A Robust Method for Blindly Estimating Speech Transmission Index using Convolutional Neural Network with Temporal Amplitude Envelope

Suradej Duangpummet*‡§, Jessada Karnjana‡, Waree Kongprawechnon§, and Masashi Unoki*
* Japan Advanced Institute of Science and Technology, Japan
‡ National Science and Technology Development Agency, Thailand
§ Sirindhorn International Institute of Technology, Thammasat University, Thailand
E-mail: {suradej, unoki}@jaist.ac.jp, jessada.karnjana@nectec.or.th, waree@siit.tu.ac.th

***Abstract*—We have developed a robust scheme for blindly estimating the speech transmission index (STI) based on a convolutional neural network (CNN) with temporal amplitude envelope as features. When assessing the quality of acoustics in a room where there are people present, STI needs to be estimated without measuring the room impulse response (RIR) or using a modulation transfer function (MTF). This estimation can be problematic because a blind method based on the MTF has low accuracy when the stochastic models of RIR and the background noise are mismatched to real sound environments. We improve the accuracy of STI estimation in noisy reverberant spaces by using a CNN that takes the entire temporal amplitude envelope of an observed speech signal as its input. Simulations were performed to evaluate the proposed scheme and results showed that it can maintain the appropriate accuracy under various realistic room acoustic conditions with an average RMSE of 0.12 and correlation of 0.87. These results demonstrate that the proposed scheme can robustly and blindly estimate STIs in noisy reverberant environments.**

I. Introduction

Background noise and reverberation in common spaces such as banks, concourses, or restaurants interfere with hearing ability. Knowing the level of listening difficulty is essential for manipulating the speech intelligibility of listeners [1]. For enclosures, many objective indices and acoustic parameters have been proposed to evaluate the listening difficulty or intelligibility level, such as reverberation time (T_{60}), the clarity index (C_{80}), and the speech transmission index (STI) [2]. The STI as an IEC 60268-16 standard, which highly correlates with listening difficulty, is an objective index to assess the speech transmission quality in a given room [3], [4]. Calculating STI is based on the concept of the modulation transfer function (MTF), which describes room acoustic characteristics as a system transfer function in the modulation frequency domain by using a set of sine-wave modulated stimuli in different frequency bands [5], [6], [7]. In addition to deriving STI from the MTF, in the time domain, a room impulse response (RIR), which uses a brief impulse signal into a room, can be used.

The problem with MTF and RIR measurements is that they have to be done in sound fields where people are excluded. In common spaces, it is thus difficult and impractical to obtain STIs, so in everyday situations, a method for estimating STIs "blindly" without measuring MTF or RIR is required. Many blind estimation techniques have been proposed and can be categorized into two groups: one based on machine learning (ML) [8], [9], [10], [12] and the other on a deterministic approach [13], [14].

Early on in the development of ML-based methods, a multilayer perceptron network (MLP) was proposed to estimate STIs. The MLP used 14 data points of envelope spectra as features [8]. An improved MLP model with additional inputs, which were the features from the power envelope of a speech signal using principal component analysis, soon followed [9]. However, the accuracy of these methods when it comes to general pronunciation is poor due to the limitations of the features and MLPs. In the last decade, a major drawback of the ML-based methods—namely, that they require a huge amount of learning data—has been solved by using synthesized RIRs, and modern ML techniques are now being applied to the estimation of room acoustic characteristics. For example, long short-term memory (LSTM) has been used to estimate reverberation time T_{60} from modulation spectra [11], where the modulation spectrogram, which is extracted from a reverberant speech signal, is the input of the LSTM. Recently, a deep convolutional neural network (deep CNN) has been utilized to directly estimate STI from a raw reverberant speech signal by means of end-to-end model [12]. While this model has high accuracy under reverberant conditions, it has not yet been evaluated in noisy environments, and thus its robustness remains in question [12].

As for the deterministic approach, an MTF-based method has been previously proposed in which the RIR and background noise are assumed as stochastic models [13], [14]. The estimated RIR and estimated noise are used to derive the corresponding STIs. The MTF-based approach delivers a good performance without requiring massive amount of data, but the problem is that the mismatch between the models and real environments reduces the accuracy.

