}s for sub-bands are mapped in accordance with the characteristics of the TAEs under reverberant conditions. RIR is approximated from the estimated T_{60}s to derive the five parameters and STI.

4.1 Sub-band analysis

The sub-band analysis for estimating room acoustic parameters is derived from the STI algorithm, which is from the basis of the MTFs in seven-octave bands. Thus, we exploit the relation between the MTF and RIR, as shown in Eq. (2.17), within the same bands as the STI. The bands have center frequencies ranging from 125 Hz to 8 kHz. The normalized reverberant-speech signal is the input. The signal is then decomposed to each sub-band using octave-band filters.

Based on the MTF concept, a temporal envelope of any signal is a smoothed version of the original signal when it is passed through a reverberant space [17]. We then utilize the seven TAEs to represent the modulation distortion characteristics caused by reverberation in the bands. The reverberation, in terms of the T_{60}s, attenuates the observed TAEs. The seven TAEs account for the accuracy enhancement of the estimating T_{60} and STI as well as the other parameters.

The TAE in each band is extracted according to Eq. (4.1). The observed signal is decomposed by using the Hilbert transform $\mathrm{H}(\cdot)$ and a lowpass filter (LPF). The LPF is

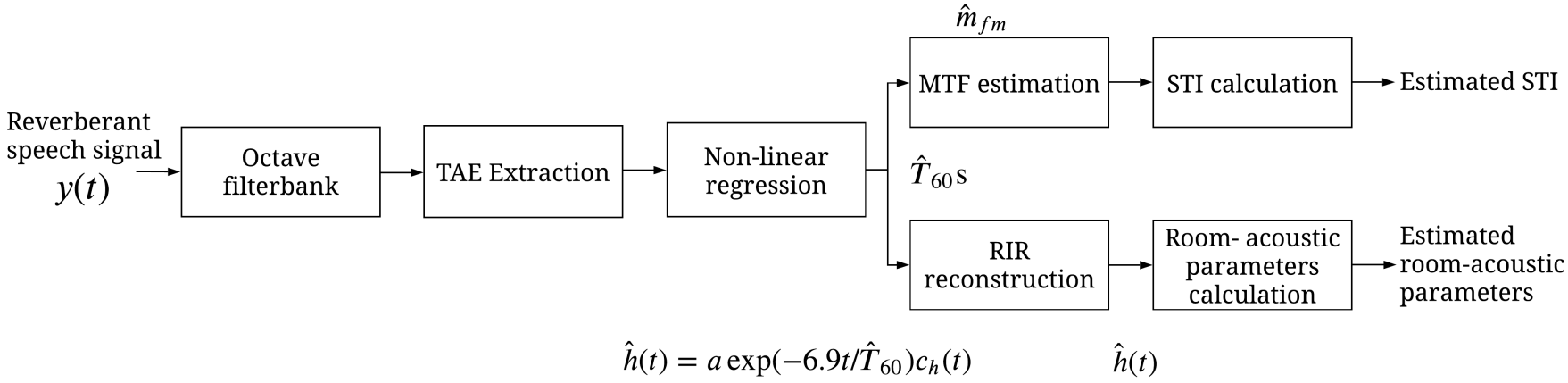

Figure 4.1: A conceptual diagram of estimating parameter of the RIR model and deriving the room acoustic parameters.

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 T_{60}s for the seven-octave bands by using CNNs.

$$e_y(t) = \text{LPF}\left[|y(t) + j\text{Hilbert}(y(t))|\right]. \tag{4.1}$$

4.2 MTF-based CNN framework

We deploy one-dimensional CNNs for mapping the characteristics of the TAEs with their associated T_{60}s in each octave band. Each model consists of four convolutional layers. The input layer takes a TAE to be convoluted with the filters. The regulated linear unit (ReLU), $f(x) = \max(x, 0)$, performs nonlinear activation in all convolutional layer. Batch normalization is applied after the first convolution. Max pooling is also used for reducing the dimensions before the next layer. The dropout rate before the last layer is set to 20%. The fully connected layer is the output layer. The seven CNNs are trained from the TAEs/T_{60}s pairs. The trained models are supervised by the T_{60}s ground-truths. The ground-truths are calculated from simulated RIRs. The output of each CNN for each sub-band is the estimated T_{60}. The details of the MTF-based CNN model is shown in Table. 4.1.

Table 4.1: Network architecture of the 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	max pooling, size $= 2$, stride $= 1$
4	Conv1D^{2nd}	16 filters, filter size $= 5 \times 1$, ReLU
5	Pooling	max pooling, size $= 2$, stride $= 1$
6	Dropout	0.2
7	Conv1D^{3rd}	8 filters, filter size $= 5 \times 1$, ReLU
8	Pooling	max pooling, size $= 2$,
9	Conv1D^{4th}	4 filters, filter size $= 5 \times 1$, ReLU
10	Fully Connected	1 output (i.e., T_{60}), ReLU
11	Regression Output	mean-square-error (MSE)

4.3 RIR approximation

The estimated T_{60}s are used to approximate RIR, $\hat{\mathbf{h}}(t)$ according to Schroeder's RIR model. As Schroeder's RIR depends on only the reverberation time, the estimated T_{60} for each octave-band is used to construct the temporal envelope of the RIR, $\hat{e_h}(t)$. The temporal envelope of each band is modulated with a band-limited Gaussian noise with a bandwidth of 1/3 of an octave. Then, the sub-band RIRs are then summed together. The approximated RIR can be expressed as

$$\hat{h}(t) = \sum_{k=1}^{K} \exp\left(-\frac{6.9t}{T_{60,k}}\right) c_{h,k}(t), \tag{4.2}$$

where $T_{60,k}$ is the estimated T_{60} in the k-th band and $K = 7$, and $c_{h,k}(t)$ is band-limited Gaussian noise. The STI can then be calculated from the estimated T_{60}s based on the basis of the MTF, as in Eq. (2.17). Also, the T_{60}, EDT, C_{80}, D_{50}, and T_s can be calculated according to the definitions, in Eqs. (2.10), (2.11), and (2.12), respectively.

4.4 Evaluations

A total of 29, 000 reverberant speech signals with a sampling rate of 16 kHz were generated from the simulated RIRs convoluted with speech signals. The simulated RIRs are based on Schroeder's RIR model. The reverberation time of the RIRs varies from 0.2 to 3.0 s with a step size of 0.1 s. Each envelope with a different T_{60} was modulated with a different random seed WGN carrier. There are a hundred different WGN carrier seeds. The speech signals were ten short (five-second) Japanese sentences uttered by five men and five women in [113]. These reverberant signals were separated into 70 % for training, and the rest for testing the model (simulated RIR).

The proposed method is then evaluated whether or not it can estimate the parameters and STI even though the acoustic characteristics might not follow Schroeder's RIR model by using reverberated speech signals in an unknown realistic environment. We utilized 43 measured RIRs from the SMILEdataset [111]. The RMSE and correlation coefficient were the metrics for indicating the accuracy of the estimation.

4.5 Results and discussion

Figures 4.2-4.6 show the estimated results of the estimated room-acoustic parameters and STI from speech signals in reverberant environments. The symbols "o" and "square" correspond to the estimated parameters in the simulated room and realistic room, respectively, where '*' is the value estimated using the previous methods. The horizontal axis indicates the parameter directly calculated from the RIRs, and the vertical axis indicates estimated values. It was found that the results from the simulated rooms were excellent in all parameters. On the other hand, in the real rooms, the results suggested that the proposed method can be used to estimate the five room-acoustic parameters and STI. However, none of the current methods can estimate these parameters simultaneously. We then directly compared only T_{60} and STI with the baseline methods proposed in [76, 98]. The others were discussed from the results compared with their ground-truths.

The results of the estimated T_{60} and STI show that the proposed method outperforms the previous methods since it provided significantly lower RMSEs. The estimated T_{60} was improved about 40%, and 25% for the STI compared with the previous methods, respectively. For C_{80}, D_{50}, and T_s, the RMSE were 1.66 dB, 11.85%, and 0.06, respectively. The results of the estimated C_{80} were close to the accuracy from the standard method from measured RIRs [3]. However, the results of the estimated D_{50} and T_s have remaining outliers. Those errors might be caused by a mismatch between the RIR model we used and the real RIRs.

Table 4.2 shows the correlation coefficients between the estimated parameters and ground-truths. The results show that the proposed method was successful in unseen simulated rooms since the correlation coefficients were close to 1. For the real rooms, the proposed method has high correlations in all parameters, but the estimated D_{50} and T_s were slightly low.

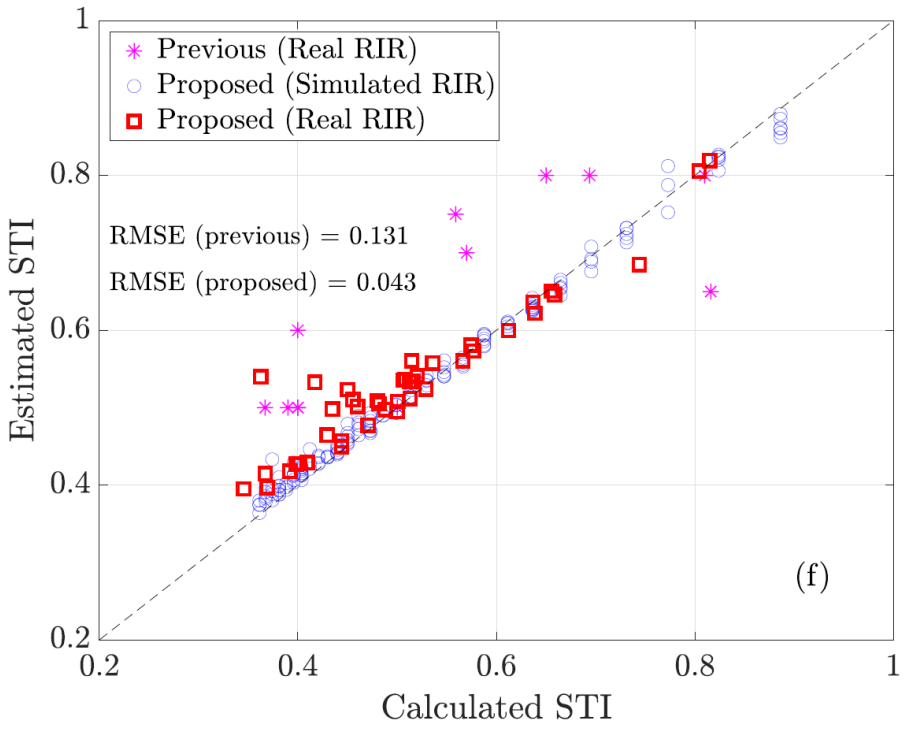

Figure 4.2: Estimated speech transmission index, STI, from reverberant speech signals. The symbol "o" corresponds to the estimated value from the simulated RIR, "square" indicates the estimated value from the measured RIR, "*" indicates the estimated result using the method proposed by Unoki *et al.* [1], and the dashed line represents the ground-truth calculated from the RIRs.

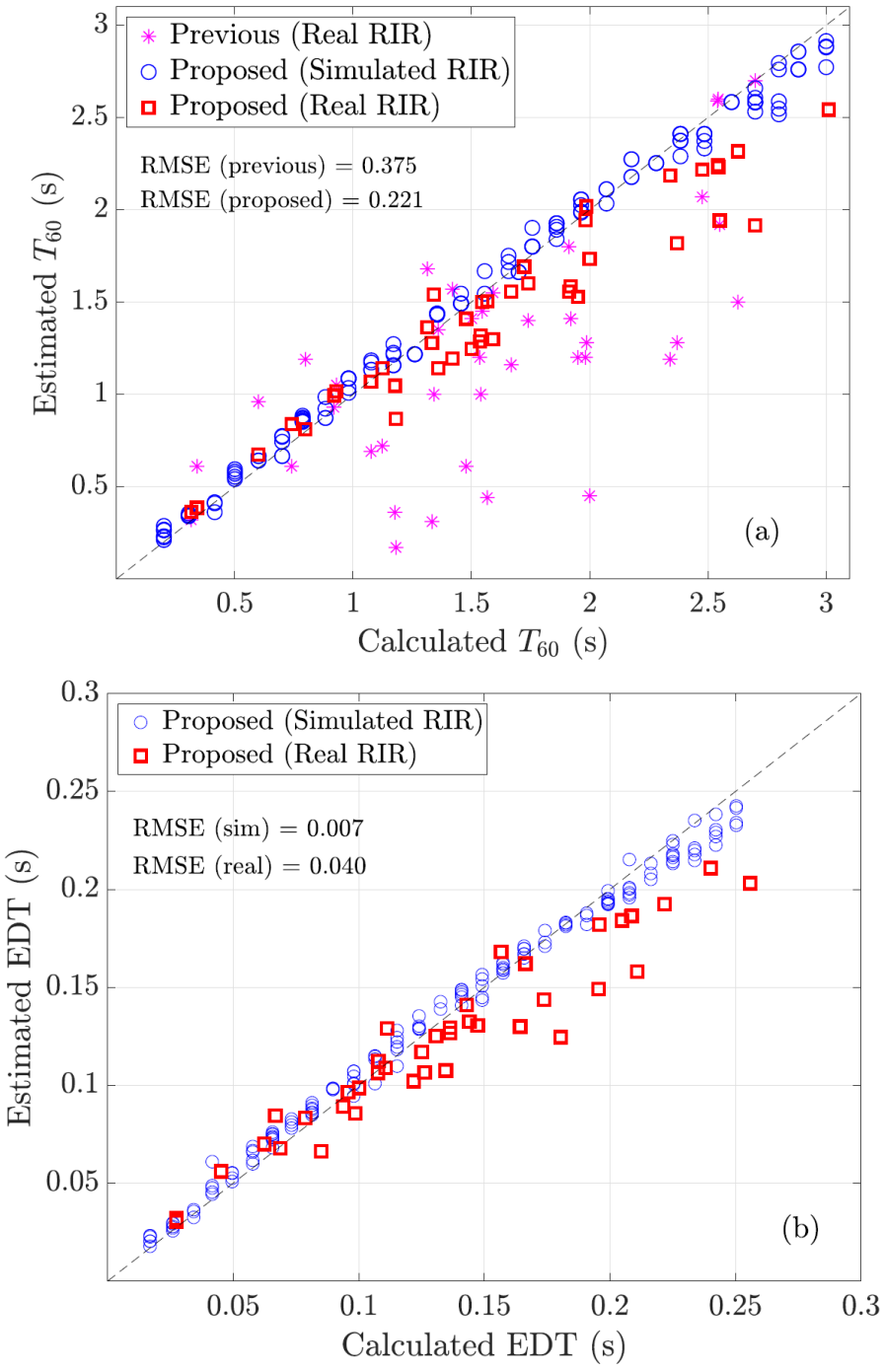

Figure 4.3: Estimated reverberation time (T_{60}: Previous method proposed by Unoki *et al.* [2]) and early decay time (EDT) from reverberant speech signals.

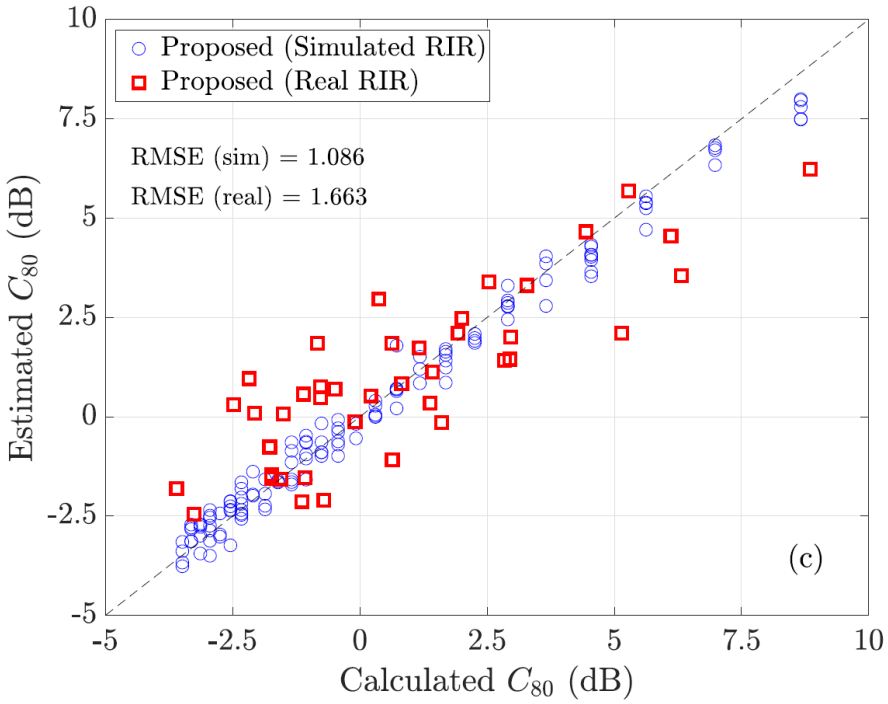

Figure 4.4: Estimated clarity index, C_{80}, from reverberant speech signals.

