<Blind Reverberation Time Estimation using A Convolutional Neural Network with Encoder>

< **Xingyu Han**[1] >

Supervisor(s): <Dr. Jorge Martinez Castaneda[1]**>, <ir. 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 24, 2024

Name of the student: <Xingyu Han>
Final project course: CSE3000 Research Project
Thesis committee: <Dr. Jorge Martinez Castaneda>, <ir. Dimme de Groot>, <Dr. Sole Pera>

Blind Reverberation Time Estimation using A Convolutional Neural Network with Encoder

Xingyu Han, Jorge Martinez Castaneda, Dimme de Groot

***Abstract*—Estimating reverberation time (RT60) accurately is crucial for enhancing the acoustic quality of various environments as it decides how you feel the sound fades away subjectively. Traditional methods, such as Sabine's equation, require extensive prior knowledge and assume ideal conditions, limiting their practicality. To address these limitations, this paper explores the application of convolutional neural networks (CNNs) enhanced with an encoder architecture based on transformer mechanisms for blind RT60 estimation. The proposed model leverages simulated and real-world datasets, incorporating environmental noise to improve robustness. Results indicate that the CNN-Encoder model achieves superior performance, with a mean squared error (MSE) as low as 0.0006 seconds for pure room impulse responses (RIRs) and 0.0011 seconds under +30dB signal-to-noise ratio (SNR) conditions. It also demonstrates potential in practical usage achieving an MSE of 0.0282 seconds under audio recordings. This approach offers a significant reduction in estimation error compared to the CNN-only architecture, demonstrating the potential for improved acoustic parameter estimation in varied environments. Future work will focus on further optimizing the model for real-world applications and reducing computational complexity while maintaining high accuracy.**

***Index Terms*—Reverberation Time Estimation, Convolutional Neural Network, Encoder Architecture, Transformer, Blind Estimation, Acoustic Environment, Signal-to-Noise Ratio.**

I. Introduction

Understanding how sound behaves in a room can be a bit abstract, but think of it this way: imagine you're in a large hall and you clap your hands. You might hear an echo that gradually fades away. This fading sound is what we call room acoustics, which describes how sound reflects, absorbs, and decays in a space. One important aspect of this is reverberation, which is how long it takes for the sound to drop to a barely noticeable level. For instance, in a concert hall, we want just the right amount of reverberation to make the music sound rich and full.

To scientifically measure and understand this, experts use a model known as the Room Impulse Response (RIR). This model captures how a sound behaves in a room and is affected by several room properties such as reflection coefficients, room geometry, and reverberation time. The RIR helps estimate the reverberation time, or RT60, which is the time it takes for the sound to decay by 60 decibels. Estimating RT60 accurately can help us control the quality of music and speech intelligibility in a space.

Commonly, RT60 is estimated using Sabine's equation [1], which derives RT60 from measured room impulse responses (RIRs). The equation is represented as:

$$RT_{60} = \frac{0.161V}{A} \tag{1}$$

where RT_{60} is the reverberation time in seconds, V is the volume of the room in cubic meters, and A is the total absorption in the room, measured in square meters of equivalent absorption area. However, this method requires prior knowledge about room characteristics and assumes an ideal diffuse soundfield, which can be challenging to achieve in practice. As an alternative, blind estimation methods have been developed to estimate RT60 directly from RIRs or audio recordings, without needing detailed room information. Various algorithms have been proposed for blind RT60 estimation based on recorded speech signals [2], [3], [4], [5], [6], achieving a maximum average estimation error of 0.11 seconds within a Signal-to-Noise Ratio (SNR) range of 10dB to 60dB [6]. However, these methods involve extensive manual work in selecting appropriate distributions and smoothing functions, and their performance may be constrained by these assumptions.

Deep neural networks (DNNs), particularly convolutional neural networks (CNNs), have shown promise in addressing these limitations due to their ability to capture underlying patterns and generalizations. A notable example is the 6-layer CNN proposed by Gamper and Tashev [7], which achieved a mean squared error (MSE) of 0.0384 seconds, outperforming the best method from the ACE challenge [8]. Further enhancements by other researchers [9], [10], [11], [12] have improved performance to an MSE of 0.0206 seconds [9] under the same ACE Challenge dataset and extended applicability to dynamic acoustic conditions. Recently, the attention mechanism from transformers [13] has gained attention for its superior performance in encoding input information and understanding patterns, further reducing MSE errors to 0.02 seconds [14] under ACE challenge [8] and 0.1541 seconds [15] with varied input length signals, albeit with increased training parameters and computational complexity.

