Estimating Reverberation Time by a Function of Intrusive Speech Intelligibility Measures

Maxim de Groot[1]

Supervisor(s): Jorge Martinez Castaneda[1]**, Dimme de Groot**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2024

Name of the student: Maxim de Groot
Final project course: CSE3000 Research Project
Thesis committee: Jorge Martinez Castaneda, Dimme de Groot, Sole Pera

***Abstract*—A Room Impulse Response (RIR) is a mathematical model for sound propagation in a room. Estimating RIR parameters such as the reverberation time (T60) allows Automatic Speech Recognition (ASR) systems to adapt to reverberation in input signals by changing their behavior based on these estimates. Currently, machine learning techniques provide the most accurate T60 estimations. We propose a novel methodology by using intrusive Speech Intelligibility Measures (SIMs) beyond their traditional application. In this study we utilize SIIB, SIIB$^{\text{Gauss}}$, STOI and ESTOI as SIMs. For each SIM we find a best fit curve with respect to the reverberation time (T60) using a statistical approach. The statistical analysis is applied on simulated RIRs obtained by using the Image Source Method. The estimator for SIIB$^{\text{Gauss}}$ achieves the lowest Mean Squared Error of 0.353 on simulated data. Although this does not outperform state-of-the-art models, we offer recommendations for possible improvements. Preliminary experiments suggest that enhancing noise robustness is crucial and that the estimators could be generalized to real-world scenarios. However, further research is necessary to confirm this.**

***Index Terms*—Reverberation time, speech intelligibility measure, room impulse response**

I. Introduction

In today's world, Automatic Speech Recognition (ASR) systems are essential for human-computer interaction. These systems typically do not perform well when distortions such as reverberation are present in the input signal [1]. Historically, ASR systems' robustness to reverberation has been improved by signal processing and dereverberation techniques [2]. More recently Deep Neural Networks (DNNs) have been successfully used for this purpose [3]. This approach is however not always desirable due to the large amount of training and validation data required. Alternative approaches applying adaptation schemes depending on estimated room acoustic properties have also shown promising results [4].

The acoustic properties in a given space, including those related to reverberation can be described by the Room Impulse Response (RIR). The room impulse response can be seen as a filter representing those acoustic properties. When this filter is convolved with an input signal, it simulates that signal in the conditions the RIR was recorded in [5]. Example acoustic properties that influence reverberation are:

- Room dimensions and geometry
- Absorption coefficients of surfaces
- Speaker and receiver locations

A particularly interesting acoustic property of a room closely related to reverberation is the reverberation time or T60 which is the time it takes for the sound energy to decay by 60 dB. This notion of a characteristic time for sound to die out in a room was first mentioned by Wallace Clement Sabine. He found an empirical equation to estimate the T60 in a room, given its acoustic properties. Formula (1) gives Sabine's equation where T is the reverberation time in seconds, V is the volume of the room in m^3 and A is the total absorption in sabins [6].

$$T = \frac{0.161V}{A} \tag{1}$$

Sabine's formula is however not always applicable as the room's acoustic properties might not be known. Traditionally the estimation of the T60 in a room given only recordings of audio signals in that room has been done by employing signal-processing techniques, but currently this is mostly done by DNNs as they are often more accurate [7].

Instead of using machine learning to estimate the T60, we propose a methodology based on statistics which exploits the relation between speech intelligiblity and reverberation. Speech intelligibility refers to how well a listener can identify spoken words, often evaluated using formal listening tests. It is established that when the T60 increases, speech intelligibility decreases [8]. Alternative to using costly and time consuming formal listening tests, speech intelligibility can be objectively estimated by various Speech Intelligibility Measures (SIMs). The traditional SIMs such as the Speech Intelligibility Index (SII) are known not to perform well when facing reverberation in the input signal, while others have been designed to take reverberation into account [9]. This supposed relation between reverberation and certain SIMs brings forth the research question this paper aims to answer:

> "Can intrusive SIMs be used to estimate the T60 in speech signals by using a statistical estimator?"

This unexpected, yet innovative application of SIMs contributes to the field of ASR by giving new means to estimate the T60 in audio, with more robust ASR systems as a final goal, given that adaptation schemes can be applied using these estimates. Literature does not show wide application of statistical estimators for estimating the T60 using SIMs. This paper discusses how we fill this research gap and is structured as follows. First, background information required to understand and justify the methodology is given in Section II. Secondly, the methodology to obtain the estimator is discussed in Section III. Thirdly, Section IV highlights efforts made to ensure ethical and responsible research. Following this, Section V discusses results obtained by performing the experiments described by the methodology. Lastly, an interpretation of the results is given along with a discussion about limitations and recommended future work in Section VI.

II. Background

Having established the hypothesis that SIMs can be used to estimate the T60, we now discuss background information to support the methodology based on statistics by explaining the relevance of statistical methods and delving into SIMs. Previous work has shown that DNNs can be used to estimate RIR parameters such as the T60 [10]. However, the usage of DNNs is not always desirable as they face limitations such as their computational complexity and "black-box" principle. [11]. Statistical estimators are tailored to a specific problem, meaning they can outperform DNNs in certain cases [12]. The possibility of finding an appropriate statistical estimator is further motivated by the fact that the converse has previously been applied to find a formula to calculate the Speech Transmission Index (STI) SIM using the T60 in [13]. This was also done by Tang and Yeung in [14], where they found the following best fit for the STI as a function of the T60 as seen in Formula (2).

$$s = (a + b\exp(-t))^2 \tag{2}$$

In this formula, s represent the STI value (between 0 and 1), t represents the T60 value in seconds, and a and b are optimized positive constants based on a curve fit. In their experiment, these optimized values are $a \approx 0.64$ and $b \approx 0.38$.