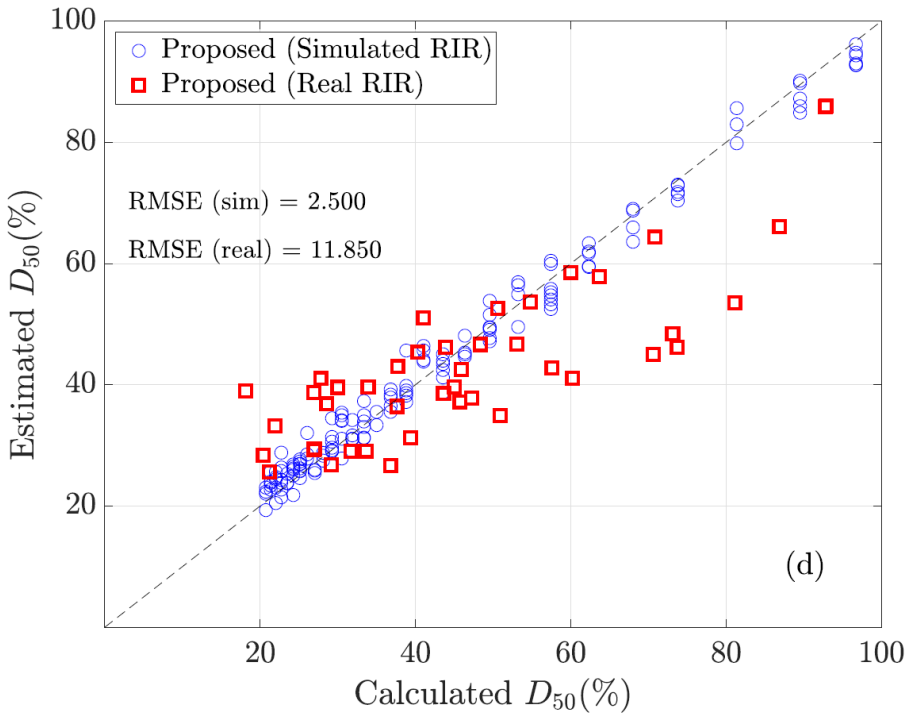

Figure 4.5: Estimated Deutlichkeit, D_{50}, from reverberant speech signals.

Table 4.2: Correlation coefficients between the estimated and calculated parameters.

	T_{60}	EDT	C_{80}	D_{50}	T_s	STI
Simulated rooms	0.996	0.996	0.992	0.994	0.996	0.997
Real rooms	0.915	0.870	0.918	0.818	0.822	0.902

4.6 Summary

In this chapter, a blind method for estimating a parameter of Schroeder's RIR model was proposed. Hence, the target room-acoustic parameters, i.e., T_{60}, EDT, C_{80}, D_{50}, and T_s as well as the STI can be derived. The proposed method leveraged the relationship between a stochastic RIR model and its MTF to estimate T_{60} for seven-octave bands. The proposed scheme estimated T_{60} from the temporal amplitude envelope of an observed signal in each band. The estimated T_{60}s were used to approximate the MTF and RIR for deriving of the room acoustic parameters and STI. Simulations were carried out to determine whether the proposed method could estimate the room acoustic parameters and STI from reverberated speech signals even if the RIRs were not the same as Schroeder's RIR model. The experimental results in terms of RMSEs and correlation coefficients showed that the proposed method yielded a better accuracy, compared with the baselines for the STI and T_{60}. Also, the estimated EDT, C_{80}, D_{50}, and T_s were also close to the standard methods.

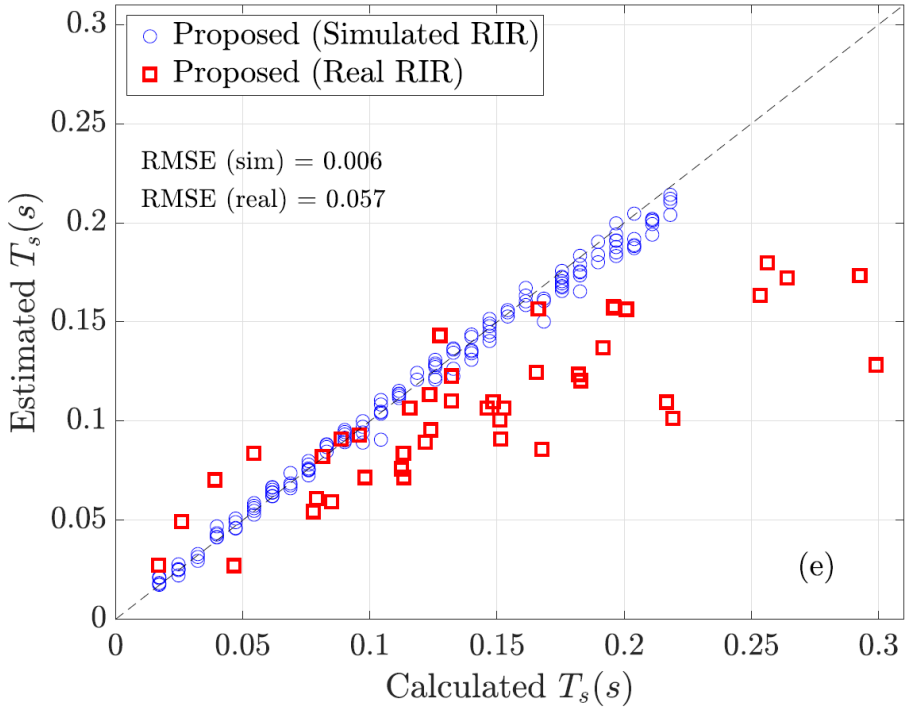

Figure 4.6: Estimated center time, T_s, from reverberant speech signals.

Chapter 5

Blindly estimating room acoustic parameters and STI based on the extended RIR model

In the previous chapter, we presented the estimating parameter of Schroeder's RIR so that multiple-room acoustic parameters and STI can be estimated by using reverberant speech signals. This chapter reports the results of the estimated acoustical parameters. Furthermore, the proposed scheme is modified such that the power envelope of the observed signal is used rather than the TAE. The simulations are conducted to verify the proposed method in both reverberant and noisy reverberant environments. In addition, the estimation under noise conditions is investigated. Since 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. 5.1. 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 [76, 77].

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. 5.4. The details of the proposed method are described as follows.

5.1 The 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

$$h(t) = h_{ext}(t - T_0), \qquad T_0 \geq 0 \tag{5.1}$$

$$h_{ext}(t) = \begin{cases} a\exp(6.9t/T_h)c_h(t), & t < 0 \\ a\exp(-6.9t/T_t)c_h(t), & t \geq 0 \end{cases} \tag{5.2}$$

where $h_{ext}(t)$ represents the extended RIR model. 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. (5.1), the time-shifting property is used to provide a causal system and stable impulse response, i.e., $h(t)=0,\ t<0$. In Eq. (5.2), the three parameters of the extended RIR model control the shape of the envelope of the RIR.

Figure 5.2 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. 5.2 (a), and Fig. 5.2 (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. Figure 5.3 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 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. (2.17), the complex MTF of the extended RIR model can be represented as

$$m(f_m, T_h, T_0, T_t) \;=\; \frac{\exp(-j2\pi f_m T_0)}{\sqrt{\left(1+(2\pi f_m(T_h/13.8))^2\right)\left(1+(2\pi f_m(T_t/13.8))^2\right)}}. \tag{5.3}$$

5.2 Core structure of the estimation

Previously, the MTF-based CNN framework has been proposed on the basis of Schroeder's RIR model. In this chapter, the proposed method is based on the previous scheme but estimate the three parameters of the extended RIR model. Instead of estimating only reverberation times, T_{60}s, the CNNs is used for mapping T_h, T_0, and T_t with the sub-band TAEs. 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.

As reverberant speech signal is decomposed into seven-octave bands, the seven identical CNN models for each band. Here, each CNN 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 5.1.

For reconstruct the RIR, this method replace replaces Eq. (4.2) of the original model with Eqs. (5.1) and (5.2) of the extended RIR model. Then, the remaining calculations are the same.

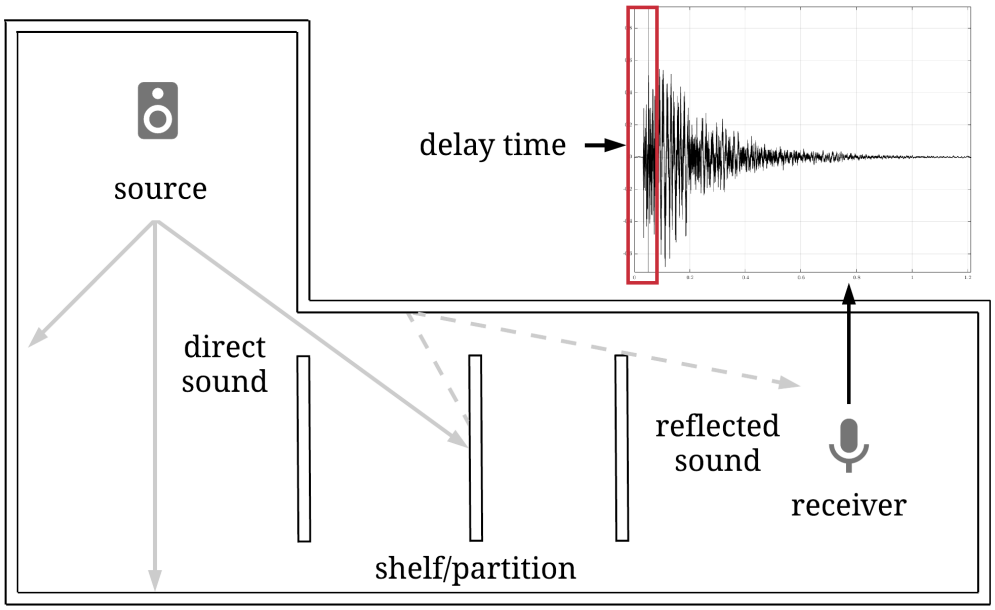

Figure 5.1: Example of complex space and its impulse response.

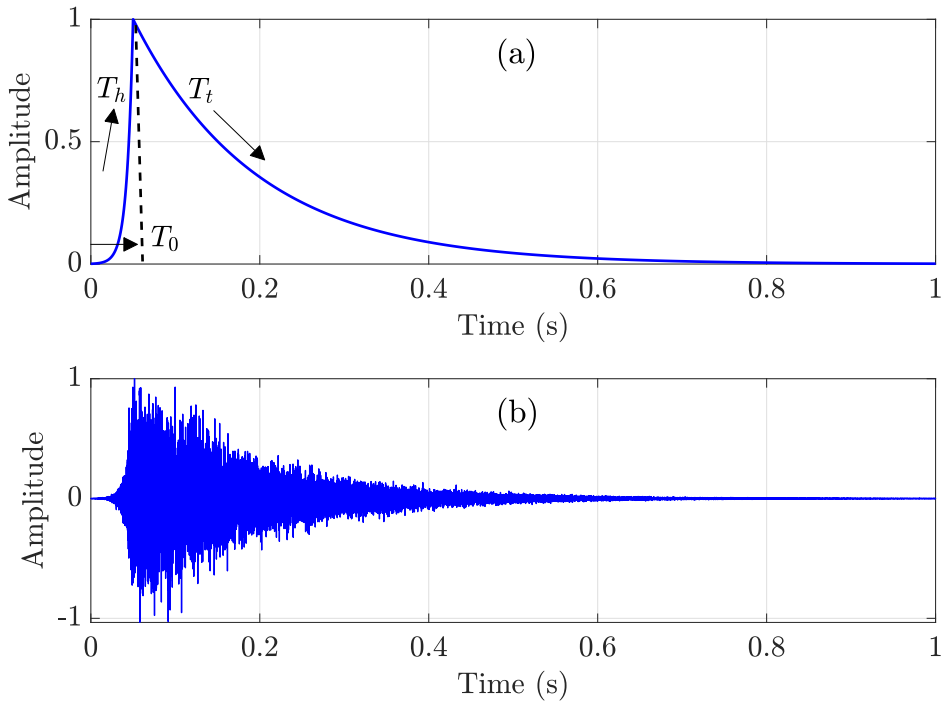

Figure 5.2: Example of extended RIR model where $T_h = 0.08$, $T_0 = 0.05$ s, and $T_t = 1.0$: (a) temporal envelope and (b) its corresponding RIR.

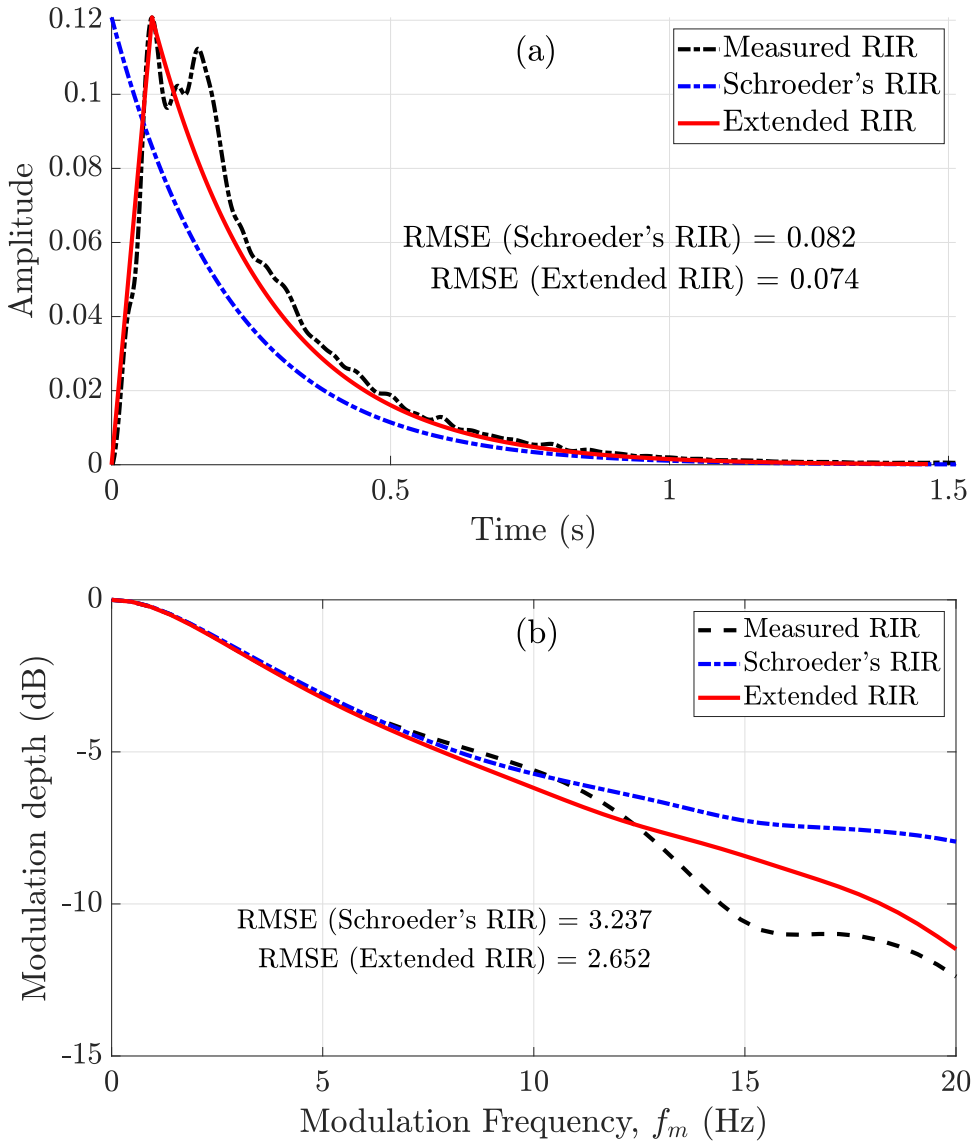

Figure 5.3: Fitting results of two RIR models with temporal amplitude envelope of measured RIR: envelopes in time domain (a) and in modulation-frequency domain (b).