In this study, we incorporate the temporal amplitude envelope of observed speech signals, i.e., the basis of the MTF, into a CNN to resolve the robustness issue. The performance is then consistent regardless of noise and reverberant conditions.

II. Background

The previous MTF-based method uses an RIR model for estimating STI [14]. In a noisy reverberant environment, an observed speech signal $y(t)$ is assumed to be a convolution between an original signal $x(t)$ and RIR $h(t)$ plus a background noise $n(t)$, as.

$$y(t) = x(t) * h(t) + n(t), \tag{1}$$

where $*$ represents the convolution of two signals $x(t)$ and $h(t)$. Schroeder's RIR model is modified here into a generalized RIR model [6], [13], i.e.,

$$h(t) = a\, t^{(b-1)} e^{-\frac{6.9t}{T_R}} c_h(t), \tag{2}$$

where a is the gain factor, b is the order of the RIR, T_R is the reverberation time, and $c_h(t)$ is a carrier of white Gaussian noise (WGN). Then, a noise signal can be modeled as

$$n(t) = e_n(t) c_n(t), \tag{3}$$

where $e_n(t)$ is the temporal amplitude envelope of a signal and $c_n(t)$ is a WGN carrier. Then, the energy of the observed signal in (1), i.e., the power envelope, is approximated as

$$e_y^2(t) = e_x^2(t) * e_h^2(t) + e_n^2(t). \tag{4}$$

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 [5]. The MTF at a modulation frequency f_m, $m(f_m)$, is defined as

$$m(f_m) = \frac{\int_0^\infty h^2(t) e^{-j2\pi f_m t} dt}{\int_0^\infty h^2(t) dt}, \tag{5}$$

where $h(t)$ is the room impulse response (RIR). The MTF is represented by modulation frequency f_m, reverberation time T_R, the order of RIR b, and signal-to-noise ratio (SNR), and is defined as

$$m(f_m, T_R, b, \mathrm{SNR}) = \left[1 + \left(2\pi f_m \frac{T_R}{13.8}\right)^2\right]^{-\frac{2b-1}{2}} \left(\frac{1}{1 + 10^{-\frac{\mathrm{SNR}}{10}}}\right). \tag{6}$$

The RIR in (2) is obtained by estimating two parameters: T_R and b. Then, we can calculate STI from this estimated RIR.

T_R 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 [14]. 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

$$\mathrm{RMS}(T_R, b) = \sqrt{\frac{1}{2}\sum_{l=1}^{2} [|E_y(f_{m_l})| - m(f_d, T_R, b)^2]}, \tag{7}$$

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 (6). 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 by the following steps.

1) Calculate MTFs in seven octave bands. Let $m_k(F_i)$ denotes the MTF of a subband k of the octave filter bank (where the center frequencies range from 125 Hz to 8 kHz for $k = 1$ to 7) with the modulation frequency F_i. Note that F_i ranges from 0.63 Hz to 12.5 Hz for $i = 1$ to 14. The $m_k(F_i)$, from the RIR in (2), can be calculated by

$$m_k(F_i) = \frac{1}{\sqrt{\left[1 + \left(2\pi F_i \frac{T_R}{13.8}\right)^2\right]^{2b-1}}}. \tag{8}$$

2) Calculate signal-to-noise ratios (SNRs). For each k and i, the SNR, $N(k, i)$, is defined as.

$$N(k, i) = 10 \log_{10}\left(\frac{m_k(F_i)}{1 - m_k(F_i)}\right). \tag{9}$$

3) Calculate transmission indices (TIs). For each k and i, the TI, $T(k, i)$, is calculated by normalizing the corresponding SNR, $N(k, i)$.

$$T(k,i) = \begin{cases} 1, & \text{if } 15 < N(k,i), \\ \frac{1}{30}\left(N(k,i) + 15\right), & \text{if } -15 \le N(k,i) \le 15, \\ 0, & \text{if } N(k,i) < 15. \end{cases} \tag{10}$$