As the relation between the STI and the T60 is already well established, it is not investigated in this work. However, many other different SIMs exist and are all developed with a specific purpose, such as estimating speech intelligibility in certain noisy environments. In an ideal situation non-intrusive SIMs would be used for the estimator as they do not require the clean speech signal and would thus allow to estimate the T60 of this signal. However, this study only considers intrusive SIMs, which require both the clean speech (no reverberation) and the reverberant speech signal. The primary reason for this limitation is that the relation between reverberation and intrusive SIMs is better established than the relation between reverberation and non-intrusive SIMs. Additionally, intrusive measures generally show a higher correlation to speech intelligibility than non-intrusive measures [15], which makes intrusive SIMs more interesting when investigating the relation between reverberation and speech intelligibility.

Intrusive SIMs are all developed for different purposes, but can generally be seen as functions which take the clean and reverbed speech signals as input and produce a number objectively describing speech intelligibility as an output. In [9] an overview of intrusive SIMs is given. From this list of SIMs a subset of SIMs is chosen to investigate for the estimator. The remainder of this section justifies the subset of SIMs chosen.

In order for a measure to be suitable for the estimator, it should first have a hinted or established relation to noise and specifically reverberation. When considering their original purpose, it is generally desirable for SIMs to be robust to all kinds of noise, but for this specific case, an ideal measure is sensitive to reverberation but robust to other types of noise so that the estimator's performance does not suffer too much under these types of noise. Secondly, SIMs with a higher correlation to speech intelligibility are preferred to further support the relation between T60 and speech intelligibility. Lastly, the SIMs should be widely applicable and available to produce a widely applicable and available estimator.

Considering Table I, a list can be established with the most generally noise robust SIMs which perform poorly in reverberant conditions as: STOI, ESTOI, SIIB, SIIB$^{\text{Gauss}}$, and sEPSM$^{\text{corr}}$. We do however not investigate sEPSM$^{\text{corr}}$ as it is not widely applied and is only implemented in Matlab, which would not lead to an accessible estimator due to the requirement of a Matlab license. The other measures are also not discussed further in this work. In the methodology section we elaborate on the calculation procedure for each SIM chosen (STOI, ESTOI, SIIB and SIIB$^{\text{Gauss}}$).

TABLE I: An overview of SIMs' robustness to noise mostly obtained from [9]. Additional references are mentioned if the information is not obtained from [9]. This overview is used to defend the chosen SIMs used for the estimator, as it is largely based on robustness to noise and sensitivity to reverberation. The measures used in this paper are underlined.

Measure	**Noise Robustness**
SII	Performs well for stationary additive noise, but poorly for modulated noise sources.
STI	Similar to SII.
HEGP	HEGP can only quantify distortion caused by additive noise signals.
CSII-mid	CSII is a generalization of the SII that can be applied to a wider range of distortions.
HASPI	Adapted to listener intelligibility data for various processing conditions, including noise suppression, additive noise, reverberation, nonlinear distortion, and hearing aid processing. [16].
NCM-BIF	Especially strongly correlated with intelligibility for speech signals exposed to post-processing enhancement.
QSTI	Good performance for non-linear distortions.
STOI	Works well for most noise forms, but performs poorly for modulated noise sources. Also fails to account for reverberation [17].
ESTOI	Developed to perform better for modulated noise sources compared to STOI.
MIKNN	Noise robustness is not documented well enough to hint towards a relationship between reverberation time and this measure [18].
SIMI	Able to reliably estimate the average intelligibility of speech signals containing by stationary and non-stationary noise. No clear relation with reverberation documented [19].
SIIB	Tested for different types of noise. However, it is suggested that reverberant channels introduce undesired statistical dependencies in its calculation. [20].
SIIB$^{\text{Gauss}}$	The noise robustness is specified but performance is said to be similar to SIIB so we assume noise robustness to be similar too.
sEPSM$^{\text{corr}}$	Accounts for the effects of stationary and fluctuating noise interferers as well as for various forms of non-linear distortions, but fails to account for the effects of reverberation [17].

III. Methods

Up until now the hypothesized relation between SIMs and the T60 has been discussed, now we propose a methodology to provide support that this relationship exists. A full overview of the methodology can be seen in Figure 1, which is discussed in detail in the remainder of this section.

A. Obtaining Input Data

Since the SIMs chosen are intrusive, both clean and reverberant speech data is required in their calculation. The clean speech data is taken from the EARS Dataset [21], which has an Attribution-NonCommercial 4.0 International license. The choice of this dataset as the clean speech dataset is motivated by its anechoic nature and high quality. The dataset also contains speech fragments sampled at 48 kHz from 107 different speakers from diverse backgrounds. We deem this to be a sufficient amount of different speakers for the purpose of this estimator as is elaborated on in Section IV. The audio dataset is pre-processed by filtering out audio fragments which do not contain speech (such as coughing) and randomly splitting the remaining speech signals into a training and

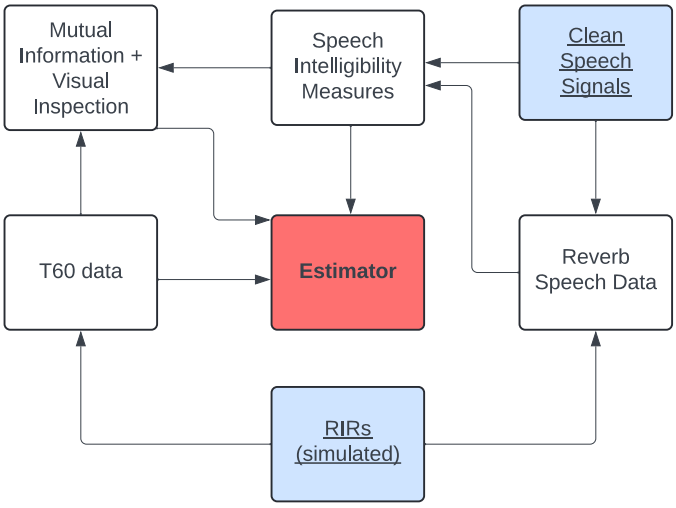


Fig. 1: Schematic methodology to construct an estimator for T60 using SIMs. Inputs are indicated in blue (underlined), while outputs are indicated in red (bold).

validation dataset with a 50/50 split, both datasets comprise 150 speech samples. In order to obtain the reverberant speech data, a dataset of RIRs is required.