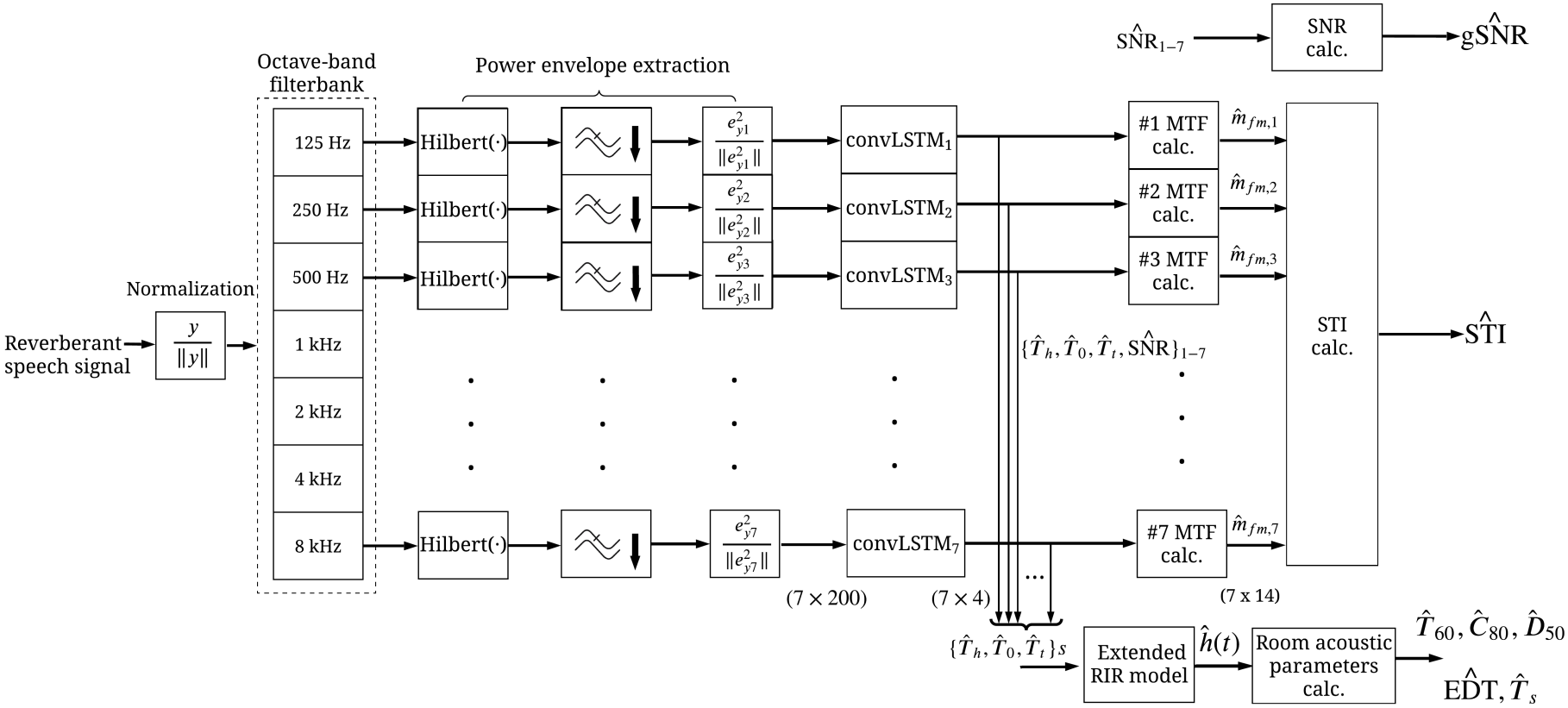

Figure 5.4: Block diagram of proposed method.

Table 5.1: Network architecture of MTF-based CNN model.

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

5.2.1 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 RMSE of the estimated parameters of the RIR model and the target acoustic parameters. It is defined as

$$J(\theta) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\alpha\left(T_{h_n} - \hat{T}_{h_n}\right)^2} + \sqrt{\frac{1}{N}\sum_{n=1}^{N}\beta\left(T_{0_n} - \hat{T}_{0_n}\right)^2} + \sqrt{\frac{1}{N}\sum_{n=1}^{N}\gamma\left(T_{t_n} - \hat{T}_{t_n}\right)^2}, \tag{5.4}$$

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.

5.3 Implementations and evaluations

Here, the similar MTF-based CNN framework were implemented using 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 RMSE and Pearson correlation

coefficient were the evaluation metrics. The errors were calculated from the difference between the ground-truths and estimated parameters.

5.3.1 Data augmentation

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 [98, 117]. 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. (2.5), 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 [113]. 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. (5.2). 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. 5.3. 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.

The CNNs were trained 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 [119]. 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 [119]. 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.

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 [64, 111]. 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 [1, 64]. 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.

5.3.2 Evaluating estimated parameters of RIR models

Simulations were carried out using reverberant speech signals to determine whether the proposed method can correctly estimate 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. Figure 5.5 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 [64]. However, the middle bands, i.e., 500 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. 5.6. Then, the estimated parameters of the RIR model were used to reconstruct the approximated RIR. Figure 5.7 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.

5.3.3 Evaluating estimated MTFs

Figure 5.8 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. (5.3). 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.

5.3.4 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. 5.9. 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 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

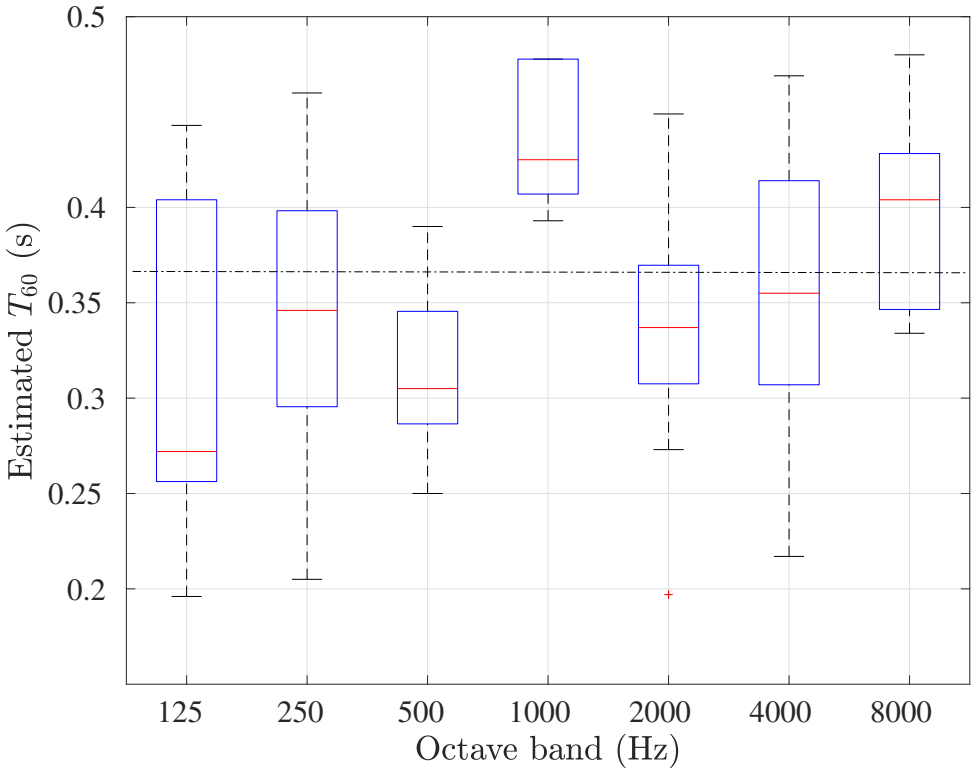

Figure 5.5: Example of estimated parameter T_{60} based on Schroeder's RIR model in octave bands. Horizontal dashed line is ground-truth calculated in full-band ($T_{60} = 0.36$ s). Solid line (red) in each box is median of samples. Size of box represents distribution of estimated values, where ten reverberant speech signals were inputs. Symbol "+" is outlier.

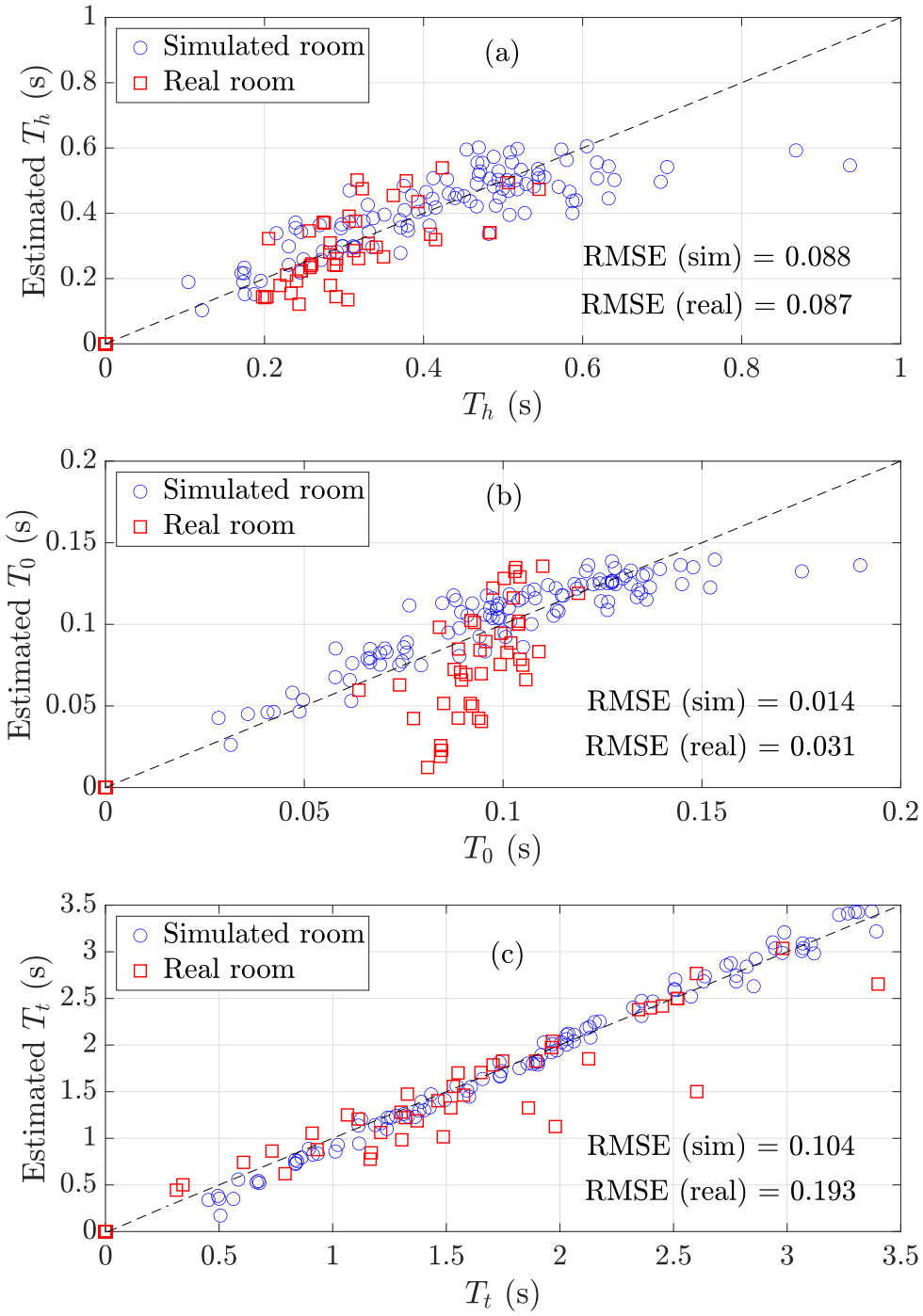

Figure 5.6: Results of estimated parameters of extended RIRs: (a) raising parameter T_h, (b) peak position parameter T_0, and (c) decay parameter T_t.

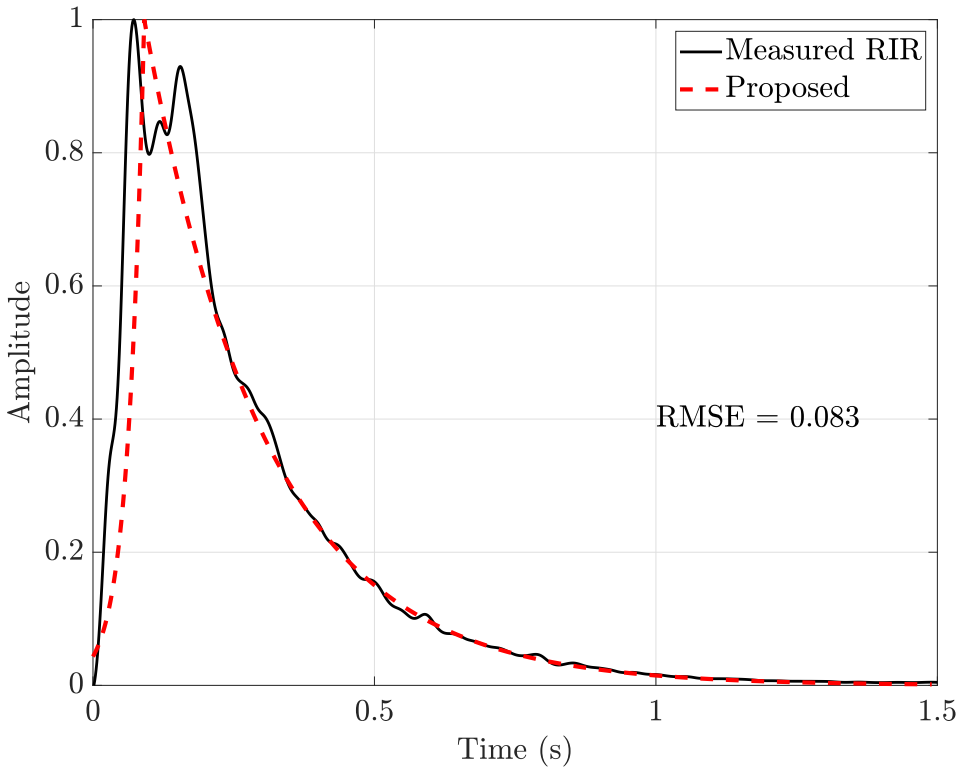

Figure 5.7: Actual and reconstructed RIR using proposed method.

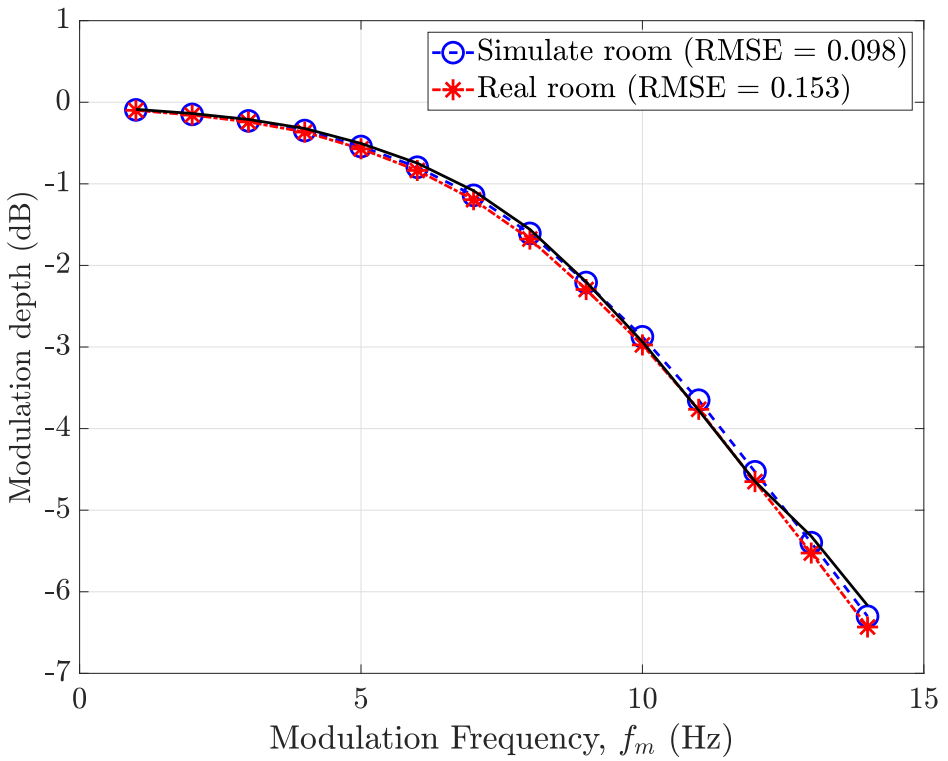

Figure 5.8: Example of MTF estimated from reconstructed RIR. Dashed lines are estimated MTFs where "o" indicates MTFs estimated from simulated room and "*" is MTF estimated from real room. Solid line is ground-truth calculated from RIR.