4) Calculate modulation transmission indices (MTIs). For each k, $\mathrm{MTI}(k)$ is the average of $T(k, i)$ for all i, i.e.,

$$\mathrm{MTI}(k) = \frac{1}{14}\sum_{i=1}^{14} T(k, i). \tag{11}$$

5) Calculate STI. The STI is a weighted average of $\mathrm{MTI(k)}$ for all k, where the weights $W(k)$ are distributed as follows: $W(1) = 0.129$, $W(2) = 0.143$, $W(3) = W(4) = 0.114$, $W(5) = 0.186$, $W(6) = 0.171$, and $W(7) = 0.143$.

$$\mathrm{STI} = \sum_{k=1}^{7} W(k)\mathrm{MTI(k)}. \tag{12}$$

Finally, the STI of a given room is represented by a number between 0 (bad listening) and 1 (excellent listening).

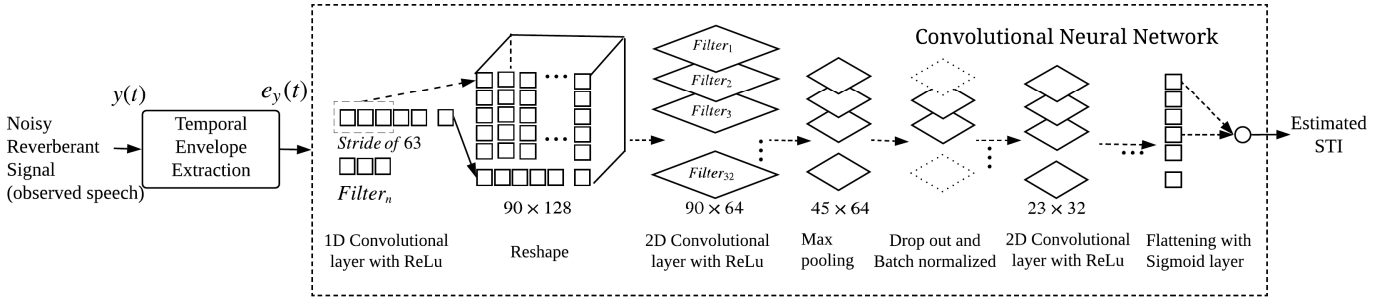

Fig. 1: Block diagram of the proposed method.

III. Proposed Method

As only information of the system is an observed speech signal, we assume the STI estimation as a blind deconvolution with regression problem. The STI is approximated from the temporal amplitude envelope of the observed speech signal. We propose the scheme that consists of the temporal envelope extraction and the CNN for estimating STIs, as shown in Fig. 1. The CNN performs the convolution operation of an output envelope signal with well-trained filters. The filters learn the relationship between the output envelope signals and their STIs. A significant feature based on the concept of the MTF, that is, temporal amplitude envelope of the observed speech signal is used as the feature of the CNN. Therefore, this method can provide a good accuracy under noisy and reverberant conditions. It means that the robustness of the STI estimator is improved.

A. Temporal Amplitude Envelope Extraction

Since noise and reverberation influence the shape of the envelopes of speech signals, an observed envelope can be used to describe the property of a speech transmission channel, which is a given room. In this study, instead of a few features of modulation spectrum, we utilize the signal of the entire temporal amplitude envelope along with their associated STIs for training the CNN. The temporal amplitude envelope of noisy reverberant speech signals, $e_y(t)$, can be extracted as

$$e_y(t) = \mathrm{LPF}\left[\left|y(t) + j \cdot \mathrm{Hilbert}(y(t))\right|\right], \qquad (13)$$

where $\mathrm{LPF}[\cdot]$ is a sixth-order Butterworth filter, which is IIR low-pass filters, with a cut-off frequency of 20 Hz, and $\mathrm{Hilbert}(\cdot)$ is the Hilbert transform. In speech perception, the significant modulation frequencies are between 1 and 16 Hz [15]. Thus, we can significantly reduce the model complexity by down-sampling the envelopes to 40 Hz. Then, the envelopes are normalized to a unit scale to avoid bias due to different amplitudes. Examples of speech signals and their temporal amplitude envelopes under different conditions are shown in Figs. 2 and 3, respectively.