Simulated RIRs of shoe-box rooms are used as large real RIR datasets with diverse T60 values are not widely available. The RIR dataset is simulated using the Image Source Method proposed in [22]. We use Habets' Matlab implementation of the Image Source Method, which has an MIT license [23]. For both the training and validation datasets we generate 2000 samples. We generate the entire dataset of 4000 samples in two batches. The first batch is for low T60s values, while the other is for high T60 values. The two batches are merged before the training and validation split. Desired T60 values are drawn from a uniform distribution with a minimum value of 0.1 s and a maximum value of 2 s in steps of 0.01 s. A uniform distribution is chosen in an attempt to make the estimator perform equally well for all values in the given range. For each desired T60 value, room dimensions are chosen with a baseline of 1.5 m and adjusted exponentially with the desired T60 and with a random deviation of at most positive or negative 0.5 m. This randomness adds more variation to the data. The final parameters required to simulate the RIRs are the sound absorption coefficients of all surfaces. Using the inverse of Formula (1) the average sound absorption coefficient in the room can be obtained. Absorption coefficients are chosen so that they alternate between positive and negative values. Random deviation is also added, while still respecting the constraints that they sum up to the average coefficient and that no reflection coefficient is larger than or equal to 1. Alternating between positive and negative absorption coefficients results in the mean amplitude of the RIR being 0 when excluding the direct path, similar to real RIRs.

Note that Sabine's formula is simply an estimation and thus not completely accurate, as the T60 also depends on other parameters like source and microphone positions in the room. For each room configuration consisting of the room dimensions and absorption coefficients, each combination of 20 microphone and 5 sound source locations in the room are drawn from a uniform distribution. The NEN-EN-ISO 3382-2:2008 standards for measuring reverberation time also state constraints on source and microphone placements for consistent results [24]. Both should be at least 0.5 m away from any surfaces in the room and no closer than 0.2 m to each other. These constraints are also respected while generating our RIR database. The RIRs generated by this method have a sampling frequency of 48 kHz.

Since Sabine's formula is not fully accurate, the T60 of the generated RIRs do not exactly match the desired T60 value used as input to the formula. For this reason the T60 values used as ground truth for the estimator are calculated based on the generated RIRs. The T60 are estimated using the rt60.measure_rt60 function of the pyroomacoustics library [25]. This function applies Schroeder's approach which uses backward integration to estimate the reverberation time given a RIR [26]. Schroeder's method struggles under certain noise conditions but is widely used in practice. When the background noise level is far outside of the measured decay range, Schroeder's method provides estimates of the T60 that we consider to be accurate enough for this work [27]. Extrapolating the T60 from the T30 (same as T60 but for 30 dB) gives us more consistent results, thus this is the approach used to obtain the ground truth T60. Especially for high T60 values this estimation method can sometimes fail, which is when the T20 (20 dB) or T15 (15 dB) values are tried in that order to obtain the ground truth T60 values by extrapolation.

Given the generated RIRs and their ground truth T60 values, the final type of input data, the reverberant speech, can be obtained by convolving each RIR with a random sample from the clean speech dataset using the fast Fourier transform. Now for each reverbed signal, the corresponding T60 value and clean speech signal are known.

B. Calculating Speech Intelligibility Measures

With the clean and reverbed speech signals we can calculate the intrusive SIMs. Below we discuss the implementations of the Short-Time Objective Intelligibility Measure (STOI), the Extended STOI (ESTOI), Speech Intelligibility in Bits (SIIB), and SIIB Gaussian ($\text{SIIB}^{\text{Gauss}}$) used in this paper.

STOI is introduced by Taal et al. in [28] and calculated using a DFT-based time-frequency-decomposition, with computational efficiency and high correlation with speech intelligibility being its primary benefits. Unlike STOI, ESTOI does not assume mutual independence between frequency bands. Additionally, ESTOI incorporates spectral correlation by comparing spectrograms of 400 ms of the distorted and the clean speech signals [19]. In this paper we use the pystoi Python library by Pariente [29] as an implementation for calculating STOI and ESTOI. The pystoi library has an MIT license.

SIIB takes a different approach as it is a measure rooted in information theory introduced by Kuyk et al. [20]. SIIB estimates the amount of information in common between a speaker and a receiver in bits per second. In its calculation, it estimates information rate between both the environment and speech production channels. Statistical dependencies within

the input vectors are undesired when estimating information rate in its calculation. For this purpose the Karhunen-Loève transform is applied. However, this does not guarantee compensation for statistical dependencies introduced by reverberant channels. The SIIB algorithm defines distortion using a KNN mutual information estimator, while SIIB$^{\text{Gauss}}$ utilises the information capacity of a Gaussian channel and is an order of magnitude faster in its calculation than SIIB [9]. We use a Python implementation of SIIB and SIIB$^{\text{Gauss}}$ in the form of the pySIIB library [30], which is a translation of the original Matlab code. This library has a GPL-3.0 license.