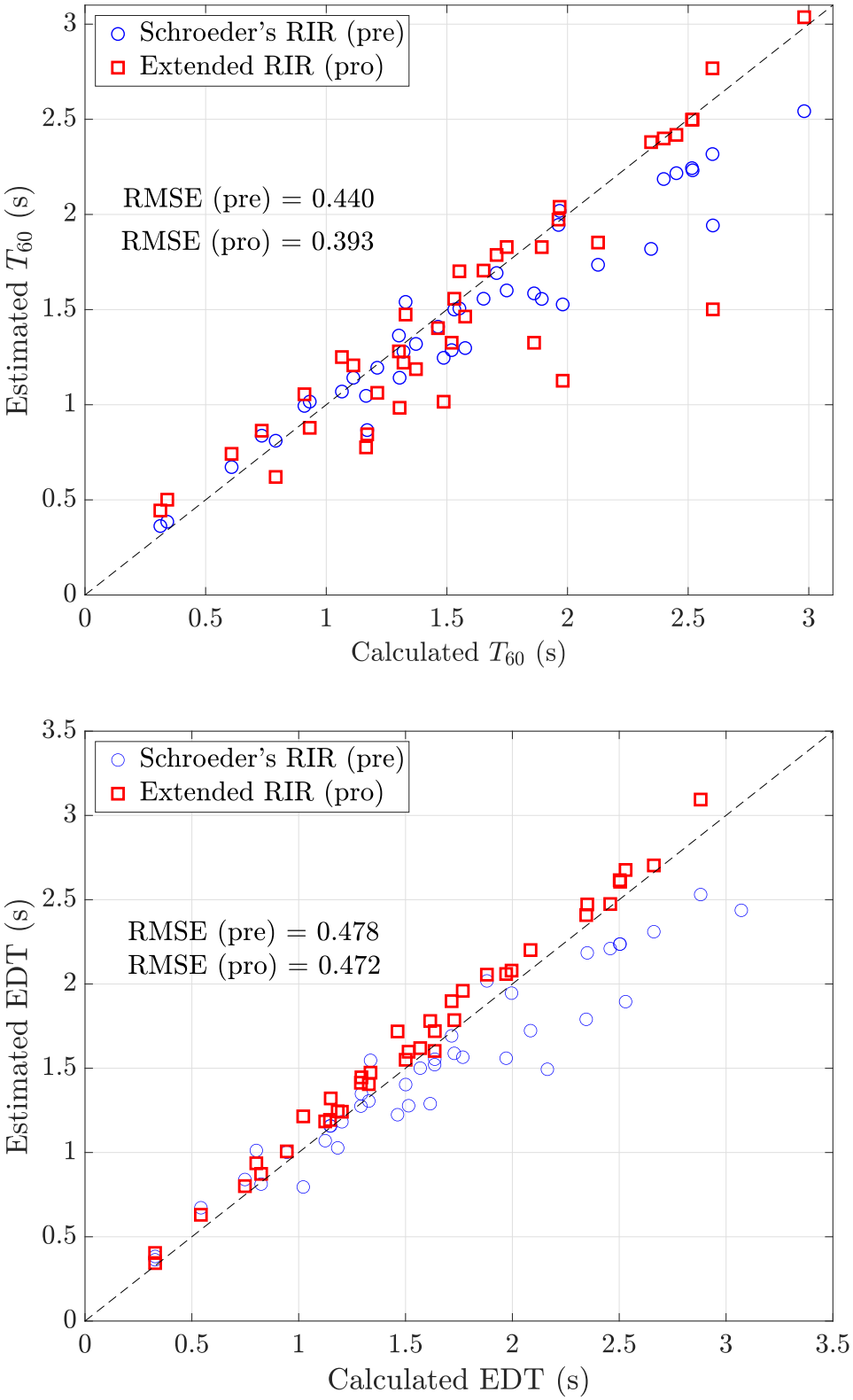

Figure 5.9: Estimated reverberation time (T_{60}) and early decay time (EDT) from reverberant speech signals.

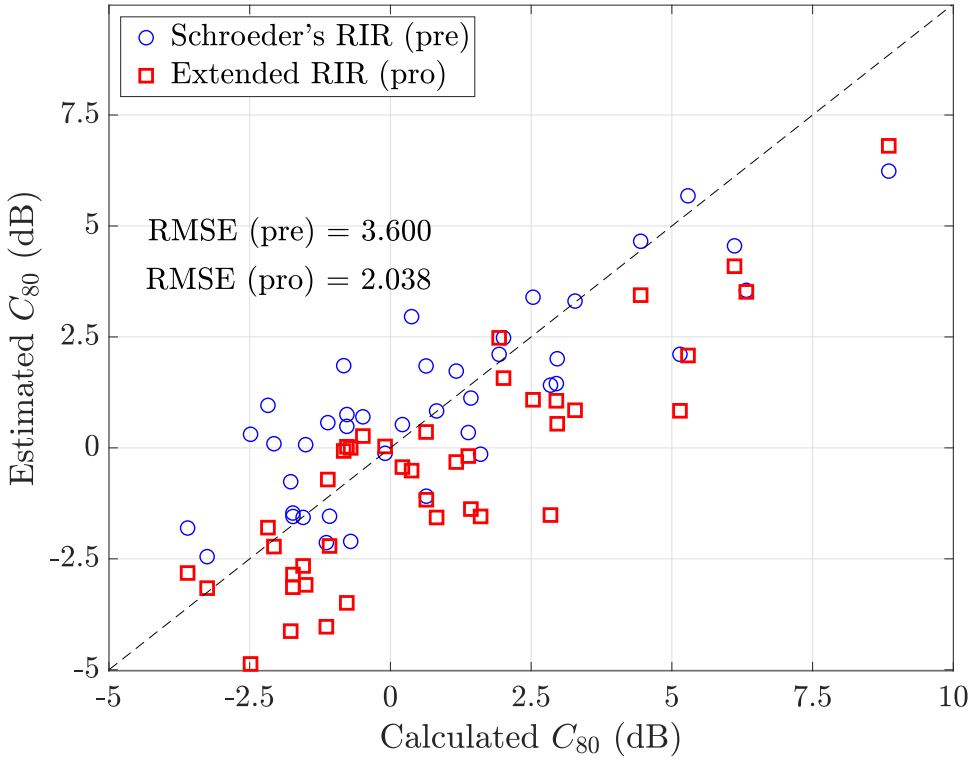

Figure 5.10: Estimated clarity index, C_{80}, from reverberant speech signals.

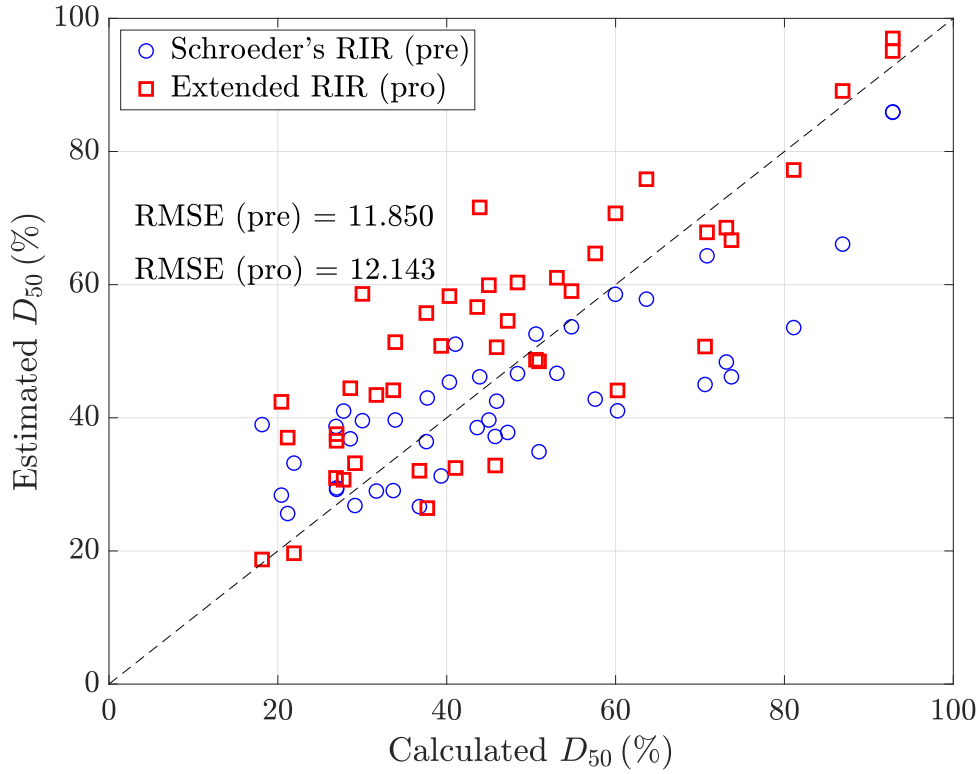

Figure 5.11: Estimated Deutlichkeit, D_{50}, from reverberant speech signals.

Table 5.2: 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 5.3: Comparison between standard derivation (SD) of estimated error and just noticeable difference (JND) of acoustical parameters [3, 4].

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

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., $\hat{T}_{60}$ and $\hat{T}_t$, were also the same. Thus, the results of the proposed method were close to those of the previous method.

The parameters related to the energy ratio of early and late reflection, C_{80}, D_{50}, and T_s, are plotted in Fig. 5.10 and Fig. 5.12, respectively. For C_{80}, the RMSEs were 2.105 with the proposed method and 3.600 with the previous method. For D_{50}, the 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.

Figure 5.13 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 5.2 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) [3]. The JNDs of all acoustical parameters are shown in Table 5.3. Then, the standard derivation of the estimated error was used for comparison with the JND of each parameter.

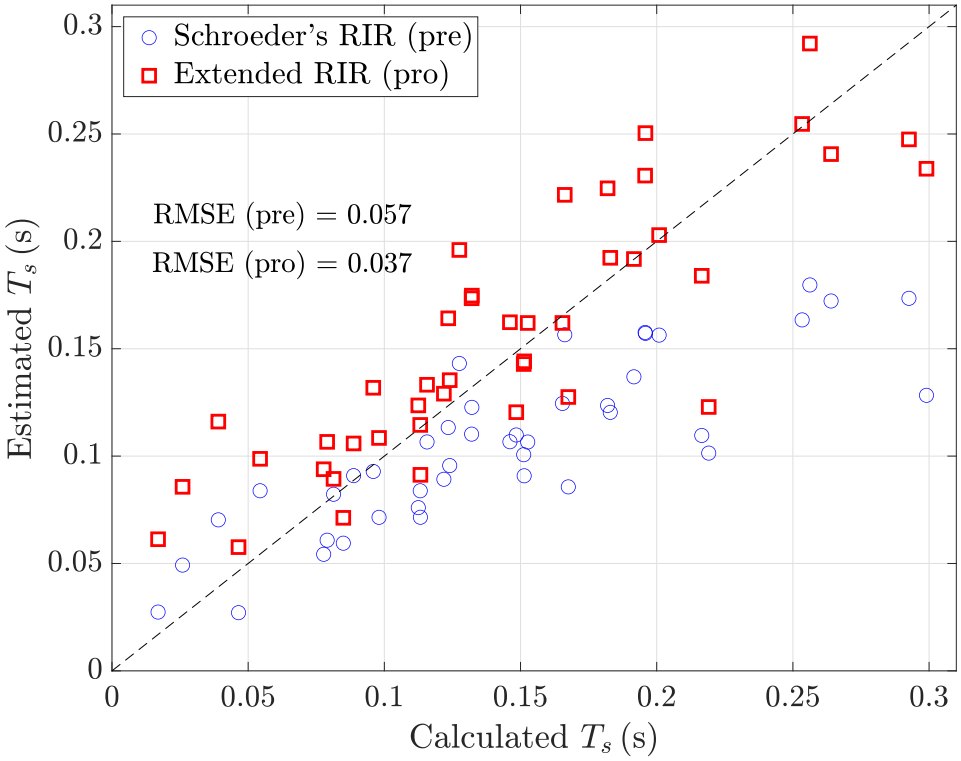

Figure 5.12: Estimated center time, T_s, from reverberant speech signals.

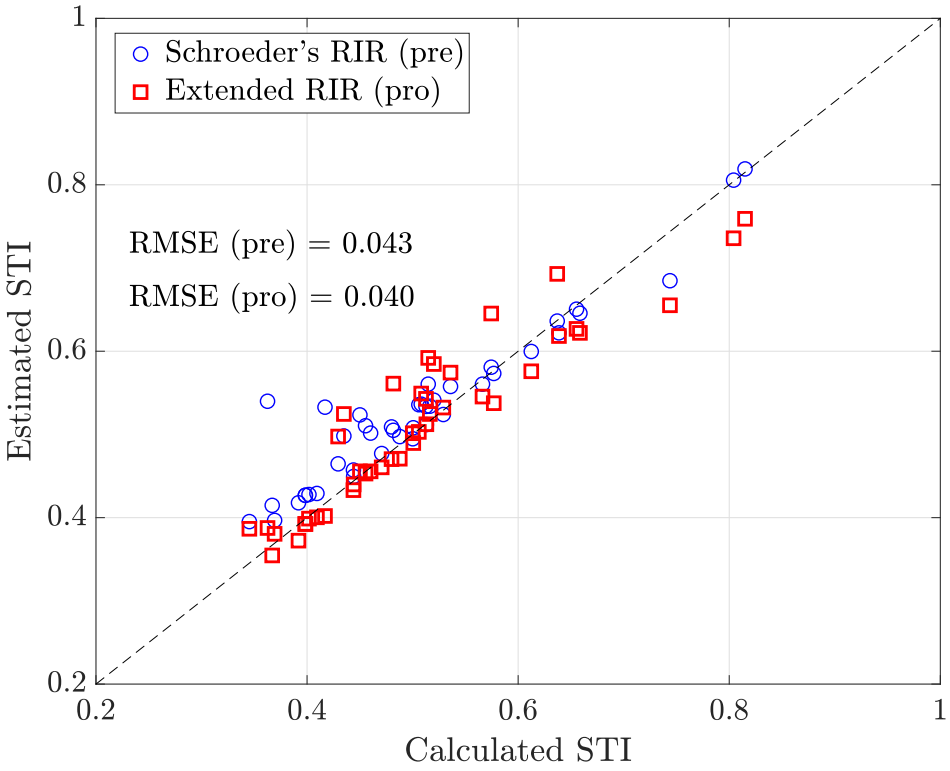

Figure 5.13: Estimated speech transmission index, STI, from reverberant speech signals.

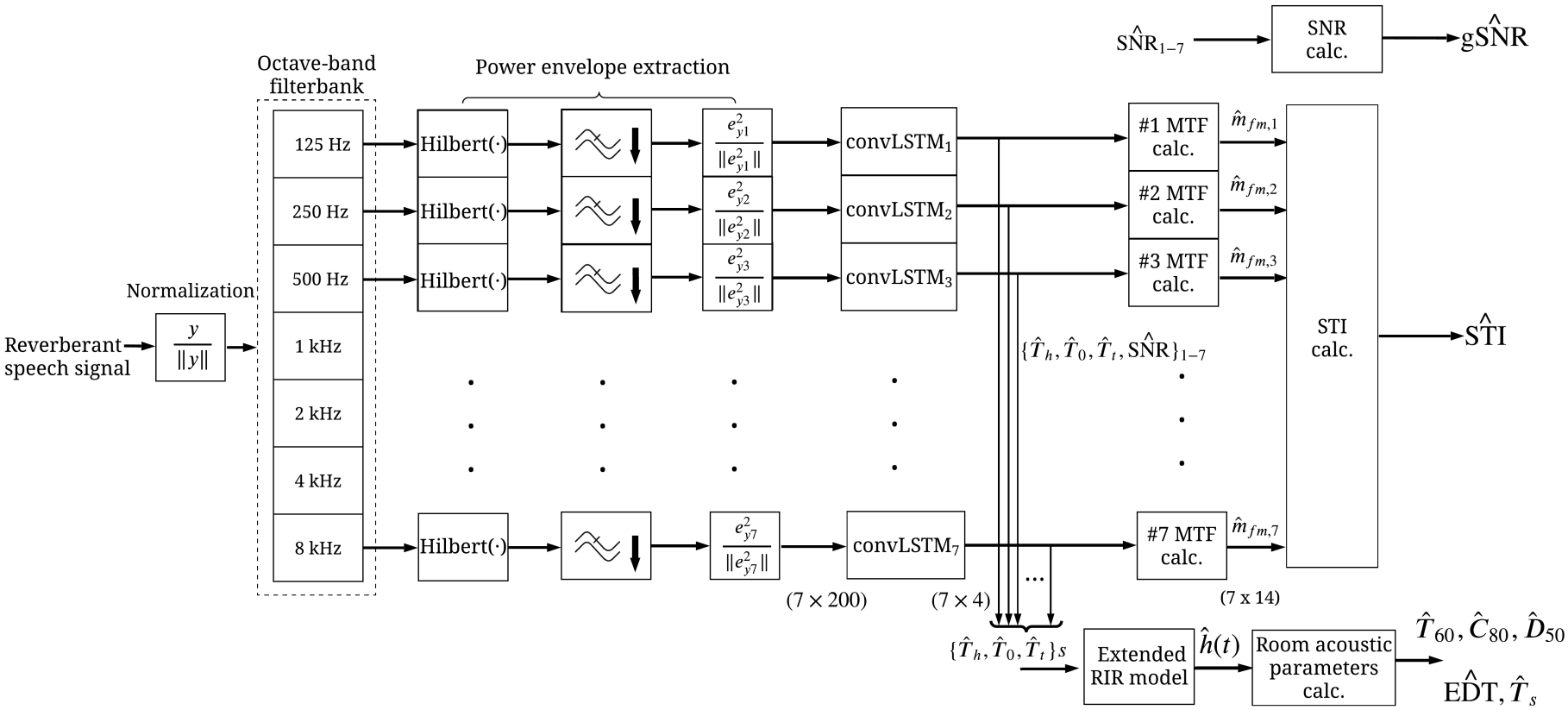

Figure 5.14: Block diagram of the proposed method (Scheme III).