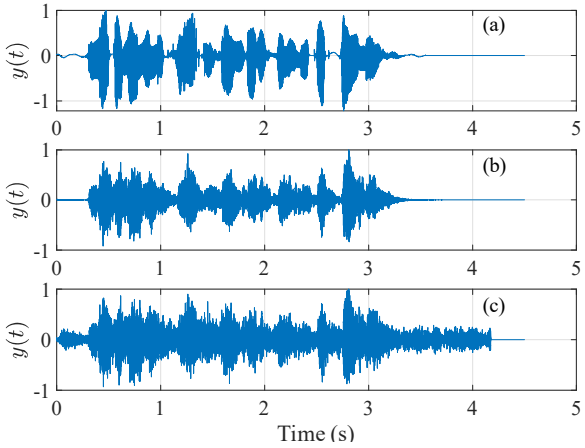

Fig. 2: Signals of (a) clean speech, (b) reverberant speech ($T_R = 0.43$ s), and (c) noisy reverberant speech (babble noise at SNR of 5 dB and $T_R = 0.43$ s).

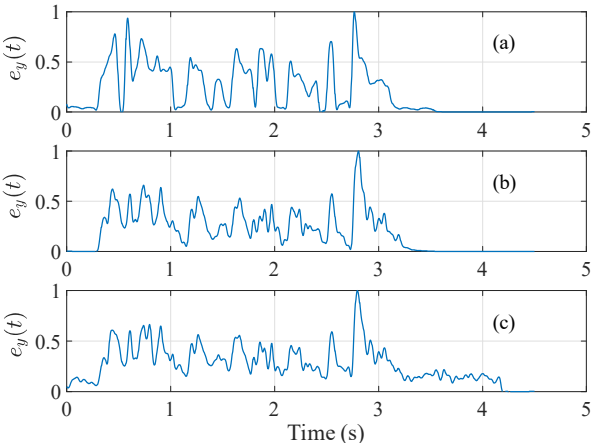

Fig. 3: Temporal amplitude envelope of (a) clean speech, (b) reverberant speech ($T_R = 0.43$ s), and (c) noisy reverberant speech (babble noise at SNR of 5 dB and $T_R = 0.43$ s).

B. Convolutional Neural Network

The CNN trained with observed reverberant envelopes under conditions of various noise types and levels can be used to determine the associated STIs. We assume a blind STI estimation as a blind deconvolution with a regression problem. The CNN performs the deconvolution operation of the observed temporal amplitude envelope and solves the regression problem. The CNN consists of three convolutional layers and complementary layers, as shown in Fig. 1

In the design of 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 [12]. The final two convolutional layers apply two-dimensional convolutional filters to perform a regression task. From a mathematical viewpoint, high-dimensional spaces expand the potential for problem-solving. Similarly, in neuroscience, the middle layer of the perceptron model contains more neurons than the other layers [16]. Thus, here we assign the two middle layers is a higher number of filters: 32 and 16, respectively.

For complementary layers, a pooling unit accompanies a convolution layer for down-sampling as well as keeping an invariance of the input. Here, max pooling, which is a non-linear operation, corrects the highest value from the neighbors. The outputs are then passed through an activation function, which is a rectified linear unit (ReLU). The ReLU function has been designed 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 utilize a batch-normalization to scale the values to 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, called a dense layer, estimates output by a sum of the products between the vectors and their weights, so that the estimated STI as the output can be presented as

$$\hat{\mathrm{STI}} = \mathrm{SIGM}\left(\sum_{i=1}^{j} W \otimes a_i + b\right), \tag{14}$$

where $\hat{\mathrm{STI}}$ is an estimated STI, SIGM is a sigmoid function, W is a weight matrix, a_i is an input from a previous layer for i to the total elements j, "$\otimes$" is the 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 I.

TABLE I: Convolutional neural network architecture.

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

IV. Experiments and Evaluations

To develop and evaluate this robust estimator, we used noisy reverberant speech signals under different conditions and their associated STIs as our datasets. These corresponding STIs as the ground truth are calculated from RIRs and noise levels as in (6). The calculated STIs and the envelopes, which are extracted from observed speech signals, are utilized for training the CNN. Then, the results of estimated STIs are analyzed on the basis of two statistical metrics: root-mean-square error (RMSE) and correlation coefficient ρ.