After calculating the measures using the pySIIB and pystoi libraries, the results are normalized by the intelligibility score of the clean speech signal. The purpose of this is to account for differences in intelligibility between the different clean signals. This normalization procedure has the additional benefit of providing values between 0 and 1, so that different estimators can more easily be compared.

C. Mutual Information

Using mutual information, we can see how much information the SIMs and T60 have in common [31]. Mutual information is a metric rooted in information theory which unlike correlation coefficients, can also identify non-linear relations. There are several ways to calculate the mutual information $I(X;Y)$ between two random variables X and Y, we use Formula (3) which is based on entropy. The entropy for a random variable X is given by $E(X)$. The work in this paper uses the mutual_info_regression function from the scikit learn Python package [32] to calculate mutual information, which estimates entropy from k-nearest neighbors distances. This methodology to calculate mutual information based on entropy is based on descriptions from [33] and [34] according to the function's documentation. A value of 0 for mutual information means that there is no shared information between X and Y. Since the mutual information in this study is not known to be bounded from above, many different interpretations of values other than 0 exist. In this work we use it is to compare different SIMs to find out which have the highest shared information with the T60.

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3}$$

D. Constructing the Estimators

For each SIM we construct an estimator for the T60 by considering the T60 as a function of SIMs. Based on the scatter plots of each SIM and T60, a possible shape of the function or curve can be determined (e.g. exponential). This function has specific parameters to be optimized. There are many optimization approaches such as maximum likelihood estimation. However, this requires the knowledge of a (log) likelihood function, which can be complex to find [35]. Instead of this we use a simpler approach where we minimize an objective function. In the case of this study the Mean Squared Error (MSE) is used as objective function. The MSE for ground truth data y and estimated data $\hat{y}$ for a dataset of size n where y_i corresponds to the i-th datapoint in y (similar for $\hat{y}_i$) is given by Formula (4).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

Alternative objective functions such as the Mean Absolute Error (MAE) may also be used, but the MSE is regarded as a suitable choice for this study as it is sensitive to strong outliers of the estimator due to the square operation. Additional motivation for the usage of the MSE is its usage as an evaluation metric in this study, as is elaborated on in the next subsection. The minimisation process of the MSE is done by applying the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. Advantages of this commonly used minimisation approach are fast convergence and high quality predictions as claimed by Xue et al. [36]. It is however also claimed that BFGS does not always perform well with many outliers present in the input data. We postulate that the usage of the MSE as objective function partially mitigates this disadvantage of the BFGS algorithm.

In this study two types of functions are considered to fit to the data X where x represents a value in X. The first is a simple hyperbolic rational function $f(x)$ given by Formula (5), where the parameter to be optimized is a. The other function is an inverse shifted exponential function $g(x)$ given by Formula (6), with parameters λ and θ to be optimized. To fit the functions, initial parameters must be given which can influence results, we chose: $a = 1$, $\lambda = 1$ and $\theta = 0$.

$$f(x) = \frac{a}{x} \tag{5}$$

$$g(x) = \theta + \lambda(\log(\frac{1}{\lambda}) - \log(x)) \tag{6}$$

E. Estimator Evaluation

To facilitate the comparison to machine learning models, our estimator is evaluated using evaluation metrics often utilised for machine learning regression tasks. According to Botchkarev [37], the three most commonly used metrics for these tasks are the MSE mentioned earlier, the MAE, and the Mean Absolute Percentage Error (MAPE). Since the MAPE values of the models the estimator is compared to are not known, it is left out of consideration for this study. The calculations of the MSE and MAE for ground truth data y and estimated data $\hat{y}$ for a dataset of size n where y_i corresponds to the i-th datapoint in y (similar for $\hat{y}_i$) are given by Formulas (4) and (7) respectively.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

We compare our model to two different state-of-the-art blind estimation deep learning approaches. The first is by Prates et al. [38], henceforth reffered to as Prates-DNN. The second is a model by Gamper and Tashev [39], referred to as Gamper-Tashev-CNN in this paper. These two "black-box" models take as input the reverberant sound, our model takes this as an input as well, but due to the intrusive nature of the SIMs it also

requires clean speech. Prates-DNN achieves an MSE of 0.0394 and an MAE of 0.152 for small rooms. Gamper-Tashev-CNN accomplishes an MSE of 0.0384, the MAE is not reported.

In addition to knowing how well the estimator performs when compared to state-of-models, it is of interest to know how generalisable it is to the real world, especially since simulated RIRs are used. For this purpose we evaluate the estimator on a set of real RIRs. We use the dEchorate dataset originally proposed in [40] as real RIR dataset. The dataset consists of 1800 multichannel RIR recordings sampled at 48 kHz recorded in a single room with different microphone and sound source configurations. We consider each of the channels as a separate recording and evaluate the trained model on a subset of 300 samples randomly taken from the dEchorate dataset. This is not an extensive evaluation on real RIRs but it can provide a general idea on how well the estimator performs in real world scenarios.

In realistic scenarios other types of additive noise may also be present. Therefore, the estimator trained on speech without noise is also evaluated on reverbed speech with additive Gaussian noise to confirm the robustness to noise based on which the SIMs were chosen. Parameters of the Gaussian noise added are calculated per signal based on a desired Signal-to-Noise Ratio (SNR). The clean speech fragment does not have noise added to it for the evaluation, so it remains true "clean speech". The noise robustness evaluation is conducted on a small sample size of 100 RIRs convolved with a clean speech signal from the validation dataset for each SNR. The Gaussian noise is added to the reverberant signal, so after convolution with the RIR. For the SNR selection we use: 10, 20 and 30 dB. This noise robustness analysis is far from elaborate, but it can hint towards the estimator's robustness to additive noise.

IV. Responsible Research

When conducting research, it is of utmost importance that it is done responsibly by promoting: reproducibility, representativeness, and ethical considerations. This section highlights how this research into SIMs and the T60 is performed responsibly.