This paper aims to enhance estimation accuracy in reverberation time estimation by addressing the following research question:

What mechanism can be introduced, and what impact does this enhancement have on blind reverberation time estimation accuracy?

To explore this, we introduce an encoder based on the transformer architecture [13]. This encoder is designed to capture more relevant feature information and effectively compress raw audio signals. We aim to connect the encoder and CNN to gain benefits from both sides. This mechanism has the potential to surpass previous performance limitations without significantly increasing training effort.

The structure of this paper is as follows: II outlines the background of our method. III explains the methodology used

in this research. IV details the proposed two different model architectures for evaluation. V discusses responsible research aspects. VI provides further analysis and discussion of the findings. Finally, VII concludes the paper, summarizing the implications of our findings and suggesting areas for future improvement.

II. Background

This section provides a detailed overview of the research path for addressing blind RT60 estimation.

Traditional empirical methods for RT60 estimation, such as Sabine's equation, require prior knowledge of room characteristics, making them impractical for many real-world applications. To overcome this limitation, Ratnam et al. [2] proposed a maximum-likelihood approach for connected speech, modeling reverberation as exponentially damped Gaussian white noise. This method achieved an estimation of 1.62 seconds compared to the ground truth of 1.66 seconds, but it also highlighted the performance differences across frequency bands and the high computational costs due to the iterative solution of the maximum-likelihood equation. To mitigate these computational demands, techniques such as downsampling and pre-selecting potential sound decays were introduced, enabling the algorithm to track time-varying RT60 with higher accuracy [3].

Another relevant improvement was proposed by Li, Schlieper, and Peissig [6], who estimated reverberation time in separate frequency bands based on recorded speech signals. They calculated the full-band RT60 by combining estimations from the 1-4kHz and 4-20kHz frequency regions. This hybrid model achieved an average estimation error ranging from 0.04 to 0.11 seconds within Signal-to-Noise Ratios from 10dB to 60dB. Despite increased accuracy and robustness to noise performance, the complexity of these traditional signal-processing approaches also grew, and the frequency-dependent estimation was limited by smaller bandwidths and lower signal energy, which might affect the full-band estimation.

Deep neural networks (DNNs), particularly convolutional neural networks (CNNs), have emerged as a competitive solution. Gamper and Tashev [7] applied a CNN with spectro-temporal features in the time-frequency domain, outperforming the best method from the ACE challenge [8]. However, their method was limited to fixed-length temporal inputs and could not accommodate time-varying scenarios. To address this, a long short-term memory (LSTM) [9] was added to the CNN model [7], maintaining the interdependent relationship within varying input temporal data and achieving a lower mean squared error (MSE) of 0.0206 seconds on the same ACE evaluation [8], though with a larger amount of training parameters.

Another improvement involved augmenting and expanding the small real acoustic impulse response dataset to a larger, more balanced one [11] which achieved a similar MSE error compared to the LSTM mechanism [9]. Both LSTM and data augmentation aimed to extract more information from the input. Following this idea, Ick, Mehrabi, and Jin [12] suggested that important information might be lost during the time-frequency transform. By reintroducing phase information, the MSE loss was further decreased.

Recognizing the superior data understanding capabilities of Large Language Models, the attention mechanism from transformer architecture [13] was explored for blind RT60 estimation. Compared to several CNN-based and CRNN-based models, the transformer-based model demonstrated better performance and even achieved an MSE loss of 0.1541 seconds under varied input length signals [15]. However, this increased accuracy and flexibility came with a significant rise in training parameters, from 0.013 million [12] to 85.256 million [15]. To address this, Saini and Peissig [14] proposed a lightweight architecture for mobile-friendly applications by combining the transformer with MobileViT V3 blocks, maintaining an MSE loss of 0.02 seconds with only 61,000 parameters under ACE Challenge corpus [8]. Despite these improvements, [15] and [14] primarily relied on the transformer, adding only a linear layer for the final regression task, leading to a heavy training burden.