A. Data Collection

Data are collected and allocated to three datasets: training, validation, and testing. These include $29,000$ four-second excerpts of noisy reverberant speech signals re-sampled at the rate of $16,000$ samples per second.

The training set and validation set are generated from RIRs, anechoic speech signals (clean signals), and noise signals. The RIRs consists of 43 realistic RIRs from the SMILE2004 dataset and one hundred RIRs synthesized using the image method [17], [18]. The 43 RIRs include different acoustic parameters (i.e. T_R from 0.38 to 3.62 second) as reported in [14]. Likewise, the synthesized RIRs are generated to cover most conceivable possibilities of STIs from various room properties. The distribution of this dataset is shown in Fig. 4. One hundred clean speech signals are randomly selected from the CSTR corpus [19]. Note that these speech signals are English sentences uttered by people of various ages, genders, accents, and regions. Then, the noisy reverberant speech signals are obtained by convolving RIRs with clean speech signals and adding noises as in (1). These added noises are a dataset from the NOISEX-92 corpus and include white noise, pink noise, factory noise, and babble noise [20]. The noise is cut so that its length is the same as that of the speech signal. Then, we add these noise signals to the reverberant signals, so that the SNR values are 5 dB, 20 dB, and infinity dB (i.e., without noise).

The test dataset consists of unknown utterances in unseen environments. These signals are the convolution between ten Japanese speech signals (five male and five female speakers) and seven RIRs plus noises [21]. These noise signals are ambient noise, fan noise, and babble noise. Note that these noises are recorded in the same room as the measured RIRs from the ACE corpus [22]. The babble noise is composed of the simultaneous speaking of ten talkers.

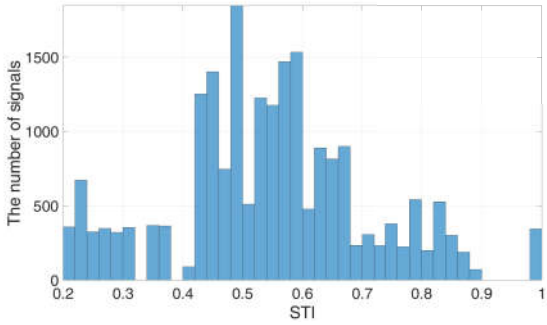

Fig. 4: Distribution of STIs in training dataset.

B. Experimental Setup

We used MATLAB for extracting the temporal envelope and Python for implementing the CNN. The optimal parameters of the CNN are trained on the Google Colaboratory for one hundred iterations. This platform is a cloud service platform run on a GPU (Tesla K80) that can complete the training process in one hour. Note that the maximum iteration is one hundred iterations, and the batch optimization is the size of 128 samples.

C. Evaluation Metrics

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

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\hat{\mathrm{STI}}_n - \mathrm{STI}_n\right)^2}, \tag{15}$$

where $\hat{\mathrm{STI}}_n$ is the estimated STI, $\mathrm{STI_n}$ is the ground truth calculated from RIR and SNR as in (5), n is an index of the observed signal, and N is the total number of signals. The second evaluation metric, i.e., correlation (ρ), is defined as

$$\rho = \frac{\sum_{n=1}^{N}(\hat{\mathrm{STI}}_n - \overline{\hat{\mathrm{STI}}_n})^2(\mathrm{STI}_n - \overline{\mathrm{STI}}_n)^2}{\sqrt{\sum_{n=1}^{N}(\hat{\mathrm{STI}}_n - \overline{\hat{\mathrm{STI}}_n})^2 \sum_{n=1}^{N}(\mathrm{STI}_n - \overline{\mathrm{STI}}_n)^2}}, \tag{16}$$

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

D. Results

In our experiments, we evaluate the proposed method using two datasets respectively assigned as known and unknown noisy reverberant environments. The baseline is the MTF-based method to determine whether the proposed method can estimate STIs in any noisy reverberant environments. The simulation of reverberant environments without noise is conducted, as shown in Fig 5. The estimated results in the 43 rooms appear that our estimated STIs slightly scatter from the dashed line of the ground truth. Our model has a little lower accuracy than the baseline of the MTF-based method, so this condition is analyzed in the discussion section. However, to evaluate our method for more realistic environments, the blind STI estimation in noisy and reverberant conditions are then examined.