In order for the scientific community to be able to validate the results this research presents, it is vital that they can easily be reproduced. Reproducibility is primarily achieved by applying open-science principles. Examples of how these principles manifest in this research project are:

- Parameters such as those used to simulate the RIRs are documented in this paper.
- All code used to obtain the RIRs, the estimator and results is made publicly available on the 4TU repository [41] under a GPL license.
- All code in this codebase is documented using code comments and a README file, including recommended parameters.
- Hardware used to obtain the results should not influence results a lot as they can be obtained in a Python virtual environment.
- External libraries used by the codebase are specified in a requirements.txt file which also defines versions of the libraries used to prevent dependency version differences from influencing results.
- Only publicly available external libraries are used in the code (MIT and GNU) this makes the estimator more accessible.
- The clean speech dataset used has an Attribution-NonCommercial license. It is therefore not allowed to use the EARS dataset used to train and validate the estimator for commercial purposes. When commercial usage is desired, the dataset must be substituted for another anechoic clean speech dataset. For more licensing information you can read the README and LICENSE files in the codebase [41].

Secondly, representativeness of the real world is a valid concern in this research. Especially since the RIRs used are simulated, one may wonder if this simulation is an accurate representation of reality. Because of this, the estimator is applied to a few real RIRs to further validate the model's generalizability to the real world.

TABLE II: Gender, ethnicity and native language division (in %) for training and validation subsets of the EARS Dataset, the division is exactly equal for both datasets.

Gender	
Male	40.19
Female	56.07
Non-Binary / Third Gender	0.93
Prefer not to Answer	2.8
Ethnicity	
Asian	3.74
White or Caucasian	68.22
Black or African American	14.95
Black or African American, White or Caucasian	1.87
Hispanic or Latino	6.54
Prefer not to Answer	1.87
Hispanic or Latino, White or Caucasian	1.87
Black or African American, Hispanic or Latino	0.93
Native Language	
Russian	0.93
American English	90.65
Spanish	0.93
Dari	0.93
Mandarin	0.93
Ukrainian	0.93
British English	0.93
German	0.93
Prefer not to Answer	2.8

Since the estimator uses speech fragments, another concern is that the estimator might express biases towards certain groups of society. Table II gives an overview of speaker characteristics of the clean speech dataset the estimator is trained on. Unfortunately, the division of groups is unevenly distributed, especially of native language which could lead to bias based on accents or dialects. We however still believe this dataset to be the right choice for this purpose, as no other large anechoic dataset with a large amount of speakers was found. To the best of our knowledge any harm the estimator could express to marginalised groups within society is minimal. As a consequence we consider this unevenly distributed dataset to be acceptable for this purpose, especially since the estimator

itself might be inherently biased due to possible bias the SIMs themselves show, which is out of our control. This is why we recommend further research into potential racial, gender or other bias the estimator possibly shows.

V. Results

As discussed we construct the estimator by minimzing the MSE. This section explains the obtained estimator using this approach and highlights evaluation results.

A. Obtaining the Estimator

Applying the methodology explained in the previous section we obtain the mutual information between each SIM and the T60 for the simulated training data as can be seen in Table III. From this table it is apparent that for the training dataset SIIB, $\text{SIIB}^{\text{Gauss}}$ and ESTOI have similar mutual information with the T60, while STOI falls behind.

SIIB	$\text{SIIB}^{\text{Gauss}}$	STOI	ESTOI
0.962	1.042	0.681	0.968

TABLE III: A table describing mutual information between each SIM used in this paper and the T60 calculated using the approach described in Section III. A value of 0 represents no shared information, the higher the value, the more shared information between a SIM and the T60. This table is used to find the SIMs with the most shared information with the T60.

Considering the training and validation datapoints in Figure 2, it is observed that the data is shaped more like the hyperbolic rational Formula (5) for SIIB and $\text{SIIB}^{\text{Gauss}}$ and more like the inverse shifted exponential Formula (6) for STOI and ESTOI. The resulting estimator lines can also be seen in the figure based on these curve fits, along with the optimized parameters.

B. Evaluating the Estimator

Given these estimator curves, we now evaluate how well the estimators perform. Table IV displays the MSE and MAE the estimators achieve for both the simulated and real RIR validation datasets. Table V highlights the performance of each estimator under Gaussian noise with SNRs of 10, 20 and 30 dB.

	SIIB	$\text{SIIB}^{\text{Gauss}}$	STOI	ESTOI
MSE Simulated	0.417	0.353	0.552	0.396
MAE Simulated	0.474	0.403	0.563	0.479
MSE Real	0.336	0.323	0.099	0.677
MAE Real	0.358	0.352	0.240	0.341

TABLE IV: Resulting MSE (squared seconds) and MAE (seconds) rounded to three decimals of the T60 estimator are reported for each SIM when evaluated on the simulated RIR validation dataset (2000 samples) and on the real RIR dataset (300 samples).

	SIIB	$\text{SIIB}^{\text{Gauss}}$	STOI	ESTOI
MSE (SNR = 10 dB)	3.644	3.05	0.795	1.50
MAE (SNR = 10 dB)	1.469	1.385	0.761	0.852
MSE (SNR = 20 dB)	2.970	3.27	0.751	0.481
MAE (SNR = 20 dB)	1.330	1.399	0.695	0.552
MSE (SNR = 30 dB)	2.760	3.768	0.742	1.206
MAE (SNR = 30 dB)	1.327	1.507	0.660	0.817

TABLE V: Resulting MSE (squared seconds) and MAE (seconds) of the estimator when evaluated on a dataset of 100 samples for different SNRs, which give a general idea of the estimator's noise robustness.