5.4 Estimation in noise environments

In common spaces (e.g., restaurants, department stores, and concourses), background noise, which is a condition of a sound field, is almost inevitable. As background noise degrades speech quality and speech intelligibility, many speech applications, such as speech recognition systems and speaker verification systems, need to consider a level of background noise [59]. Those applications might estimate noise level in terms of a signal-to-noise ratio (SNR) to adapt their parameters and to maintain the performance. Similarly, robustness to background noise is also important for estimating room acoustic parameters.

Background noise is a critical problem not only in estimating acoustic parameters from speech signals but also in the standard procedure from measuring the RIR [3, 21]. The previous chapter studies a method of blindly estimating the extended RIR model parameters for deriving five room-acoustic parameters and STI under reverberant conditions only [120,121]. However, background noise was not included in the model. The accuracy of the estimated parameters might be drastically reduced in noisy reverberant environments.

Thus, background noise is taken into account for estimating room acoustic parameters and STI. The main goal of this study is deployed to estimate the five acoustical parameter and STI. In addition, SNR is estimated along with those parameters since the proposed method takes the SNR into account. As a result, the SNR is then one of the estimated values. The results of estimated SNRs for each band and the global SNRs are the following.

5.5 Discussion

In the above evaluations, the proposed method incorporating the extended RIR model was compared to the previous method that is based on Schroeder's RIR model. The overall estimated 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 model and its estimated parameters. The estimated five room-acoustic parameters and STI were determined from the approximated RIR. We approximate the unknown RIR based on 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 provided the remaining significant errors. The errors were caused by the mismatch between Schroeder's RIR model and many of the realistic RIRs. In contrast, the extended RIR model was taken into account 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 is more accurate than Schroeder's RIR model. Consequently, we hypothesis that the accuracy of the estimated room-acoustic parameters based on the extended RIR model would be better. The model mismatch problem is resolved by incorporating the extended RIR into the proposed framework.

Second, the estimated acoustical parameters and STI have been improved. The parameters that are mainly related to reverberation, i.e., T_{60}, EDT, and STI, were slightly improved. It 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, has been improved by about 35%. This significant improvement of estimated T_s is from the correct estimation of the peak

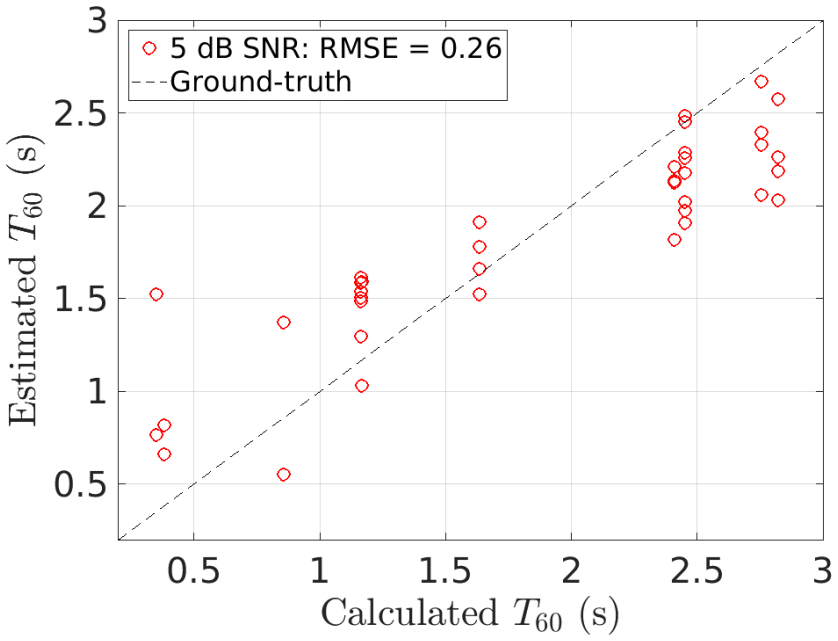

Figure 5.15: Results of the estimated T_{60} in noisy reverberant environments.

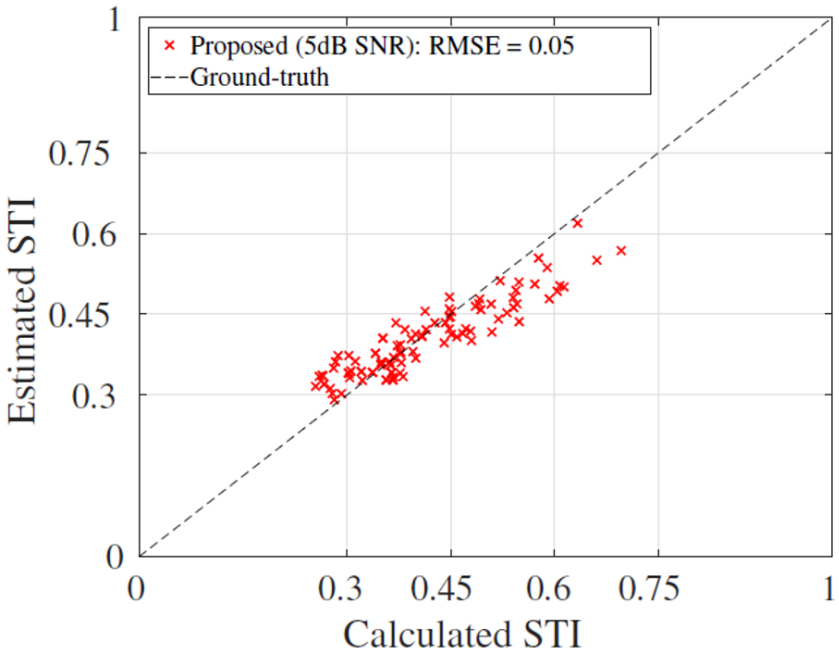

Figure 5.16: Results of the estimated STI in noisy reverberant environments.

positions of the RIRs. Also, the estimated C_{80} has been improved by about 40%. It was revealed that the proposed method could overcome the previous issue. It could correctly estimate those acoustical parameters in realistic acoustic environments even their RIRs are non-exponential decay. Note that since the proposed scheme approximated an unknown RIR in seven-octave bands, the estimated acoustical parameters could be showed for each band according to the requirements of architects, as reported in [64]. However, the estimated D_{50} was insignificantly improved. The reason for the minor improvement is still unclear. Furthermore, a few outliers of the estimated parameters have remained. These outliers are caused by some complicated environments that the RIR models could not represent well. Therefore, precisely estimating the parameters of the extended RIR model and dealing with complicated impulse responses need further investigation.

5.6 Summary

This chapter presents a more accurate method for blindly estimating five-room acoustic parameters and STI. The main idea is an accurate stochastic RIR model for representing the unknown RIR. The extended RIR model is used such that its parameters are estimated from the reverberant speech signal. The CNNs are employed for mapping the relationship between TAE of reverberant speech signal and parameters of the RIR model for seven-octave bands. Later, the power envelope of the observed signal is applied, corresponding to the concept of the MTF. Furthermore, robustness against background noise was investigated whether or not the concept of the MTF can be fully applied. Thus, the proposed method does estimate not only the RIR but also the SNR. Simulations were carried out by using speech signals under realistic reverberant and noisy reverberant conditions. The experimental results suggest that the proposed method can blindly estimate the RIR and the MTF to derive the five-room acoustic parameters and STI correctly. However, the result of the estimated SNRs and the room acoustic parameters under noisy conditions need further consideration since the accuracy of the estimated SNRs remained lower than the baseline.

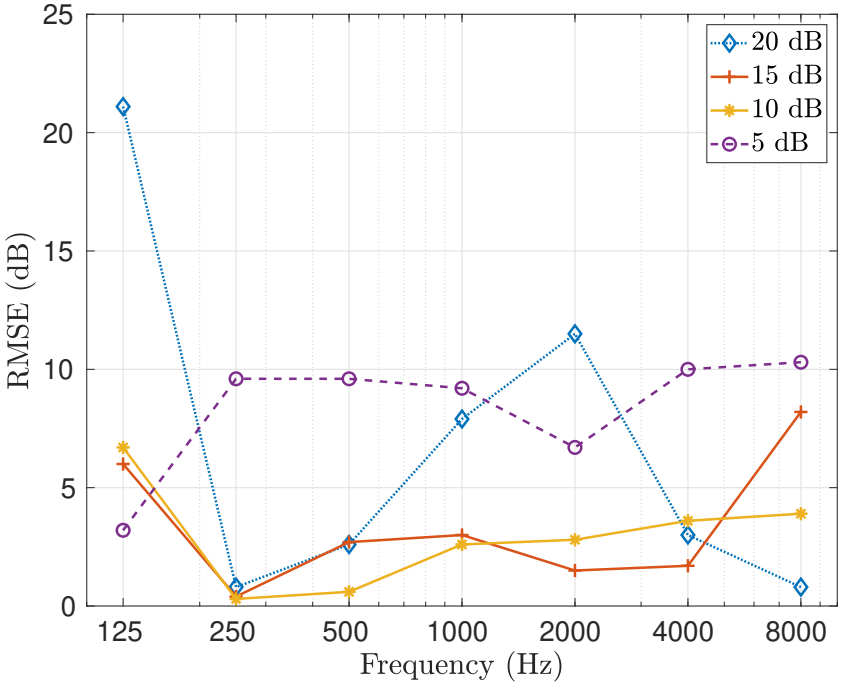

Figure 5.17: Results of the estimated SNRs for sub-bands from reverberant speech signals.

Chapter 6

Conclusion

This chapter summarizes the studies reported in Chapter 3 to Chapter 5. The contributions to knowledge are highlighted. Finally, some remaining limitations and interesting for further improvements are discussed.

6.1 Summary

This dissertation proposed methods for blindly estimating room-acoustic parameters from a speech signal in noisy reverberant environments. Since it is difficult to conduct a direct measurement in places where people exist, the STI was first studied. The STI was estimated by using TAE as a feature incorporated into a CNN model. The CNN models motivated by convolution operation were employed for mapping a TAE of the observed speech signal to the STI. Later, instead of estimating the STI or a single acoustical parameter, the unknown RIRs were approximated by using stochastic RIR models. Therefore, multiple room-acoustic parameters can be estimated. From the evaluation results, the objectives of this research have been achieved.

To be crystal clear, the basis of this work can be broken down as follows.

- The concept of the MTF that represents the effect of reverberation and background noise of room acoustics is used through the temporal amplitude envelope and power envelope of the observed signal, so-called the MTF-based feature.
- TAE and one-dimensional CNN, as a preliminary scheme, was proposed to estimate the STI [98]. The robustness of the estimator can be improved by taking noisy reverberant TAEs into account.
- The power envelope of a noisy reverberant speech as the input can provide a similar result as the TAE.
- Sub-band analysis by following the STI algorithm, i.e, seven-octave bands can improve the accuracy of the estimation and provide estimated parameter of each band.
- The CNN models motivated by convolution operator in the theory of a linear system were exploited, by considering as a blind deconvolution operation for estimating parameters of the RIR models.

- The impulse response of a room can be modeled by using stochastic models. Schroeder's RIR model and the extended RIR model were used for approximating the unknown RIR.
- The parameter(s) of these RIR model were estimated on the basis of the MTF.
- Five room-acoustic parameters, including T_{60}, EDT, C_{80}, D_{50}, and T_s, and STI (an objective index), as well as SNR (background noise condition) can be calculated from the RIR and the MTF.
- The unknown RIRs were reconstructed from blindly estimating their model parameters so that the above acoustical parameters can be blindly estimated simultaneously.
- The extended RIR model can deal with a more variations of sound fields compared with Schroeder's RIR. Consequently, the proposed method using the extended RIR model incorporated into the MTF-based CNN framework outperforms than that of Schroeder's RIR [121].
- Simulations were carried out to determine whether the proposed methods can correctly estimate the acoustic parameters and STI in unseen reverberant or noisy reverberant environments. A few realistic RIRs measured from various sound fields and artificial RIRs from image-source method and stochastic RIR models were used.
- The experimental results in terms of RMSEs and correlation coefficients suggest that the proposed methods can correctly estimate the STI and five room-acoustic parameters. The estimated results were close to the standard method using measured RIR.

6.2 Contributions

This study contributes to the areas of room acoustics and speech signal processing. The significant and original contribution to knowledge from this work can be listed as follows.

- Physical features based on the concept of the MTF (TAE/PE) with convolution-like operator using a machine learning (CNNs) for solving blind estimation problem
- Blindly and correctly estimating the parameters of the extended RIR model
- Simultaneously estimating five room-acoustic parameters and speech transmission index from speech signals in noisy reverberant environments.
- Robust estimation for the quality of sound transmission channel under background noise
- Applicable to use in common spaces with people for quasi-real-time applications

6.3 Recommendation and future works

The following interesting issues are recommended to be investigate further.

1. The quality of sound sources has not been considered yet. This study was limited to estimating the physical properties of a sound field to express subjective aspects. Such subjective perceptions predicted from those room acoustic parameters and STI are only one of three parts of speech communication in an enclosure.

2. Phase response in the MTF domain should be taken into account, e.g., phase shift of the observed signal.

3. Besides noisy reverberant speech signals, a signal from music (e.g., music instruments and/or generic audio signals) is interesting. Based on the literature review, none of the current work could achieve good accuracy of estimating room acoustic parameters and STI from music.

4. The variation of source and receiver position need to be verified the robustness of the proposed method. Since ISO 3382 defines a minimum distance at 1.5 m for measuring T_{60} and other critical distances for other parameters, these critical distances are related to the sound pressure level.

5. The extended RIR model need further verification whether or not it can correctly represent non-diffused spaces.

6. The number of people and their positions might affect the absorption and reflection of a sound wave. Hence, the proposed method should be evaluated in real environments with people.

7. Multichannel approaches using a microphone array (e.g., beamformming algorithm) are interesting for blindly estimating room-acoustic parameters. It might be used to improve the propose method based on a single channel estimation. Moreover, some acoustical parameters related to the spatial domain (e.g., IACC and LEF) should be included.

Appendix A

Controlling estimated STI for protecting privacy of conversation

This study is one of the applications that utilize the estimated STI with the extended RIR model for controlling a level of speech intelligibility. In this case, maintaining the low intelligibility for outside the conversation zone can provide speech privacy.

Protecting the privacy of conversations containing confidential and sensitive information in semi-open rooms, such as in banks and hospitals, is essential because their acoustical characteristics, such as room impulse response (RIR) and background noise, are unknown and prone to change [122]. In the previous work [123], a method, manipulating parameters T_h and T_t of the extended RIR model was used to simulate RIR with low STI. This study proposes a scheme for protecting the privacy of conversations on the basis of feedback control of an estimated speech transmission index (STI). The STI is an objective index related to listening difficulty and is a function of RIR. Without measuring the RIR of the environment where a supposedly private conversation occurs, an STI-estimation method and one RIR model are utilized.

Figure A.1 shows the scheme of the proposed method. The scheme modifies speech signals in such a way that, for an unintended listener, the signals are as unintelligible as they would be in a room with a low STI. To control the late reverberant parameter of the RIR model, a proportional-integral-derivative (PID) controller is used whose controller gains are tuned by using a differential evolution optimizer [124]. The algorithm of the differential evolution is provided. For more details of this study, please see in [30].