In response, this paper proposes a novel approach that uses the attention encoder as part of the feature extraction rather than the entire transformer. Two different model architectures connecting the encoder and CNN are explored. By dividing the tasks for each component, we aim to combine the benefits of attention mechanisms and CNNs to achieve high accuracy while maintaining relatively low training requirements.

III. Methods

The objective of this experiment is to estimate reverberation time (RT60) using a Convolutional Neural Network (CNN) with an encoder architecture. To achieve this, a comprehensive experimental setup has been developed, incorporating both simulated and real-world datasets. The following sections outline the simulation process, dataset characteristics, and the methodology employed for training and evaluating the model.

A. Simulated RIRs

For the simulated dataset, we utilize the RIR-Generator library [16], which is proficient at generating Room Impulse Responses (RIRs). This library provides tools for creating accurate acoustic models of various environments by simulating sound propagation and reflection within a defined space. In our experiment, simulated RIRs are generated with target reverberation times (RT60) ranging from 0.1 to 2.0 seconds, covering a wide range of acoustic environments from relatively dry to highly reverberant spaces. To account for potential biases in the generation process, the measured RT60 based on the Schroeder equation [17] is used as the ground truth for training and testing.

Fig. 1 illustrates an example where the intended RT60 is 2.0 seconds, but the actual generated RT60 is approximately 2.6 seconds.

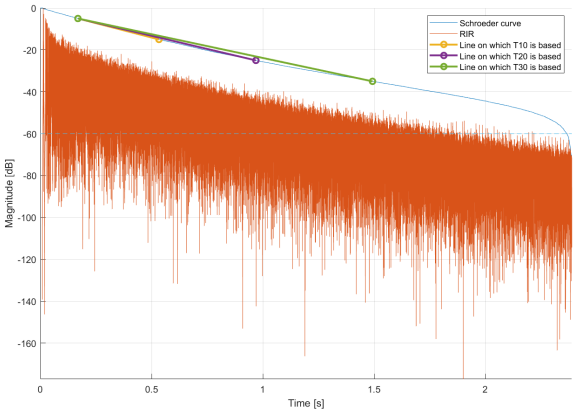

Fig. 1: An example of bias in RIR simulation. The target RT60 is 2.0 seconds, while the actual generated sample is around 2.6 seconds, determined by the intersection of -60 dB and the Schroeder Curve.

To ensure the diversity and robustness of our dataset, several parameters are regulated during the simulation process, including:

- **Room Geometry**: The dimensions of the shoebox rooms (length, width, and height) are scaled exponentially with the targeted RT60 to avoid unrealistic conflicts, ranging from 2.0 to 9.5 meters in all dimensions. This variation ensures the dataset includes a variety of spatial configurations, enhancing the model's ability to generalize across different room shapes and sizes.
- **Reflection Coefficients**: Three positive and corresponding negative coefficients are assigned to the six walls of the room to obtain more realistic RIRs. These coefficients are averaged to the mean coefficient calculated by the inverted Sabine's equation [1], influencing the amount of sound energy reflected off surfaces and affecting the overall reverberation time.
- **Source and Receiver Positions**: The positions of the sound source and the receiver within the room are randomly assigned, maintaining a minimum distance of 50 cm from each wall. The distance between the source and receiver is at least 20 cm to ensure a better simulation. Each room configuration includes 100 channel samples generated from 20 receivers and 5 sources.

B. Environmental Noise

In real-world scenarios, it is challenging to measure clean RIRs due to environmental noise. To validate the practical capabilities of the model, we simulate environmental noise as Gaussian white noise at a signal-to-noise ratio (SNR) level of +30 decibels. The noised RIR $\mathbf{r}$ is calculated as:

$$\mathbf{r} = \mathbf{h} + \mathbf{n} \tag{2}$$

where $\mathbf{h}$ is the clean RIR generated from the simulated dataset III-A, and $\mathbf{n}$ represents the Gaussian white noise. Both $\mathbf{h}$ and $\mathbf{n}$ have the same length.

C. Simulated Audio Recordings

In addition to the clean RIR data, we use anechoic speech datasets from the ACE Challenge corpus [8] to simulate real audio signals. These anechoic speeches, recorded with minimal reflected sound energy, are convolved with the generated RIRs to create audio signals under various environmental conditions. The generated audio signals $\mathbf{y}$ are represented as:

$$\mathbf{y} = \mathbf{s} * \mathbf{h} + \mathbf{n} \tag{3}$$

where $\mathbf{s}$ is the anechoic speech, $\mathbf{h}$ is the generated RIR from III-A, and $\mathbf{n}$ is the Gaussian white noise from III-B. These convolved recordings enable the model to learn realistic signal patterns and are evaluated from a practical perspective.