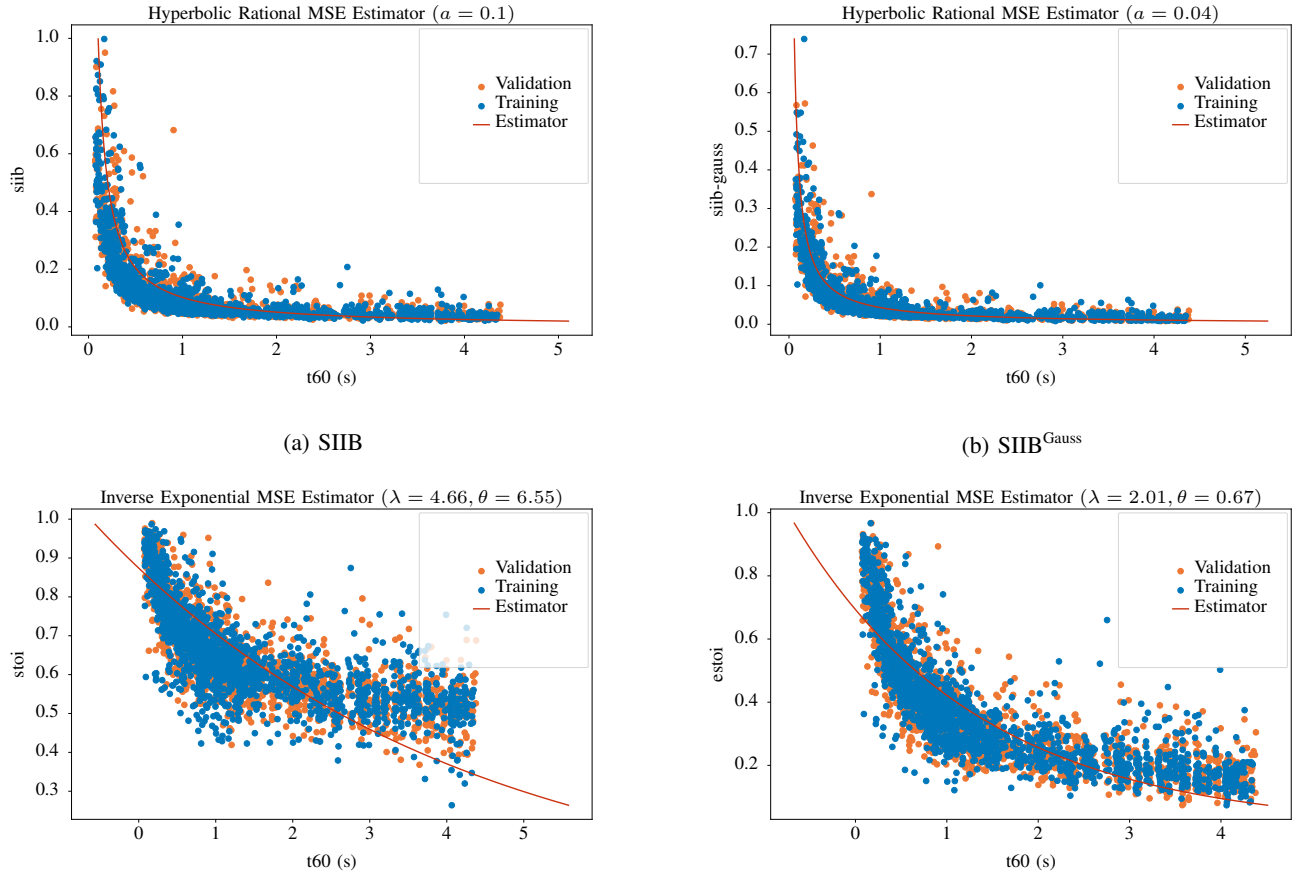
VI. Conclusions and Future Work

When comparing our lowest achieved MSE of 0.353 by the $\text{SIIB}^{\text{Gauss}}$ to those of Prates-DNN (0.0394) and Gamper-Tashev-CNN (0.0384), we notice that it is worse by approximately an order of magnitude. Our lowest achieved MAE of 0.403 is also not close to the MAE Prates-DNN achieved (0.152). Note that this comparison between models should be taken with a grain of salt as the models are validated on different datasets.

Despite the lesser accuracy, we still believe the insights of this research to be valuable due, as we especially for SIIB and $\text{SIIB}^{\text{Gauss}}$ observe a tight curve fit. The statistical estimator has an advantage over machine learning models that it is less computationally complex than the machine learning models. While constructing the estimator this high calculation speed compared to a typical machine learning model was especially apparent for STOI and ESTOI. Therefore, if the accuracy can be improved the estimator can compete with state-of-the-art models. We propose that combining both $\text{SIIB}^{\text{Gauss}}$ and ESTOI into a single estimator could improve results. This is due to their calculation approaches being very different, thus we believe that they capture different information about the T60, which increases mutual information as a consequence. To further improve accuracy we suggest attempting fitting functions other than those proposed in this paper and attempting different parameter optimization strategies such as maximum likelihood estimation or finding a minimum variance unbiased estimator. This would especially be beneficial for the ESTOI estimator as for small T60 the fit is not great. Objective functions other than the MSE should also be explored to potentially improve the curve fits.

By considering at the results of the estimators evaluated under the conditions of additive Gaussian noise, we observe that results suffer heavily under noise. Most noticeable for SIIB and least noticeable for ESTOI. The sample size for noise robustness evaluation is very small so we cannot draw any significant conclusions from this experiment. However, it does lead us to believe the initial idea that these SIMs provide enough noise robustness to produce a noise robust estimator might be a naive approach. A future improvement could be to add a pre-processing layer before inputting the speech data to the SIMs to potentially decrease the negative impact additive noise has on results.