A.1 Feedback controller

This work applies a well-known control algorithm that is proportional-integral-derivative controller (PID). A PID or three-term controller is a closed loop system incorporating feedback to control the process variables within a set point. The controller comprises three components serving different purposes. First, the proportion term, a function of present value of the error, provides an overall proportion to the error. Second, the integral term which accumulates the error overtime, minimizes the steady-state error. Third, the derivative term improves transient response using differential compensation. These three parallel terms are defined as

$$u(t) = K_P e(t) + K_I \int e(t)dt + K_D \frac{d}{dt} e(t), \tag{A.1}$$

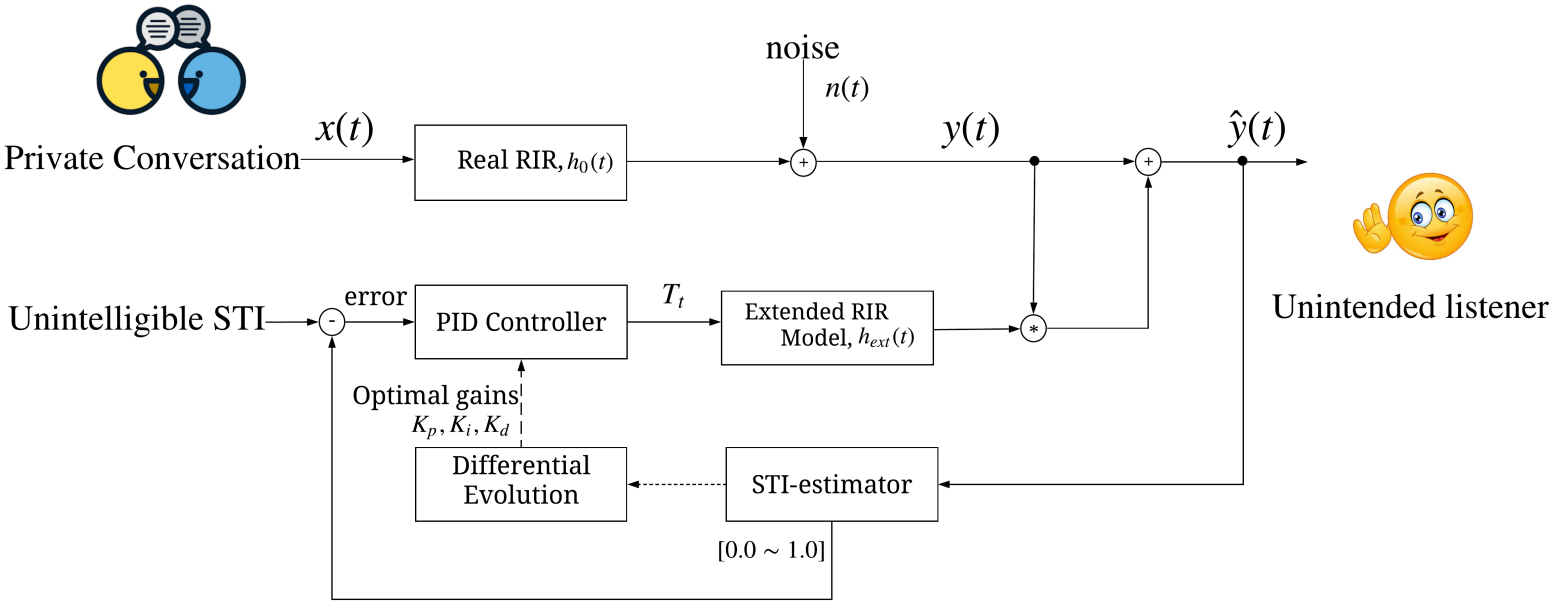

Figure A.1: Block diagram of the privacy control based on optimally controlling estimated STI.

where $e(t)$ represents the difference between estimated STI and the desired value (error). K_P, K_I, and K_D are non-negative values denoted as proportional, integral, and derivative gain, respectively. PID parameters are tuned experimentally to achieve the desired STI of 0.3 with an acceptable error margin of 0.01 within 20 iterations of the controller.

A.2 Differential evolution optimization

The differential evolution algorithm is a parallel direct search method that optimizes a problem by iterative improving solutions with a cost function and a set of constraints. Differential evolution is an appreciate optimization algorithm for the following reasons. First, differential evolution is a multi-point optimizer. Thus, it can effectively handle the starting point problem. Second, differential evolution is a derivative-free approach. Third, it is the fastest algorithm in the evolutionary computation class. Without loss of generality, we assume that the problem is to find a D-dimensional target vector $\mathbf{x}$ such that the cost value $C(\mathbf{x})$ is minimized. Differential evolution algorithm consists of four processes, that are, initialization, mutation, crossover, and selection, as illustrated in Fig. A.2.

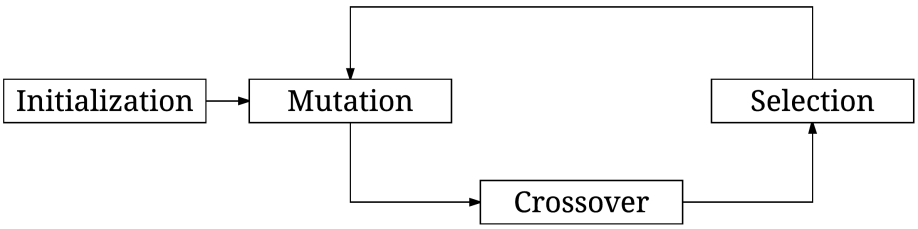

Figure A.2: Differential evolution algorithm.

1. *Initialization.* The initial target vectors $\mathbf{x}_{i,G}$ of the generation $G = 1$, for $i = 1, 2, .., NP$, where NP is the total population, are randomly generated. Note that these initial target vectors should cover the entire solution space.

2. *Mutation.* Each target vector $\mathbf{x}_{i,G}$ is used to generate a mutant vector $\mathbf{v}_{i,G+1}$ by the following formula.

$$\mathbf{v}_{i,G+1} = \mathbf{x}_{r_1,G} + F \cdot (\mathbf{x}_{r_2,G} - \mathbf{x}_{r_3,G}), \tag{A.2}$$

where i, r_1, r_2, and r_3 are distinct and randomly chosen from $\{1, 2, .., NP\}$, and the predefined constant F is in the interval $[0, 2]$. This constant determines the convergence rate of the algorithm.

3. *Crossover.* Each pair of target vector $\mathbf{x}_{i,G}$ and its mutant vector $\mathbf{v}_{i,G+1}$ is used to generate a trial vector $\mathbf{u}_{i,G+1}$ by the following formula.

$$\mathbf{u}_{i,G+1} = \begin{bmatrix} u_{1i,G+1} \\ u_{2i,G+1} \\ \vdots \\ u_{Di,G+1} \end{bmatrix},$$

$$\mathbf{u}_{ji,G+1} = \begin{cases} \mathbf{v}_{ji,G+1}, & \text{if } \Xi(j) \leqslant CR \text{ or } j = \upsilon, \\ \mathbf{x}_{ji,G}, & \text{otherwise,} \end{cases} \tag{A.3}$$

where $\Xi(j)$ is a uniform random number generator with a result in the interval $[0, 1]$, the crossover constant CR is a predefined constant in $[0, 1]$, and v is chosen randomly from $\{1, 2, ..., D\}$ to ensure that the trial vector gets at least one element from the mutant vector.

4. *Selection.* $C(\mathbf{x}_{i,G})$ and $C(\mathbf{u}_{i,G+1})$ are compared, where C is the cost function. The one with the less cost value survives for the next generation $G+1$. That is, if $C(\mathbf{x}_{i,G}) \leq C(\mathbf{u}_{i,G+1})$, then $\mathbf{x}_{i,G+1} = \mathbf{x}_{i,G}$; otherwise, $\mathbf{x}_{i,G+1} = \mathbf{u}_{i,G+1}$. Once all NP members of generation $G+1$ are obtained, they iterative continue until some condition or constraint is satisfied. The solution is the vector $\mathbf{x}_i$ from the last generation that yields the lowest cost.

A.3 Evaluation and discussion

Simulations of the controlling estimated STIs and subjective tests are conducted to evaluate the performance of the proposed method. The stimuli consist of 16 Thai short commands of three syllables spoken by Thai female announcer [125]. For the subjective test, the stimuli are divided into two groups. The first group includes eight stimuli that are generated from $h_L(t)$ with a fixed value of 10.5 second. This constant value is selected from the appropriate value for masking a signal in a specific room. The second group includes eight stimuli that are generated as a result of the proposed method, i.e., PID-DE control algorithm. These two groups are compared to study whether the proposed method outperforms the open-loop method.

The objective metric is the actual estimated STIs, which is controlled to be poor intelligibility. The datasets used in this test include 12 three-syllable Thai voice command, 43 RIRs in [111], pink noise with a signal-to-noise ratio (SNR) equal to 20 dB. The target speech signals are generated by convoluting speech with RIR and adding noise.

The optimal gains of three PID parameters (i.e., K_P, K_I, and K_D) are obtained from the DE. The system performance is then evaluated when it reaches the maximum iteration or converses to the target STI. The error is determined by the absolute difference between the target and the actual of the estimated STIs. The simulation result of the estimated STIs in the reverberant room and of the target signal corrupted with pink noise (i.e., 20 dB SNR) are shown in Fig. A.3. In addition, the proposed method in various rooms, which is different RIRs, is shown in Fig. A.4. The results showed that the average error between the actual and target STIs converse to zero within ten iterations. The average error at steady-state of clean reverberant signal and signal with pink noise is 0.01 and 0.02, respectively. Note that these errors are not significant according to the limitation of the STI-estimator.

From the subjective results, as shown in II, the proposed method can manipulate the target speech signals to make them unintelligible. The proposed method has a lower WIR and a higher LDR than the open-loop method. However, the proposed method provides slightly higher annoyance than the open-loop method. From the proposed method results, the higher annoyance might cause the lower STI (which is better for masking the information in conversations). The results of the study in [123] are shown as a reference for supporting the relationship between the controlled STI and the three subjective indicators. According to the three subjective indicators, an STI of 0.3 causes the speech signal $y(t)$ to be unintelligible. Hence, the proposed method takes this STI value as the setpoint.

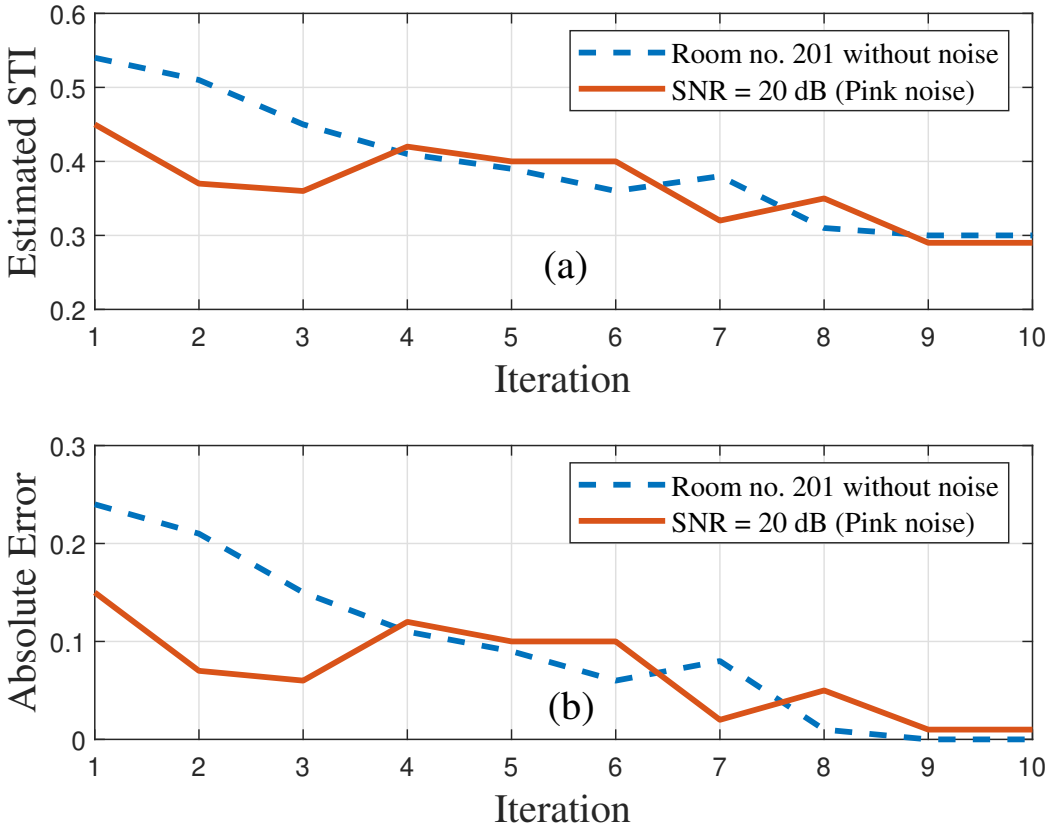

Figure A.3: The proposed method under two conditions of background noise: (a) the estimated STI at each iteration and (b) the average error.

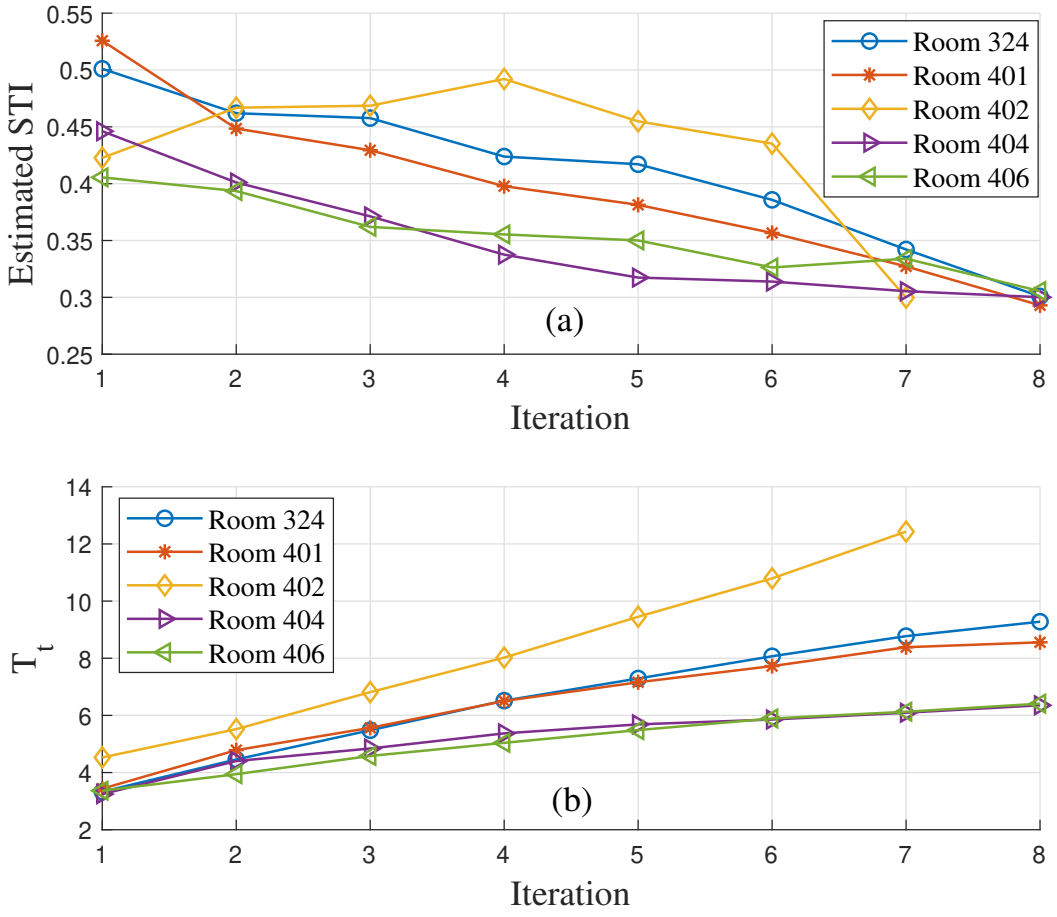

Figure A.4: The proposed method under variations room conditions: (a) the output of the estimated STIs and (b) the controlled parameter of the extended RIR model, T_t.

A.4 Summary

The proposed method extended the method proposed by Unoki *et at.*. This scheme is available to control the estimated STI in real-time. Hence, it can be used under variations of noisy reverberant conditions. Note that the results of the method in [123] used a different language (Japanese) and with different subjects (native Japanese speakers). Moreover, the calculation and the familiarity of words are different. Therefore, biases in the comparison exist, and more research is needed.

Appendix B

Room impulse responses: SMILEdataset

Table B.1: Dataset of room impulse responses (RIRs).