D. Generalization Evaluation Dataset

To assess the learning and generalization abilities of the models, two datasets are selected for evaluation. The single-channel RIR dataset is used to evaluate the model trained solely on RIRs. For the model trained on both RIRs and simulated audio recordings, the evaluation dataset is constructed by convolving each channel of the RIRs in the EM32 dataset with randomly chosen speeches from the anechoic speech dataset. The corresponding ambient noise is also added to the convolved signals. All datasets, including the single-channel RIR dataset, EM32 dataset, and anechoic speech dataset, are sourced from the ACE Challenge corpus [8].

E. Evaluation Metrics

The model's performance is evaluated using standard metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE) between the predicted and actual RT60 values. MAE provides an intuitive understanding of the estimation error across all data samples, while MSE reflects how the model handles outliers, highlighting individual large errors through the square calculation. These metrics offer a clear indication of the model's accuracy and robustness across different acoustic environments.

Additionally, the Pearson coefficient (ρ) is introduced to measure the similarity between the estimated and ground truth values. Higher similarity indicates a better ability to learn underlying patterns. Therefore, a superior model is represented by lower MSE and MAE values, along with a higher ρ, indicating greater accuracy and similarity.

F. Data Preprocessing

Both simulated and real-world datasets undergo preprocessing before being fed into the model. This phase involves normalizing the audio signals and converting them into a suitable format. Specifically, Mel spectrograms are computed from the RIRs, as they provide a rich representation of the time-frequency characteristics essential for accurately estimating reverberation time and handling complex speech signals. This process also reduces the data representation scale, easing the training workload.

The input signals are truncated or padded to 2 seconds (96,000 samples at 48kHz) to maintain uniform data size,

then transformed into 128×188 Mel spectrograms. These spectrograms are shuffled and divided by the data loader to form batches of 32 for the training process. A similar process is applied during testing.

By integrating simulated and real-world datasets and employing a well-structured CNN with an encoder architecture, our experiment aims to develop a reliable model for estimating reverberation time in diverse acoustic settings.

IV. Model Architectures

In this section, we introduce two different model architectures, Encoder-CNN and CNN-Encoder, for blind reverberation time estimation using both simulated RIRs and audio recordings. These architectures differ in the sequence of connecting the CNN and encoder, impacting the estimation accuracy and generalization ability.

A. Encoder-CNN

In the Encoder-CNN architecture, the CNN follows the encoder layer, taking its output as input to estimate RT60. This design is based on the intuition that the early encoder layer can progressively capture the interrelationships and underlying patterns of the input using the self-attention mechanism. The CNN then filters and maps these extracted features to a regression estimation through multiple convolution layers. It is anticipated that the CNN will benefit from the preserved global input information, achieving higher estimation accuracy. Given the excellent performance of the Vision Transformer (ViT) [18] in image feature extraction and classification, it is used as the encoder block to enhance performance.

Fig. 2 illustrates this architecture. A 128×188 Mel spectrum from III-F, corresponding to an RIR signal, is the input to the encoder. Within the encoder, the spectrum information is enhanced based on learned attention weights. Then, three convolution layers compress dimensions and output extracted features. A max-pooling layer and a softmax layer are sequentially connected to each convolution layer to meet the requirements. The last two linear layers map the flattened data array to a single regression estimation used for evaluation and comparison.

B. CNN-Encoder

The CNN-Encoder architecture reverses the sequence used in Encoder-CNN. It first convolves the input data with several convolution layers, then resizes the output as a series of data tokens. The ViT encoder then extracts features and maps them to the estimation through linear layers. This design is inspired by findings that early convolution can help ViT converge quickly and improve robustness under different optimizers [19].

Fig. 3 provides an example of the CNN-Encoder architecture. This process takes the same input as Encoder-CNN but first convolves the spectrogram with several convolutional layers, then reshapes its format to fit the encoder. One of the last two linear layers is removed to avoid losing important information within multiple linear mappings from the encoder output.