First, in known noisy reverberant environments, the results of estimated STIs from reverberant speech signals with different noise types, which are WGN, pink noise, babble noise, and factory noise, are shown in Fig. 6. The horizontal axis indicates STIs ground truth, and the vertical axis indicates the estimated STIs. Ideally, the estimation results should be close to the ground truth STIs, which is the diagonal dashed line. The color symbols indicate two SNR levels. The blue color stands for SNR of 20 dB, and the red color stands for SNR of 5 dB. The symbols "$\circ$" and "$\times$" correspond to the estimation methods. The comparison with the baseline in terms of RMSEs and correlation coefficients is shown in Table II. These results demonstrate that the proposed method had lower RMSE and higher correlation than the baseline in all noise types.

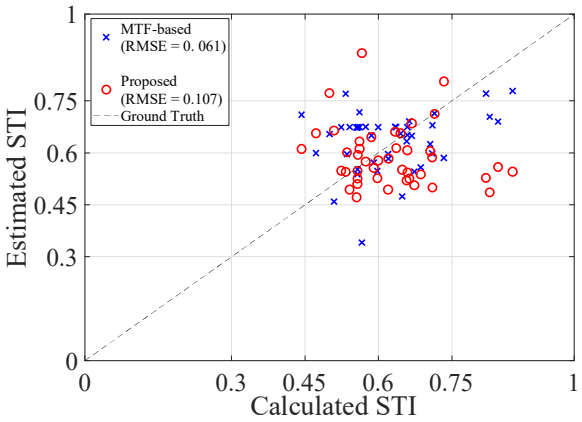

Fig. 5: Estimated STIs from reverberant speech signals.

Second, to evaluate whether the proposed model is not over-fitting and memorizing, the test dataset is utilized, which is unknown utterances and unknown environments with background noise. These background noise, including ambient noise, fan noise, and babble noise, represent the realistic background noise. The estimating results in terms of RMSE and ρ are summarized in Table III. We can see that the proposed model provided low RMSEs at an average of $0.09, 0.08$, and 0.14 as well as high correlations of 0.85 in the seven reverberant rooms with three real noise conditions. Therefore, in these noisy reverberant environments, the proposed method outperforms the MTF-based method in robustness.