No.	Index	Description	T_{60}	STI
1	301	Multi-purpose hall 1 (with reflex board)	1.09	0.80
2	302	Multi-purpose hall 1 (without reflex board)	0.80	0.63
3	303	Multi-purpose hall 2 (with reflex board)	1.44	0.51
4	304	Multi-purpose hall 2 (without reflex board)	1.04	0.57
5	305	Multi-purpose hall 3 (with reflex board)	1.93	0.44
6	306	Multi-purpose hall 3 (without reflex board)	1.35	0.53
7	307	Multi-purpose hall 3 (with absorption board)	1.42	0.55
8	308	Multi-purpose hall 4 (with absorption board)	1.54	0.52
9	319	Multi-purpose hall 5 (14000 m^3)	1.47	0.53
10	321	Multi-purpose hall 6 (19000 m^3)	2.16	0.43
11	309	Concert hall 1 (5600 m^3)	2.52	0.40
12	310	Concert hall 1 ($d = 6\,\mathrm{m}$)	2.39	0.41
13	311	Concert hall 1 ($d = 11\,\mathrm{m}$)	2.51	0.40
14	312	Concert hall 1 ($d = 15\,\mathrm{m}$)	2.45	0.40
15	313	Concert hall 1 ($d = 19\,\mathrm{m}$)	2.60	0.39
16	314	Concert hall 2 (6100 m^3)	1.16	0.57
17	315	Concert hall 3 (20000 m^3)	1.96	0.43
18	316	Concert hall 4 (with absorption curtain)	1.86	0.47
19	317	Concert hall 4 (without absorption curtain)	2.60	0.43
20	323	Concert hall 5 (17000 m^3)	2.35	0.62
21	324	Concert hall 6 (front)	1.65	0.47
22	325	Concert hall 6 (side)	1.74	0.46
23	326	Concert hall 6 (3F)	1.89	0.45
24	201	Lecture room with flatter echoes	1.30	0.54
25	318	Theater hall (3900 m^3)	0.91	0.61
26	401	Meeting room (130 m^3)	0.60	0.69
27	402	Lecture room 1 (400 m^3)	0.93	0.60
28	403	Lecture room 2 (2400 m^3)	1.17	0.65
29	404	General speech hall (11000 m^3)	1.55	0.48
30	405	Church 1 (1200 m^3)	0.73	0.65
31	406	Church 2 (3200 m^3)	1.53	0.50
32	407	Event hall 1 (28000 m^3)	3.40	0.36
33	408	Event hall 2 (41000 m^3)	3.60	0.34
34	409	Gym 1 (12000 m^3)	2.98	0.37
35	410	Gym 2 (29000 m^3)	1.70	0.47
36	411	Living room (110 m^3)	0.34	0.80
37	412	Movie theater (560 m^3)	0.31	0.81
38	413	Atrium (4000 m^3)	1.32	0.50
39	414	Tunnel (5900 m^3)	3.88	0.40
40	415	Concourse in train station	1.98	0.50
41	416	General speech hall 2 (1F front)	1.52	0.52
42	417	General speech hall 2 (1F center)	1.57	0.50
43	418	General speech hall 2 (1F balcony)	1.48	0.52
44	C-11	Concert hall of about 500 people	7.34	0.67
45	C-12	Echoic chamber	1.83	0.73

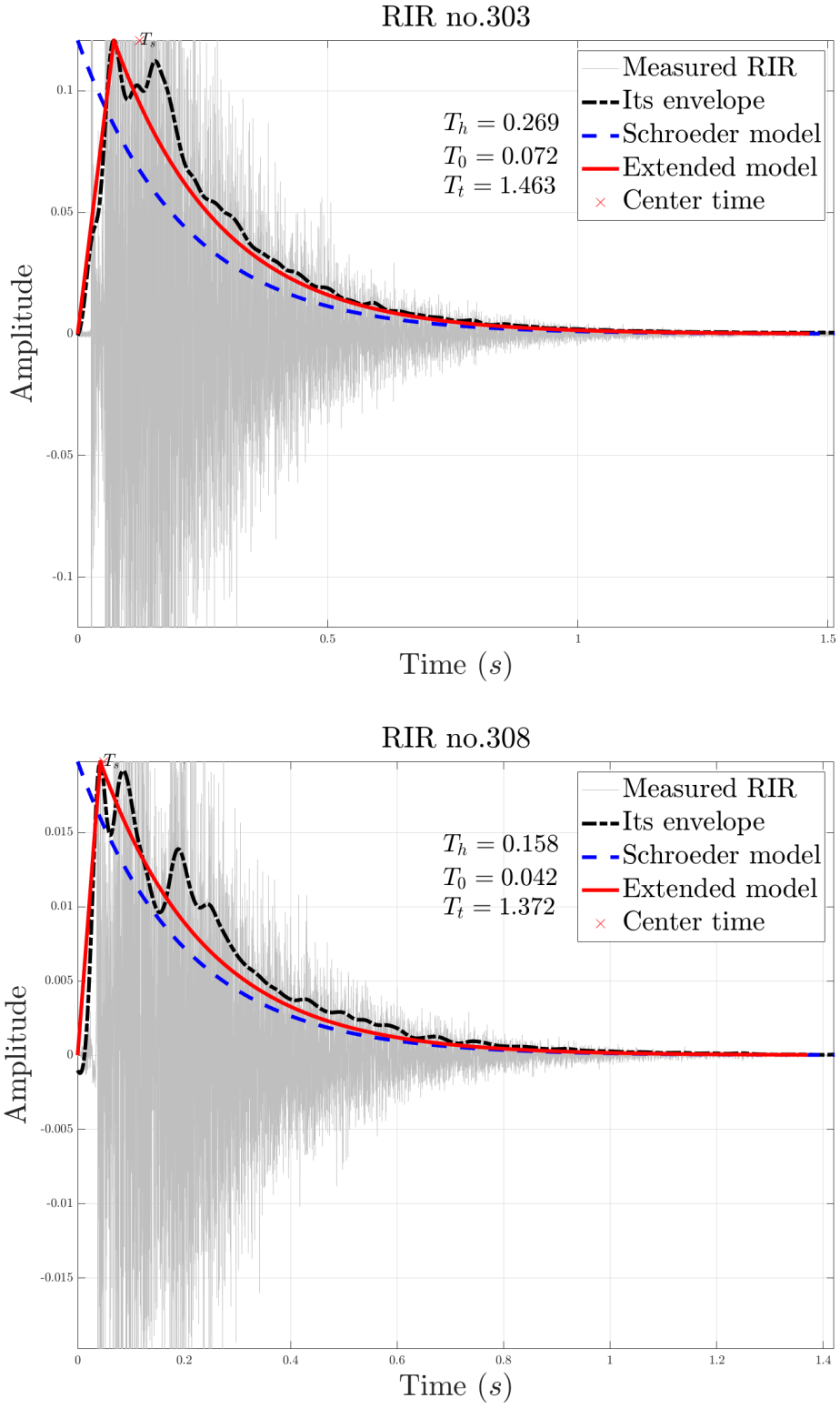

Figure B.1: Examples (I): room impulse responses (RIRs), its envelopes (black dotted line), and RIR models. Solid line is the extended RIR model with model's parameters and dashed line is Schroeder's RIR model. Cross symbol is a position of the Center time, Ts.

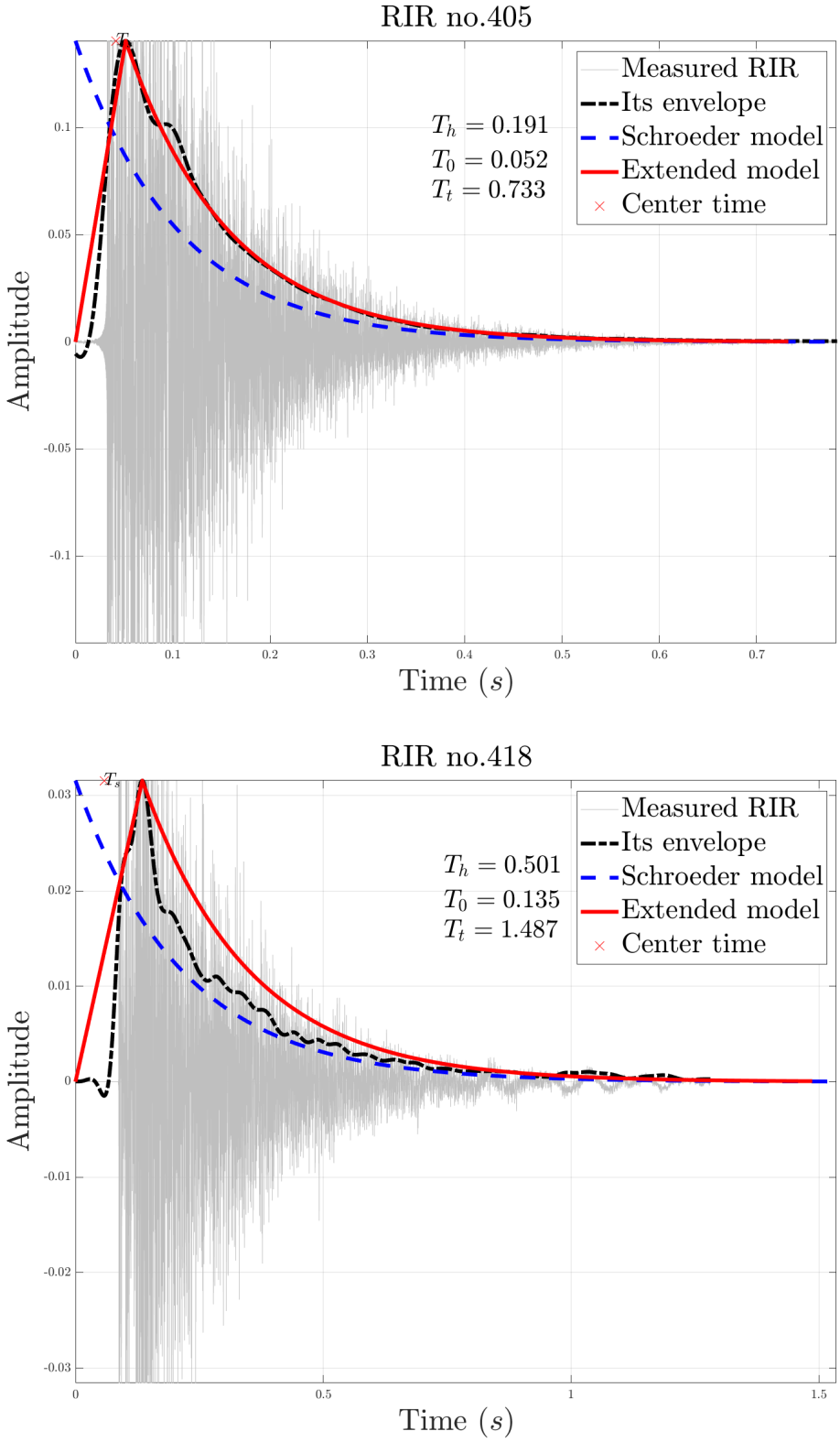

Figure B.2: Examples (II): room impulse responses (RIRs), its envelopes (black dotted line), and RIR models. Solid line is the extended RIR model with model's parameters and dashed line is Schroeder's RIR model. Cross symbol is a position of the Center time, Ts.

Appendix C

Variation of CNNs for blind parameter estimation

In this dissertation, a few neural network models have been proposed. Most of them are CNN because the convolution operation of the CNN is similar to the basis of the operation between the input signal and the RIR. However, estimating RIR models are focused. Other different CNN models are investigated in this section.

C.1 Experiments and evaluations

Since there are many variations of the CNN architecture, this section presents a few experiments regarding the effect of different CNN models and performance of the estimator. Some of state-of-the-art CNN architectures are investigated, such as VGG, ResNet, DenseNet, and EfficientNet. Unfortunately, all of them are developed for image processing. In this study takes a power envelope of a speech signal as one-dimension input feature. In addition, the dataset used in this study is small comparing with the original data of a few million images to train these DNNs. Thus, the transfer learning and fine-tuning technique is used to avoid the overfitting problem. The baseline CNN model have been extended to different models.

The computational time and complexity of the proposed method were considered. The proposed method is applicable for real-time assessments for two reasons. The first is a short period of recording a reverberant speech signal. The proposed method needs only five seconds of a reverberant speech signal. The second is that a few computational time is required. We evaluated the computational time on general 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 need significantly least 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. The evaluation results can be summarized, as shown in the table and figure below.

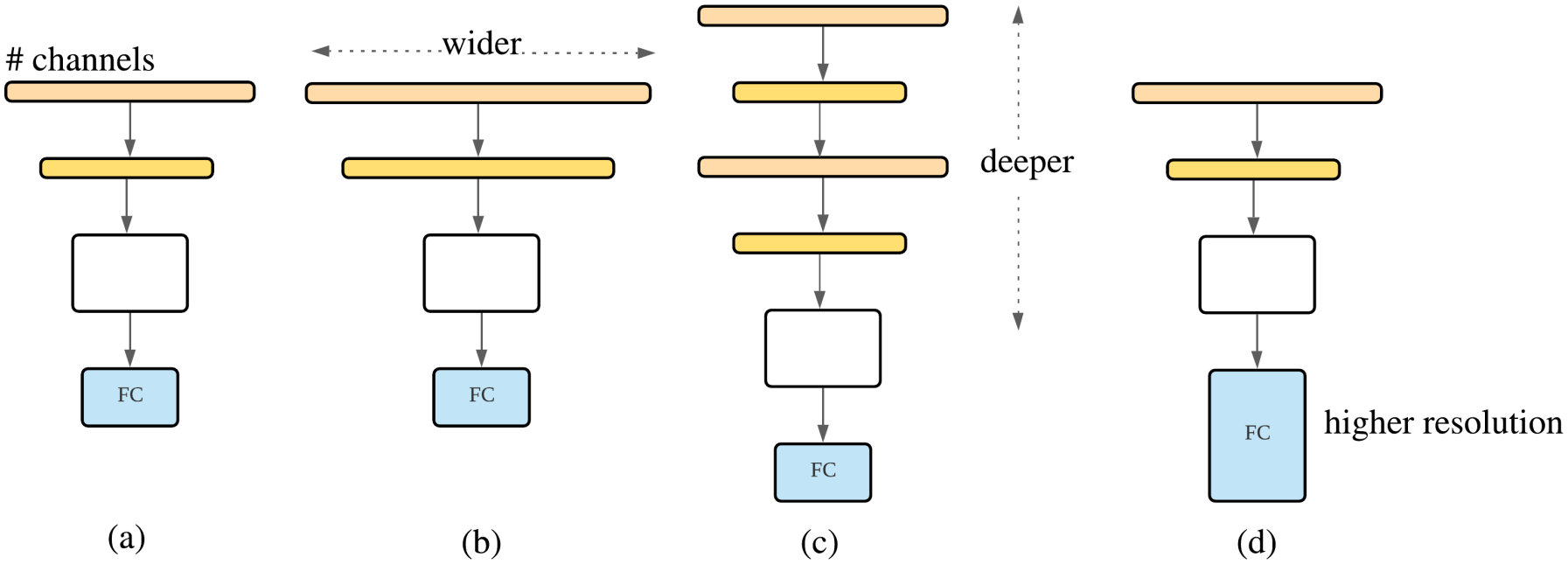

Figure C.1: CNN models with a variation of scaling: (a) baseline, (b) width scaling, (c) depth scaling (d) resolution scaling.

	baseline	**wider**	**deeper**	**higher res**	**all scaled up**
No. of parameters	**6,419**	893,699	3,919,043	125,467	**4,734,403**
RMSE	0.022	0.013	0.008	0.020	**0.006**
Improvement	-	41%	64%	10%	73%
Computational time	0.26 s	0.38 s	0.41 s	0.31 s	0.74 s

Table C.1: Comparison results of variation of CNN Scaling in the MTF-based parameter of the extended RIR model estimation.

C.2 Discussion

From the result, the different CNN models have significantly different performances. The baseline used in Chapter $4-5$ provides acceptable accuracy. At the same time, it is the fastest in terms of computational time. It is also more generalized than the deeper CNNs, since the baseline is the smallest parameter. On the other hand, the all-dimension scaling up model contains a huge number of parameters (4.7M), which is 737 times larger than the baseline. Although this deep CNN can provide the highest accuracy (the lowest RMSE), we have to trade-off with the higher computational time as well as it has a likelihood to overfit to the training set.

Appendix D

Supplementary materials

The supplementary materials developed in this dissertation are provided in the following online sources.

- **Source code:**

  https://github.com/GolfSuradej/Room-acoustic-parameters-estimation-based-on-the-concept-of-the-MTF

- **RIR database:** Duangpummet, Suradej (2021), "Room Impulse responses", Mendeley Data, V1, doi: 10.17632/28hfxyrnwj.1

- **Speech and noise database:** https://github.com/GolfSuradej/Noise-dataset

- **Demo and application:** http://www.suradejresearch.com/

- **Recorded Presentations:** https://youtu.be/X6HoNpKualA

Bibliography

[1] 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, no. 6, pp. 1430–1445, 2017.

[2] M. Unoki, Y. Yamasaki, and M. Akagi, "MTF-based power envelope restoration in noisy reverberant environments," in *2009 17th European Signal Processing Conference (EUSIPCO)*. IEEE, 2009, pp. 228–232.

[3] ISO 3382:2009, "Acoustics – measurements of room acoustics parameters – part 1: Performance spaces," 2009.

[4] J. Bradley, "Review of objective room acoustics measures and future needs," *Applied Acoustics*, vol. 72, no. 10, pp. 713–720, 2011.