V. Responsible Research

Our research adheres to responsible research principles, ensuring that our findings are transparent, reproducible, and ethically sound. To facilitate reproducibility and transparency, we have implemented several measures detailed below:

A. Data Accessibility and Transparency

- All code and datasets used in this study will be made publicly available through the 4TU.Centre for Research Data repository. This repository is chosen for its compliance with the FAIR principles, ensuring that our data is Findable, Accessible, Interoperable, and Re-usable.
- The datasets included in our study are publicly available and collected under licenses that permit their use for research purposes. This guarantees that there are no privacy or sensitive information concerns associated with the data used in our experiments.

B. Reproducibility

- We have used random seeds to control all random processes within the experiments, ensuring that the results can be consistently reproduced. The specific random seeds and their applications in various stages of the experiments are thoroughly documented.
- Detailed documentation of our experimental setup, including data preprocessing steps, model architectures, training procedures, and evaluation metrics, will be provided. This allows other researchers to replicate our experiments precisely and verify our findings.
- We describe the use of the RIR-Generator library for simulating Room Impulse Responses (RIRs), detailing the parameters and configurations used to ensure diverse and robust datasets.

C. Ethical Considerations

- Our research does not involve any data collection from private individuals or the use of sensitive personal data. The datasets used, such as the ACE Challenge corpus, are publicly available and used in accordance with their respective licenses.
- There is no risk of harm to individuals or communities from our research. Our focus is purely on the technical aspects of reverberation time estimation and does not involve any human subjects or private information.

D. Bias Mitigation

- We have ensured the diversity and robustness of our dataset by regulating parameters such as room geometry, reflection coefficients, and source and receiver positions. This helps mitigate potential biases and improves the generalization ability of our models across different acoustic environments.
- The methodology used to generate and preprocess data is designed to minimize biases and ensure that the model's performance is not unduly influenced by any specific configurations or conditions.

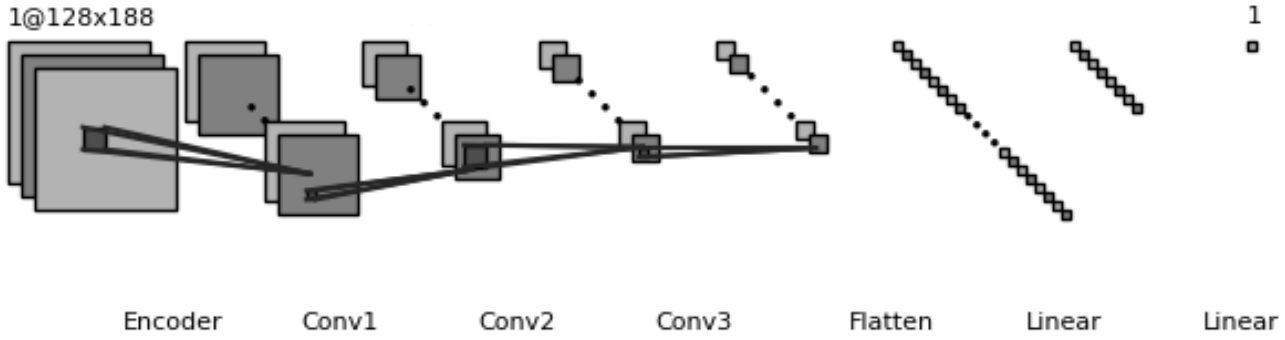

Fig. 2: The model architecture of the Encoder-CNN. The input is the Mel spectrum from III-F. The encoder recalculates the spectrum based on the attention mechanism and feeds its output to CNN. The estimation is mapped by the last two linear layers from the flattened CNN output.