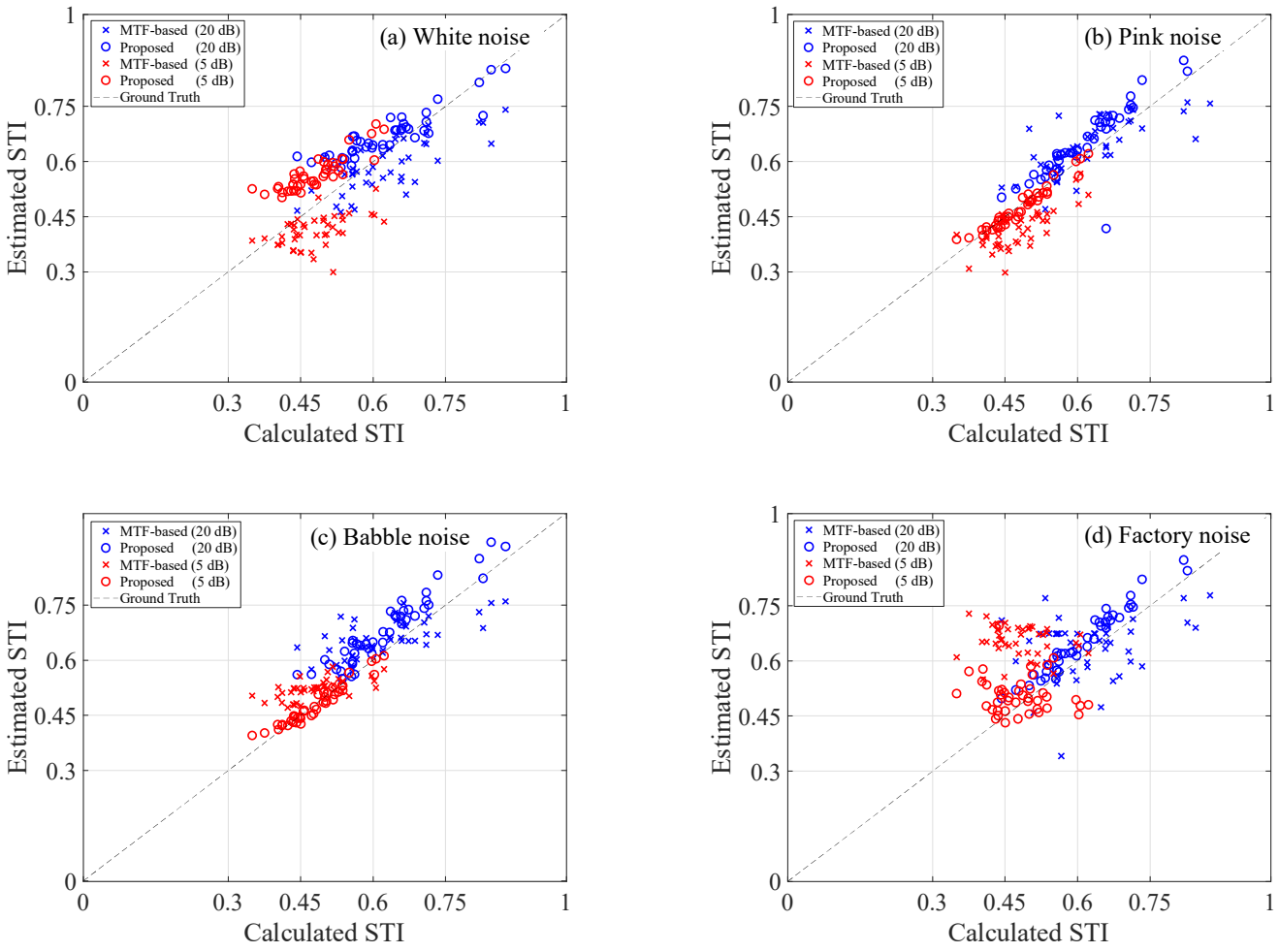

Fig. 6: Estimated STIs from observed speech signals under background noise and reverberant conditions where the four noise types are: (a) white noise, (b) pink noise, (c) babble noise, and (d) factory noise.

TABLE II: Estimated STIs under various conditions from SMILE corpus in the metrics of RMSE and correlation (ρ) [17].

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

TABLE III: Estimated STIs of speech signals in RIR and background noise from acoustic characteristic corpus [22].

Noise	Method	RMSE 20 dB	RMSE 5 dB	ρ
Ambient	MTF-based	0.14	0.35	0.64
	Proposed	0.07	0.11	0.86
Fan	MTF-based	0.17	0.18	0.73
	Proposed	0.08	0.09	0.79
Babble	MTF-based	0.18	0.26	0.63
	Proposed	0.13	0.15	0.86

V. Discussion

There are some advantages, limitations, and issues of this work we would like to discuss. First, in the noise-free experiment, there are a few reverberant speech signals without additive noise in the training set. Hence, the proposed model has a little lower accuracy than the baseline of the MTF-based method. Second, for the noisy condition, since babble noise and factory noise are non-stationary noise, these noise types are different from the noise model of the MTF-based method. Hence, the model mismatch causes inadequate accuracy, whereas the CNN learned from various noise types can overcome this problem. The accuracy of our method 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 reverberant speech signals with non-stationary noise might need for training separately to reduce the outliers. Furthermore, with that said, our proposed not only satisfies the accuracy and robustness, but also has advantages over the existing methods in additional aspects, as follows.

First, the proposed model can reduce the operation time from the conventional STI measurement time of 15 minutes [7]. Our method, which uses the envelope of a short four-second speech segment, can provide accuracy comparable to that of the conventional method [4]. 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. On the other hand, the ahead-of-time optimal filters of the CNN can calculate STIs promptly.