Another limitation the research faces is that the estimators are trained on simulated RIR data. Looking at the results of the small real RIR evaluation experiment, it hints towards the



(a) SIIB

(b) SIIB$^{\text{Gauss}}$

(c) STOI

(d) ESTOI

Fig. 2: For each SIM the ground truth T60 data is plotted for both the training and validation datasets. The resulting curve of minimizing the MSE is also plotted for each SIM. SIIB and SIIB$^{\text{Gauss}}$ are fitted to Formula 5, while STOI and ESTOI are fitted to Formula (6), optimized parameters of the functions a, θ and λ are also given for each estimator.

estimator being generalisable to real RIRs due to similar results compared to simulated RIR results. However, just like the additive noise experiment, the sample size is small, thus further research is required to confirm this idea of generalisability.

As mentioned in Section IV, the speech data the estimator is trained on is inherently biased, which can also influence generalisability of input data. We recommend further experimentation to research how well the estimator performs for speakers of different genders and speakers with different dialects or accents.

We believe that when the recommendations highlighted in this paper are implemented, a good estimator for the T60 using SIMs can be obtained. This work also motivates more research into the relation between non-intrusive SIMs and the T60. If these non-intrusive SIMs show similar or higher mutual information with the T60 than those considered in this work, it could lead to an estimator of the T60 which only requires reverberant speech as an input. If successful, this would allow for a framework where speech fragments could be inserted into the model, and the reverberation time could be estimated from solely this speech fragment. If ASR systems were to apply adaptation schemes based on these estimated T60 values, it could lead to improved ASR system performance.

In conclusion this paper presents four statistical estimators for the T60 using Speech Intelligibility Measure (SIM), which are currently limited in their application by their intrusive nature and performance. Despite the limitations, more research into the usage of SIMs for estimating the T60 is strongly motivated by this work.

References

[1] M. E. Sadeghi, H. Sheikhzadeh, and M. J. Emadi, "A proposed method to improve the WER of an ASR system in the noisy reverberant room," *Journal of the Franklin Institute*, vol. 361, no. 1, pp. 99–109, Jan. 2024, ISSN: 00160032. DOI: 10.1016/j.jfranklin.2023.11.039.

[2] V. Mitra, J. Van Hout, W. Wang, *et al.*, "Improving robustness against reverberation for automatic speech recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, Dec. 2015, pp. 525–532, ISBN: 978-1-4799-7291-3. DOI: 10.1109/ASRU.2015.7404840.

[3] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural Network-Based Spectrum Estimation for Online WPE Dereverberation," in *Interspeech 2017*, ISCA: ISCA, Aug. 2017, pp. 384–388. DOI: 10.21437/Interspeech.2017-733.

[4] F. Xiong, S. Goetze, and B. T. Meyer, "Estimating room acoustic parameters for speech recognizer adaptation and combination in reverberant environments," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2014, pp. 5522–5526, ISBN: 978-1-4799-2893-4. DOI: 10.1109/ICASSP.2014.6854659.

[5] M. Pezzoli, D. Perini, A. Bernardini, F. Borra, F. Antonacci, and A. Sarti, "Deep Prior Approach for Room Impulse Response Reconstruction," *Sensors*, vol. 22, no. 7, p. 2710, Apr. 2022, ISSN: 1424-8220. DOI: 10.3390/s22072710.

[6] M. Long, "Sound in Enclosed Spaces," *Architectural Acoustics*, pp. 313–344, Jan. 2014. DOI: 10.1016/B978-0-12-398258-2.00008-8.

[7] K. Zheng, C. Zheng, J. Sang, Y. Zhang, and X. Li, "Noise-robust blind reverberation time estimation using noise-aware time–frequency masking," *Measurement*, vol. 192, p. 110 901, Mar. 2022, ISSN: 02632241. DOI: 10.1016/j.measurement.2022.110901.

[8] J. Xia, B. Xu, S. Pentony, J. Xu, and J. Swaminathan, "Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1523–1533, Mar. 2018, ISSN: 0001-4966. DOI: 10.1121/1.5026788. [Online]. Available: /asa/jasa/article/143/3/1523/609625/Effects-of-reverberation-and-noise-on-speech.

[9] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Evaluation of Intrusive Instrumental Intelligibility Metrics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2153–2166, Nov. 2018, ISSN: 2329-9290. DOI: 10.1109/TASLP.2018.2856374.

[10] W. Yu and W. B. Kleijn, "Room Acoustical Parameter Estimation From Room Impulse Responses Using Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 436–447, 2021, ISSN: 2329-9290. DOI: 10.1109/TASLP.2020.3043115.

[11] Y. Liang, S. Li, C. Yan, M. Li, and C. Jiang, "Explaining the black-box model: A survey of local interpretation methods for deep neural networks," *Neurocomputing*, vol. 419, pp. 168–182, Jan. 2021, ISSN: 09252312. DOI: 10.1016/j.neucom.2020.08.011.

[12] E. A. Hussein, M. Ghaziasgar, C. Thron, M. Vaccari, and A. Bagula, "Basic Statistical Estimation Outperforms Machine Learning in Monthly Prediction of Seasonal Climatic Parameters," *Atmosphere*, vol. 12, no. 5, p. 539, Apr. 2021, ISSN: 2073-4433. DOI: 10.3390/atmos12050539.

[13] A. Nowoświat and M. Olechowska, "Fast estimation of speech transmission index using the reverberation time," *Applied Acoustics*, vol. 102, pp. 55–61, Jan. 2016, ISSN: 0003682X. DOI: 10.1016/j.apacoust.2015.09.001.