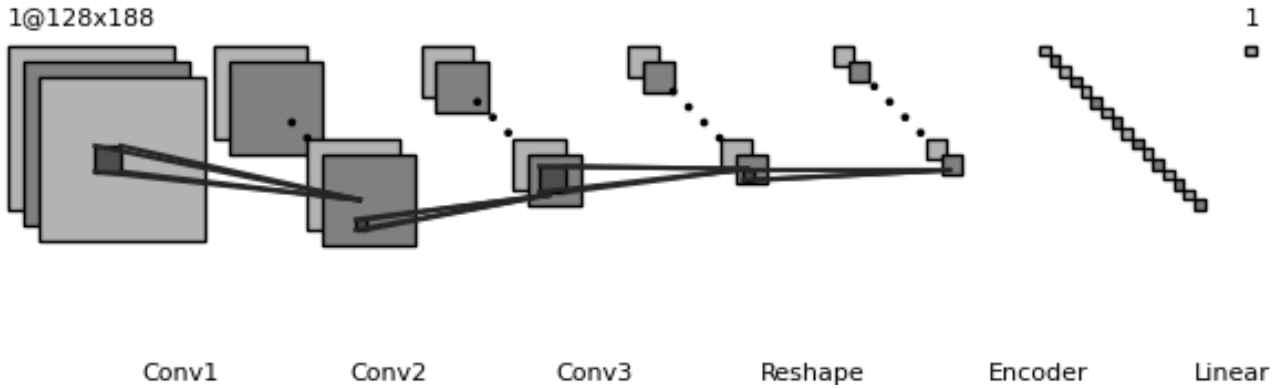

Fig. 3: The model architecture of the CNN-Encoder. The input is the Mel spectrum from III-F. The CNN first convolves the spectrum through several convolutional layers and feeds its reshaped output to the encoder. The estimation is mapped by the last linear layer from the encoder output.

E. Methodological Transparency

- Comprehensive details about the methodologies employed in our study, including the design and implementation of the convolutional neural network (CNN) and encoder architectures, are provided. This ensures that our research process is transparent and can be critically evaluated by peers.
- All hyperparameters, training procedures, and evaluation metrics are explicitly documented, allowing for exact replication of our experiments.

By adhering to these responsible research practices, we aim to contribute to the scientific community in a meaningful and ethical manner, ensuring that our research is both credible and beneficial to future studies in the field of acoustic parameter estimation.

VI. Results

This section discusses the estimation accuracy of the models based on the simulated RIR and audio recording datasets. The discussion is divided into two parts: the first part evaluates the model performance on clean and noised RIRs, while the second part assesses the performance based on simulated audio recordings. The generalization ability is also examined in both parts. All training configurations are set as the learning rate of 0.0001, Adam optimizer, MSE Loss and 10 epochs to avoid overfitting.

A. RT60 Estimation on RIRs

A dataset with 18,100 RIR samples (181 room configurations with 100 channels each) was generated based on the method described in III-A. To better evaluate model performance and minimize potential bias, the training and testing sets were randomly divided into equal sizes. The RT60 ground truth distribution is shown in Fig. 4, covering a range from 0.1 to 2.9 seconds. The wide distribution range also indicates possible biases in the generation, especially for longer RT60 values, which increased from 2.0 to 2.9 seconds. Such biases might confuse the desired pattern during training, lowering the accuracy and generalization of the estimation.

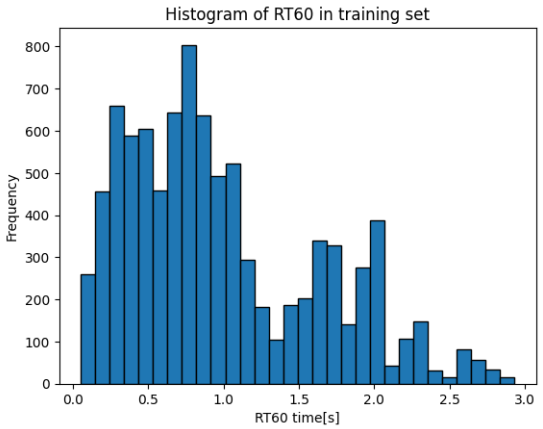

Fig. 4: The distribution of RT60 ranging from 0.1 to 2.9 seconds corresponds to the generated training data samples. The X-axis represents the range of RT60 and the Y-axis represents the number of samples. The testing dataset has a similar distribution due to the equal division.

Table I presents the performance of models on both clean and noised RIRs, with the clean CNN serving as the baseline, which is essentially the CNN part of both Encoder-CNN and CNN-Encoder. The noised RIRs were generated using the same 18,100 samples as described in III-B. All models were retrained and tested to avoid overfitting. Despite differences in the Pearson coefficients among the models, the close-to-1 values indicate their ability to capture similar signal patterns. CNN-Encoder achieves the highest precision in both MAE and MSE, followed by Encoder-CNN. This performance surpasses the baseline CNN model, highlighting the potential of combining the encoder and CNN to leverage global information structure and local pattern exploration.