However, there is one concerning issue in this work that we should point out here. 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 new and random data. The generalized model needs to compensate for the trade-off between high accuracy and model complexity. For instance, we found that the longer envelope input the CNN takes (i.e., from one second to four seconds), the more accurate the performance of the CNN. In this study, we thus empirically propose the CNN architecture to maintain the acceptable performance. However, the model can be fine-tuned so as to deliver an even better performance.

VI. Conclusion

We have presented a scheme to improve the robustness of blind STI estimation in noisy reverberant environments. Previously, even though the MTF-based method can estimate STI according to the MTF concept, it suffers a mismatch between the stochastic models and realistic environments, thus resulting in unsatisfactory accuracy under some noise conditions. To resolve this issue, we developed a robust scheme that incorporates the entire temporal amplitude envelope into a CNN. The CNN is trained by the temporal envelope features, which are temporal amplitude envelopes of observed speech signals and their associated STIs under various noisy reverberant conditions. We carried out simulations to evaluate the proposed model using observed speech signals in known and unknown reverberant environments with many noise types. The results showed that the proposed method delivers good accuracy with the average RMSE of 0.12 and the correlation of 0.87, thus demonstrating that this method is robust against reverberation and background noise without the need for retraining. Additionally, an entire temporal amplitude envelope signal is suitable features for training the CNN, so the proposed method succeeds in blindly estimating STI in general public room acoustics. In the future, we will apply our scheme to the estimation of other room-acoustic parameters and indices and extend it for use in speech enhancement or speech privacy control algorithms.

Acknowledgment

This work was supported by a grant in the Secom Science and Technology Foundation, JST-Mirai Program (Grant Number: JPMJMI18D1), SIIT-JAIST-NSTDA Dual Doctoral Degree Program, and Thammasat University Basic Research Grant.

References

[1] 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," *APSIPA ASC*, pp. 1199-1204, 2017.

[2] H. Kuttruff, "Room acoustics 5th ed.," *CRC Press,* 2016.

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

[4] International Electrotechnical Commission et al., "IEC 60268-16 (2003)," Sound system equipment-Part, vol. 16, 2003.

[5] T. Houtgast and H. J. M. 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.

[6] MR. Schroeder, "Modulation transfer function. Definition and measurement," *Acustica,* vol. 4 no. 3, pp. 179-82, 1981.

[7] 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, 1980.

[8] F. F. Li and T. J. Cox, "Speech transmission index from running speech: A neural network approach," *J. Acoust. Soc. Am.*, vol. 113, 2003.

[9] 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.

[10] P. Kendrick, T. J. Cox, Y. Zhang, J. A. Chambers, and F. F. Li, "Room acoustic parameter extraction from music signals," *in Proc. ICASSP,* 2006.

[11] J. F. Santos and T. H. Falk, "Blind room acoustics characterization using recurrent neural networks and modulation spectrum dynamics," *In Dereverberation and Reverberation of Audio, Music, and Speech,* 2016.

[12] P. Seetharaman, G. J. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," IEEE International *in Proc. ICASSP,* pp. 591-595, 2018.

[13] M. Unoki, K. Sasaki, R. Miyauchi, M. Akagi, and N. S. Kim, "Blind method of estimating speech transmission index from reverberant speech signals," *Proc. EUSIPCO2013,* pp. 1-5, 2013.

[14] 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.

[15] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication,* vol. 28, 1999.

[16] M. Ito, "The cerebellum and neural control," *Raven Press*, 1984.

[17] K. Kawai, K. Fujimoto, T. Iwase, H. Yasuoka, T. Sakuma, and Y. Hidaka, "Development of a sound source database for environmental/architectural acoustics: Introduction of SMILE 2004," *in Proc. ICA,* 2004.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.,* vol. 65, no. 4, 1979.

[19] C. Veaux, J. Yamagishi, K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, " University of Edinburgh. The Centre for Speech Technology Research, 2017.

[20] A. Varga and H. J. M. 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.

[21] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," 2017.

[22] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of Room Acoustic Parameters: The ACE Challenge," *IEEE/ACM Trans. Audio Speech Lang. Process.,* vol. 24, no. 10, pp. 1681-1693, 2016.