[5] C. J. Plack, *The sense of hearing*. Routledge, 2018.

[6] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

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

[8] W. Yang and J. Bradley, "Effects of room acoustics on the intelligibility of speech in classrooms for young children," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 922–933, 2009.

[9] R. Ljung, P. Sörqvist, A. Kjellberg, and A.-M. Green, "Poor listening conditions impair memory for intelligible lectures: implications for acoustic classroom standards," *Building Acoustics*, vol. 16, no. 3, pp. 257–265, 2009.

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

[11] M. Barron, *Auditorium acoustics and architectural design*. Routledge, 2009.

[12] W. J. Murphy and N. Xiang, "Room acoustic modeling and auralization at an indoor firing range," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, pp. 3868–3872, 2019.

[13] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[14] S. K. Mitra and Y. Kuo, *Digital signal processing: a computer-based approach*. McGraw-Hill New York, 2006, vol. 2.

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

[16] T. Houtgast, H. J. M. Steeneken, and R. Plomp, "Predicting speech intelligibility in rooms from the modulation transfer function. I. general room acoustics," *Acta Acustica united with Acustica*, vol. 46, no. 1, pp. 60–72, 1980.

[17] T. Houtgast and H. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acta Acustica United with Acustica*, vol. 28, no. 1, pp. 66–73, 1973.

[18] 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.

[19] W. C. Sabine, *Architectural acoustics*. Cambridge: Harward University Press, 1923.

[20] V. L. Jordan, "Acoustical criteria for auditoriums and their relation to model techniques," *The Journal of the Acoustical Society of America*, vol. 47, no. 2A, pp. 408–412, 1970.

[21] IEC 60268-16:2020, "Sound system equipment–part 16: Objective rating of speech intelligibility by speech transmission index," International standard, 2011.

[22] K. L. Payton and L. D. Braida, "A method to determine the speech transmission index from speech waveforms," *The Journal of the Acoustical Society of America*, vol. 106, no. 6, pp. 3637–3648, 1999.

[23] Z. Ding and Y. Li, *Blind equalization and identification*. CRC press, 2001.

[24] D. Kundur and D. Hatzinakos, "Blind image deconvolution," *IEEE signal processing magazine*, vol. 13, no. 3, pp. 43–64, 1996.

[25] M. Schutte, S. D. Ewert, and L. Wiegrebe, "The percept of reverberation is not affected by visual room impression in virtual environments," *The Journal of the Acoustical Society of America*, vol. 145, no. 3, pp. EL229–EL235, 2019.

[26] C. Yu, C. Zhang, and L. Xie, "An envelope signal based deconvolution algorithm for ultrasound imaging," *Signal Processing*, vol. 92, no. 3, pp. 793–800, 2012.

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

[28] Y.-L. You and M. Kaveh, "Blind image restoration by anisotropic regularization," *IEEE Transactions on Image Processing*, vol. 8, no. 3, pp. 396–407, 1999.

[29] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation.* Springer Science & Business Media, 2010.

[30] S. Duangpummet, P. Kraikhun, C. Phunruangsakao, J. Karnjana, M. Unoki, and W. Kongprawechnon, "Speech privacy protection based on optimal controlling estimated speech transmission index in noisy reverberant environments," in *28th European Signal Processing Conference (EUSIPCO).* IEEE, 2020, pp. 76–80.

[31] C. Phunruangsakao, P. Kraikhun, S. Duangpummet, J. Karnjana, M. Unoki, and W. Kongprawechnon, "Speech privacy protection based on controlling estimated speech transmission index," in *2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON).* IEEE, 2020, pp. 628–631.

[32] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "A method based on the MTF concept for dereverberating the power envelope from the reverberant signal," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003.

[33] N. R. Shabtai, Y. Zigel, and B. Rafaely, "Towards room-volume classification from reverberant speech using room-volume feature extraction and room-acoustics parameters," *Acta Acustica United with Acustica*, vol. 99, no. 4, pp. 658–669, 2013.

[34] N. K. Bui, D. Morikawa, and M. Unoki, "Method of estimating direction of arrival of sound source for monaural hearing based on temporal modulation perception," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5014–5018.

[35] R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S. Q. Lee, and H.-M. Park, "Sound source localization based on GCC-PHAT with diffuseness mask in noisy and reverberant environments," *IEEE Access*, vol. 8, pp. 7373–7382, 2020.

[36] L. Ljung, "System identification," *Wiley encyclopedia of electrical and electronics engineering*, pp. 1–19, 1999.

[37] H. Nomura, H. Miyata, and T. Houtgast, "Speech intelligibility and MTF in non-exponential decay fields," in *International Conference on Acoustics, Speech, and Signal Processing,*, 1989, pp. 1985–1988 vol.3.

[38] H. Sato, M. Morimoto, H. Sato, and M. Wada, "Relationship between listening difficulty and acoustical objective measures in reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2087–2093, 2008.

[39] 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.

[40] R. F. Lyon, *Human and machine hearing*. Cambridge University Press, 2017.

[41] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 585–592, 1995.

[42] T. Ngo, R. Kubo, and M. Akagi, "Increasing speech intelligibility and naturalness in noise based on concepts of modulation spectrum and modulation transfer function," *Speech Communication*, vol. 135, pp. 11–24, 2021.

[43] J. E. Preminger and D. J. V. Tasell, "Quantifying the relation between speech quality and speech intelligibility," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 3, pp. 714–725, 1995.

[44] A. Gabrielsson and H. Sjögren, "Perceived sound quality of sound-reproducing systems," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 1019–1033, 1979.

[45] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)–a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[46] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, no. 4, pp. 482–499, 1989.

[47] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *The Journal of the acoustical society of America*, vol. 61, no. 5, pp. 1337–1351, 1977.

[48] S. Amano, S. Sakamoto, T. Kondo, and Y. Suzuki, "Development of familiarity-controlled word lists 2003 (fw03) to assess spoken-word intelligibility in japanese," *Speech Communication*, vol. 51, no. 1, pp. 76–82, 2009.

[49] K. D. Kryter, "Methods for the calculation and use of the articulation index," *The Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1689–1697, 1962.

[50] ANSI, "S3. 5-1997, Methods for the calculation of the speech intelligibility index," *New York: American National Standards Institute*, vol. 19, pp. 90–119, 1997.

[51] C. Pavlovic, "SII–speech intelligibility index standard: ANSI s3. 5 1997," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1906–1906, 2018.

[52] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *The journal of the Acoustical society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[53] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[54] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[55] K. L. Payton and M. Shrestha, "Comparison of a short-time speech-based intelligibility metric to the speech transmission index and intelligibility data," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3818–3827, 2013.

[56] C. Pavlovic, "The speech intelligibility index standard and its relationship to the articulation index, and the speech transmission index," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3326–3326, 2006.

[57] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[58] H. Sato, M. Morimoto, and M. Wada, "Relationship between listening difficulty rating and objective measures in reverberant and noisy sound fields for young adults and elderly persons," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4596–4605, 2012.

[59] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[60] 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.

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

[62] P. Guidorzi, L. Barbaresi, D. D'Orazio, and M. Garai, "Impulse responses measured with MLS or swept-sine signals applied to architectural acoustics: an in-depth analysis of the two methods and some case studies of measurements inside theaters," *Energy Procedia*, vol. 78, pp. 1611–1616, 2015.

[63] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo, "Impulse response estimation for robust speech recognition in a reverberant environment," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 1668–1672.

[64] Architectural Institute of Japan, "Benchmark problems for acoustical parameters." [Online]. Available: http://news-sv.aij.or.jp/kankyo/s24/benchmark/

[65] J. Mourjopoulos and M. Paraskevas, "Pole and zero modeling of room transfer functions," *Journal of Sound and Vibration*, vol. 146, no. 2, pp. 281–302, 1991.

[66] M. Tohyama, T. Koike, and J. F. Bartram, "Fundamentals of acoustic signal processing," 2000.

[67] M. Tohyama, "Chapter 6 - room reverberation theory and transfer function," in *Acoustic Signals and Hearing*. Academic Press, 2020, pp. 111–140.

[68] N. Xiang and G. M. Sessler, *Acoustics, Information, and Communication: Memorial Volume in Honor of Manfred R. Schroeder*. Springer, 2014.

[69] 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.

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

[71] M. Barron, "Interpretation of early decay times in concert auditoria," *Acta Acustica united with Acustica*, vol. 81, no. 4, pp. 320–331, 1995.

[72] 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.

[73] D. J. Eaton, "Non-intrusive estimation of acoustic parameters from degraded speech," Ph.D. dissertation, 2015.

[74] P. Zahorik, "Direct-to-reverberant energy ratio sensitivity," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2110–2117, 2002.

[75] 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.

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

[77] M. Unoki, K. Sasaki, R. Miyauchi, M. Akagi, and N. S. Kim, "Blind method of estimating speech transmission index from reverberant speech signals," in *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE, 2013, pp. 1–5.

[78] S. Morita, M. Unoki, X. Lu, and M. Akagi, "Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 163–173, 2016.

[79] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr, 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.

[80] P. Kendrick, F. F. Li, T. J. Cox, Y. Zhang, and J. A. Chambers, "Blind estimation of reverberation parameters for non-diffuse rooms," *Acta Acustica united with Acustica*, vol. 93, no. 5, pp. 760–770, 2007.

[81] 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.

[82] P. Kendrick, "Blind estimation of room acoustic parameters from speech and music signals," Ph.D. dissertation, University of Salford, 2009.

[83] J. Wen, "Reverberation: models, estimation and application," Ph.D. dissertation, Imperial College London, 2009.

[84] J. Aldrich *et al.*, "RA fisher and the making of maximum likelihood 1912-1922," *Statistical science*, vol. 12, no. 3, pp. 162–176, 1997.

[85] S. Li, R. Schlieper, and J. Peissig, "A hybrid method for blind estimation of frequency dependent reverberation time using speech signals," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 211–215.

[86] F. Li and T. Cox, "Speech transmission index from running speech: A neural network approach," *The Journal of the Acoustical Society of America*, vol. 113, no. 4, pp. 1999–2008, 2003.

[87] P. Kendrick, T. Cox, Y. Zhang, J. Chambers, and F. Li, "Room acoustic parameter extraction from music signals," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006.

[88] 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.

[89] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, 2010.

[90] P. Kendrick, N. Shiers, R. Conetta, T. J. Cox, B. M. Shield, and C. Mydlarz, "Blind estimation of reverberation time in classrooms and hospital wards," *Applied Acoustics*, vol. 73, no. 8, pp. 770–780, 2012.

[91] 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)*. IEEE, 2018, pp. 591–595.

[92] S. Deng, W. Mack, and A. P. Habets, "Online blind reverberation time estimation using CRNNS," in *IEEE International Speech Communication Association (INTERSPEECH)*, 2020.

[93] P. P. Parada, D. Sharma, and P. A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4718–4722.

[94] P. P. Parada, D. Sharma, J. Lainez, D. Barreda, T. van Waterschoot, and P. A. Naylor, "A single-channel non-intrusive C50 estimator correlated with speech recognition performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 719–732, 2016.

[95] J. F. Santos and T. H. Falk, "Blind room acoustics characterization using recurrent neural networks and modulation spectrum dynamics," in *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*. Audio Engineering Society, 2016.

[96] H. Gamper and I. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 136–140.

[97] F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, "Exploring auditory-inspired acoustic features for room acoustic parameter estimation from monaural speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1809–1820, 2018.

[98] S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "A robust method for blindly estimating speech transmission index using convolutional neural network with temporal amplitude envelope," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1208–1214.

[99] H. Gamper, "Blind C50 estimation from single-channel speech using a convolutional neural network," in *22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020, pp. 1–6.

[100] D. Sharma, L. Berger, C. Quillen, and P. A. Naylor, "Non-intrusive estimation of speech signal parameters using a frame-based machine learning approach," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 446–450.

[101] F. F. Li and T. J. Cox, *Digital signal processing in audio and acoustical engineering*. CRC Press, 2019.

[102] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[103] T. Prego, A. A. Lima, Z. L. R., and S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *2015 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[104] D. Looney and N. D. Gaubitch, "Joint estimation of acoustic parameters from single-microphone speech observations," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 431–435.

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

[106] J. Tchorz and B. Kollmeier, "Estimation of the signal-to-noise ratio with amplitude modulation spectrograms," *Speech Communication*, vol. 38, no. 1-2, pp. 1–17, 2002.

[107] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[108] S. Morita, X. Lu, M. Unoki, and M. Akagi, "Method of estimating signal-to-noise ratio based on optimal design for sub-band voice activity detection," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 8, no. 6, pp. 1446–1459, 2017.

[109] M. Unoki, M. Toi, and M. Akagi, "Development of the MTF-based speech dereverberation method using adaptive time-frequency division," in *Proc. Forum Acusticum*, vol. 2007, 2005, pp. 51–56.

[110] M. Unoki, S. Morita, A. Miyazaki, and M. Akagi, "Preliminary study on blind estimation of room acoustic parameters in noisy reverberant environments," in *Proc. 12th Western Pacific Acoustics Conferences 2015 (WESPAC 2015)*, 2015, pp. 428–435.

[111] Architectural Institute of Japan, "Sound library of architecture and environment," in *Gihodo Shuppan Co., Ltd., Tokyo*, 2004.

[112] K. Kinoshita, M. Delcroix, S. Gannot *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research."

[113] T. Takeda, "Speech database User's manual ATR technical report," *TR-I-0028*, 1988.

[114] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017.

[115] N. J. Bryan, "Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1–5.

[116] P. Masztalski, M. Matuszewski, K. Piaskowski, and M. Romaniuk, "StoRIR: Stochastic room impulse response generation for audio data augmentation," *Proc. Interspeech*, 2020.

[117] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[118] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[119] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. The MIT Press, 2018.

[120] S. Duangpummet, J. Karnjana, M. Unoki, and W. Kongprawechnon, "Blind estimation of room acoustic parameters and speech transmission index using MTF-based CNNs," in *29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 181–185.

[121] S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Blind estimation of room acoustic parameters and speech transmission index based on the extended impulse response model," *Applied Acoustics*, vol. 185, no. 108372, pp. 1–2, 2022.

[122] Y. J. Choi, "Effects of the distribution of occupants in partially occupied classrooms," *Applied Acoustics*, vol. 140, pp. 1–12, 2018.

[123] M. Unoki, Y. Kashihara, M. Kobayashi, and M. Akagi, "Study on method for protecting speech privacy by actively controlling speech transmission index in simulated room," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 1199–1204.

[124] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[125] Human Language Technology Laboratory of National Electronics and Computer Technology Center (NECTEC, Thailand), "Thai speech corpus." [Online]. Available: www.nectec.or.th/corpus/

Publications

Journal paper

[1] Suradej Doungpummet, Jessada Karnjana, Waree Kongprawechnon, and Masashi Unoki, "Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response," *Applied Acoustics,* vol. 185, no. 108372, pp. 1-12, 2022.

International conference

[2] Suradej Doungpummet, Jessada Karunjana, Waree Kongprawechnon, and Masashi Unoki, "Blind Estimation of Room Acoustic Parameters and Speech Transmission Index using MTF-based CNNs," *29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland,* pp.181-185, 2021.

[3] Suradej Doungpummet, Jessada Karnjana, Waree Kongprawechnon, and Masashi Unoki, "Speech Privacy Protection based on Optimal Controlling Estimated Speech Transmission Index in Noisy Reverberant Environments," *28th European Signal Processing Conference (EUSIPCO), Amsterdam, the Netherlands,* pp.76-80, 2020.

[4] Suradej Doungpummet, Jessada Karnjana, Waree Kongprawechnon, and Masashi Unoki, "A Robust Method for Blindly Estimating S peech Transmission Index using Convolutional Neural Network with Temporal Amplitude Envelope," *Int. Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), China,* pp.1208-1214, 2019.

Domestic conference

[5] Suradej Doungpummet, Jessada Karnjana, Waree Kongprawechnon, and Masashi Unoki, "Room Acoustic Parameters Estimation using MTF-based CNNs," *in Acoustic Society of Japan, Spring meeting,* 2021.

[6] Suradej Doungpummet, Jessada Karnjana, Waree Kongprawechnon, and Masashi Unoki, "Study on Robust Method for Blindly Estimating Speech Transmission Index using Convolutional Neural Network with Temporal Amplitude Envelope," *in Engineering Acoustics, IEICE at Tohoku Univ.,* no.163 pp.47-52, 2019.

[7] Suradej Doungpummet, Jessada Karnjana, Waree Kongprawechnon, and Masashi Unoki, "Blind Estimation of Speech Transmission Index in Noisy Reverberant Environment using Deep Learning with Modulation Spectrum," *in Acoustic Society of Japan, Spring meeting,* 2019.