TABLE I: Performance comparison of Models on clean RIRs and RIRs at SNR +30 decibels

Model	**clean RIR**			**SNR +30**		
	MSE[s]	**MAE[s]**	ρ	**MSE[s]**	**MAE[s]**	ρ
Clean CNN	0.0044	0.0513	0.9969	0.0051	0.0487	0.9941
Encoder-CNN	0.0013	0.0237	0.9983	0.0027	0.0367	0.9968
CNN-Encoder	**0.0006**	**0.0198**	**0.9993**	**0.0011**	**0.0246**	**0.999**

It is noteworthy that the order of CNN and encoder impacts accuracy. One explanation is that while the encoder preserves interrelationships within the context, the local relationships might be altered due to the attention mechanism, potentially harming CNN performance due to its restricted kernel size view. Conversely, CNN-Encoder avoids this issue, as the CNN first explores local data within the kernel, leaving the global relationship unchanged. The encoder then extracts interrelationships within the context, benefiting from the already filtered local information.

B. RT60 Estimation on Audio Recordings

Although the CNN-Encoder and Encoder-CNN models show better accuracy, their performance on audio recordings remains unknown. To address this, 10 out of 100 channels from each of the 181 room configurations were randomly selected and convolved with randomly chosen anechoic speech from the ACE Challenge corpus [8], forming a training set of simulated audio recordings. These 1,810 convolved audio recordings were further set at an SNR of +30 dB to simulate environmental noise.

Instead of retraining all models, these recordings were used to fine-tune the models already trained on clean and noised RIRs. This approach reduces training costs and improves performance, given the smaller simulated audio recording set compared to the 18,100 RIRs, while transformer architectures require large datasets. The evaluation set from III-D was used to measure performance and assess generalization ability.

TABLE II: Generalization Performance on ACE Audio Recording Datasets for RT60 Estimation

Model	**MSE [s]**	**MAE [s]**	ρ
Clean CNN	0.3555	0.4979	0.6729
Encoder-CNN	0.035	0.1418	0.8253
CNN-Encoder	0.0282	0.1143	0.8524
AudMobNet L [14]	0.02	-	0.9

As shown in Table II, CNN-Encoder outperforms Encoder-CNN and clean CNN, demonstrating the advantages of its specific architecture order and potential for practical applications. Additionally, CNN-Encoder achieves performance close to the state-of-the-art model AudMobNet L [14], also evaluated on the ACE Challenge corpus [8]. However, our training and evaluation datasets share the same anechoic speech dataset, which may cause potential overfitting during training. Although each simulated audio recording is a randomly selected 2-second sequence from the convolved signal described in III-D to reduce the overfitting influence, the actual performance of CNN-Encoder remains to be fully understood. Furthermore, AudMobNet L is optimized for model size and training speed for mobile applications, while our models still face constraints in these areas and require further improvement.

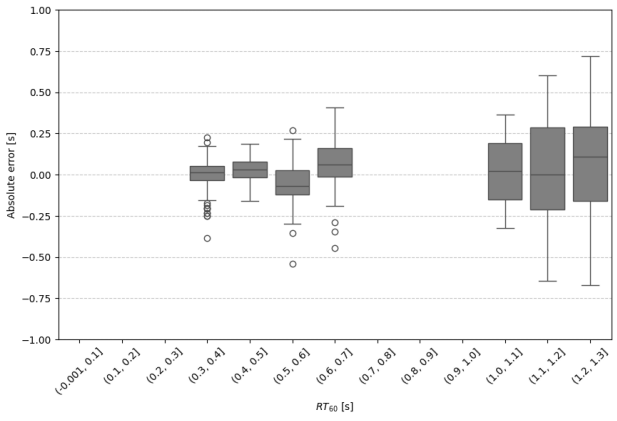

Fig. 5: The distribution of absolute error for generalization performance on ACE audio recording datasets for RT60 Estimation. The X-axis represents the RT60 groundtruth range and the Y-axis represents the error between estimation and groundtruth. Higher RT60 shows larger variance and instability.