[14] S. K. Tang and M. H. Yeung, "Reverberation times and speech transmission indices in classrooms," *Journal of Sound and Vibration*, vol. 294, no. 3, pp. 596–607, Jun. 2006, ISSN: 0022-460X. DOI: 10.1016/J.JSV.2005.11.027.

[15] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomedical Signal Processing and Control*, vol. 71, p. 103 204, Jan. 2022, ISSN: 1746-8094. DOI: 10.1016/J.BSPC.2021.103204.

[16] J. M. Kates and K. H. Arehart, "An overview of the HASPI and HASQI metrics for predicting speech intelligibility and speech quality for normal hearing, hearing loss, and hearing aids," *Hearing Research*, vol. 426, p. 108 608, Dec. 2022, ISSN: 03785955. DOI: 10.1016/j.heares.2022.108608.

[17] H. Relaño-Iborra, T. May, J. Zaar, C. Scheidiger, and T. Dau, "Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2670–2679, Oct. 2016, ISSN: 0001-4966. DOI: 10.1121/1.4964505.

[18] J. Taghia and R. Martin, "Objective Intelligibility Measures Based on Mutual Information for Speech Subjected to Speech Enhancement Processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 6–16, Jan. 2014, ISSN: 2329-9290. DOI: 10.1109/TASL.2013.2281574.

[19] J. Jensen and C. H. Taal, "Speech Intelligibility Prediction Based on Mutual Information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 430–440, Feb. 2014, ISSN: 2329-9290. DOI: 10.1109/TASLP.2013.2295914.

[20] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, Jan. 2018, ISSN: 1070-9908. DOI: 10.1109/LSP.2017.2774250.

[21] J. Richter, Y.-C. Wu, S. Krenn, *et al.*, *EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation*, 2024.

[22] 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, Apr. 1979, ISSN: 0001-4966. DOI: 10.1121/1.382599.

[23] E. Habets, *ehabets/RIR-Generator: RIR Generator*, Oct. 2020. DOI: 10.5281/ZENODO.4117640. [Online]. Available: https://zenodo.org/records/4117640 (visited on 06/18/2024).

[24] *NEN-EN-ISO 3382-2: Acoustics - Measurement of room acoustic parameters - Reverberation time in ordinairy rooms*. Delft: Nederlands Normalisatie-instituut, 2008. [Online]. Available: https://www.nen.nl/nen-en-iso-3382-2-2008-en-123846.

[25] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms," Oct. 2017. DOI: 10.1109/ICASSP.2018.8461310.

[26] M. R. Schroeder, "New Method of Measuring Reverberation Time," *The Journal of the Acoustical Society of America*, vol. 37, no. 6_Supplement, pp. 1187–1188, Jun. 1965, ISSN: 0001-4966. DOI: 10.1121/1.1939454.

[27] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," 2003. DOI: 10.1121/1.1616578.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 4214–4217, ISBN: 978-1-4244-4295-9. DOI: 10.1109/ICASSP.2010.5495701.

[29] M. Pariente, *Python implementation of STOI*, 2023. [Online]. Available: https://github.com/mpariente/pystoi (visited on 06/18/2024).

[30] N. Kamo, *pySIIB: A python implementation of speech intelligibility in bits (SIIB)*, 2022. [Online]. Available: https://github.com/kamo-naoyuki/pySIIB (visited on 06/18/2024).

[31] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066 138, Jun. 2004, ISSN: 1539-3755. DOI: 10.1103/PhysRevE.69.066138.

[32] *scikit-learn: machine learning in Python — scikit-learn 1.5.0 documentation*, 2024. [Online]. Available: https://scikit-learn.org/stable/index.html.

[33] B. C. Ross, "Mutual Information between Discrete and Continuous Data Sets," *PLoS ONE*, vol. 9, no. 2, e87357, Feb. 2014, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0087357.

[34] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[35] R. K. Pace, "Maximum likelihood estimation," *Handbook of Regional Science*, pp. 1553–1569, Jan. 2014. DOI: 10.1007/978-3-642-23430-9{_}88/TABLES/2. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007/978-3-642-23430-9_88.

[36] C. Xue, T. Zhang, and D. Xiao, "An Advanced Broyden–Fletcher–Goldfarb–Shanno Algorithm for Prediction and Output-Related Fault Monitoring in Case of Outliers," *Journal of Chemistry*, vol. 2022, no. 1, p. 7 093 835, Jan. 2022, ISSN: 2090-9071. DOI: 10.1155/2022/7093835. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1155/2022/7093835%20https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/7093835%20https://onlinelibrary.wiley.com/doi/10.1155/2022/7093835.

[37] A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, pp. 45–76, 2019, ISSN: 15551237. DOI: 10.28945/4184.

[38] R. L. Prates, M. R. Petraglia, J. C. B. Torres, and A. Petraglia, "Blind estimation of reverberation time by neural networks," in *Proceedings of the 23rd International Congress on Acoustics : integrating 4th EAA Euroregio*, 2019, pp. 9–13. DOI: https://doi.org/10.18154/RWTH-CONV-239859.

[39] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," *16th International Workshop on Acoustic Signal Enhancement, IWAENC 2018 - Proceedings*, pp. 136–140, Nov. 2018. DOI: 10.1109/IWAENC.2018.8521241.

[40] D. D. Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, "dEchorate: a calibrated room impulse response dataset for echo-aware signal processing," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–15, Dec. 2021, ISSN: 16874722. DOI: 10.1186/S13636-021-00229-0/TABLES/8. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-021-00229-0.

[41] Maxim de Groot, *Codebase underlying the BSc thesis: Estimating Reverberation Time by a Function of Intrusive Speech Intelligibility Measures*, Jun. 2024. DOI: 10.4121/8411a570-fb8e-451a-8228-1a5806f7b650.