Fig. 5 illustrates the distribution of absolute errors in seconds for different RT60 bins from the evaluation dataset. Each box represents the interquartile range (IQR) of errors for the corresponding RT60 bin, with the median shown by the line inside each box, and whiskers extending to 1.5 times the IQR. Generally, the errors are centered around zero, indicating accurate estimations. However, there is noticeable variability in higher RT60 bins, particularly in ranges such as 1.1 to 1.3 seconds, indicating less consistent performance in these regions. This inconsistency may be caused by the loss of important information due to the truncation of the input signals, as higher RT60 values may require longer input signals to reveal their patterns. Lower RT60 bins exhibit tighter error distributions, reflecting better estimation accuracy. Outliers are present across most bins, highlighting occasional significant deviations from true values. Overall, while the estimation method shows good accuracy for lower RT60 values, its performance for higher values could benefit from further refinement to reduce variability and outliers.

VII. Conclusions and Future Work

In conclusion, this study presents a novel approach to blind reverberation time estimation by integrating a convolutional neural network with an encoder architecture based on the transformer mechanism. The CNN-Encoder model demonstrates superior accuracy and generalization ability compared to alternative Encoder-CNN and standalone CNN models. Our findings indicate that the proposed architecture effectively captures complex acoustic patterns, making it suitable for practical applications in diverse acoustic environments. Future work will focus on optimizing the model for mobile applications, reducing computational complexity while maintaining high accuracy. Additionally, expanding the dataset with real-world recordings, assessing bias between target and generated RT60 results, and obtaining high-precision labels will help validate and enhance the model's performance in more diverse scenarios.

References

[1] H. Kuttruff, *Room Acoustics, Fifth Edition*. CRC Press, 4 2014. [Online]. Available: https://www.taylorfrancis.com/books/9781482266450

[2] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time." *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–92, 11 2003. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/14650022

[3] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An Improved Algorithm for Blind Reverberation Time Estimation," Tech. Rep. [Online]. Available: https://www.researchgate.net/publication/229008151_An_improved_algorithm_for_blind_reverberation_time_estimation

[4] J. Y. Wen, E. A. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3 2008, pp. 329–332. [Online]. Available: http://ieeexplore.ieee.org/document/4517613/

[5] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5 2013, pp. 161–165. [Online]. Available: http://ieeexplore.ieee.org/document/6637629/

[6] 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, 5 2019, pp. 211–215. [Online]. Available: https://ieeexplore.ieee.org/document/8682661/

[7] H. Gamper and I. J. Tashev, "Blind Reverberation Time Estimation Using a Convolutional Neural Network," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 9 2018, pp. 136–140. [Online]. Available: https://ieeexplore.ieee.org/document/8521241/

[8] 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, 10 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7486010/

[9] S. Deng, W. Mack, and E. A. Habets, "Online Blind Reverberation Time Estimation Using CRNNs," in *Interspeech 2020*, vol. 2020-October. ISCA: ISCA, 10 2020, pp. 5061–5065. [Online]. Available: https://www.isca-archive.org/interspeech_2020/deng20c_interspeech.html

[10] P. Gotz, C. Tuna, A. Walther, and E. A. P. Habets, "Blind Reverberation Time Estimation in Dynamic Acoustic Conditions," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2022-May. IEEE, 5 2022, pp. 581–585. [Online]. Available: https://ieeexplore.ieee.org/document/9746457/

[11] 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, 5 2020, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9052970/

[12] C. Ick, A. Mehrabi, and W. Jin, "Blind Acoustic Room Parameter Estimation Using Phase Features," 3 2023. [Online]. Available: http://arxiv.org/abs/2303.07449

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 2017-December, 2017.

[14] S. Saini and J. Peissig, "Blind Room Acoustic Parameters Estimation Using Mobile Audio Transformer," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, vol. 2023-October. IEEE, 10 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10248186/

[15] C. Wang, M. Jia, M. Li, C. Bao, and W. Jin, "Exploring the power of pure attention mechanisms in blind room parameter estimation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 23, 4 2024. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-024-00344-8

[16] Habets, "RIR-Generator," 6 2024. [Online]. Available: https://github.com/ehabets/RIR-Generator

[17] M. R. Schroeder, "New Method of Measuring Reverberation Time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 3 1965. [Online]. Available: https://pubs.aip.org/jasa/article/37/3/409/720995/New-Method-of-Measuring-Reverberation-Time

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 10 2020. [Online]. Available: http://arxiv.org/abs/2010.11929

[19] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in neural information processing systems*, vol. 34, pp. 30 392–30 400, 2021.