Blind Reverberation Time Estimation Using A Convolutional Neural Network With Encoder

Xingyu Han[A], Dr. Jorge Martinez[B], ir. Dimme de Groot[C]
A Author, X.HAN-5@student.tudelft.nl
B Responsible Professor
C Supervisor

Introduction

- Imaging clapping hands in different rooms, the sound is also perceived differently by humans. This perception difference comes from the specific room characteristics.
- **Room Impulse Response (RIR)** is such a scientific model describing how a sound behaves in a room and is affected by several room properties.
- **Reverberation time (RT60),** defined as the time it takes for the sound to decay by 60 decibels, is a key parameter deciding how you feel the sound gradually fades away.
- RIR can help estimate RT60, but it is still challenging due to the complexity of physical acoustic systems in real-world scenarios.
- The common empirical approaches exist but require prior room knowledge and are under idealized assumptions.
- **Blind estimation** (without prior room knowledge) is expected by deep learning techniques to estimate accurate RT60 directly from RIRs or audio recordings.
- Existing models are mainly CNN-based and transformer-based. CNN-based models are easier to converge and training-friendly while transformer-based models achieve higher accuracy at the cost of heavy training burden.

TU Delft Delft University of Technology

Objective

The main research question is ***What mechanism can be introduced, and what impact does this enhancement have on blind reverberation time estimation accuracy?***

Connecting both CNN and Encoder(transformer):

- **Assumption 1:** CNN can overcome the limitation of the kernel size by benefiting from the interrelationship captured by the attention mechanism in transformer encoder.
- **Assumption 2:** The Encoder can extract better data features with the deeply explored local information generated by the convolution process of CNN.
- **Assumption 3:** The introduced transformer is an encoder-only architecture which reduces the training difficulty.

Data Generation

- A RIR dataset with 18100 samples covering a RT60 range from 0.1 to 2.9 seconds is generated based on the image-source method following restrictions on room geometry, reflection coefficients, and source and receiver positions.
- 1810 samples from the dataset are selected to convolve with ambient speeches to finetune models for evaluating performance in real applications.
- Gaussian white noise at a signal-to-noise ratio of +30 dB is simulated as the environmental noise.
- ACE Challenge corpus is chosen as the generalization evaluation dataset containing RIRs, ambient speeches, and environmental noise recorded in the real world.

Model Architectures

- Two different model architectures are proposed based on the **Assumption 1** and **Assumption 2**.

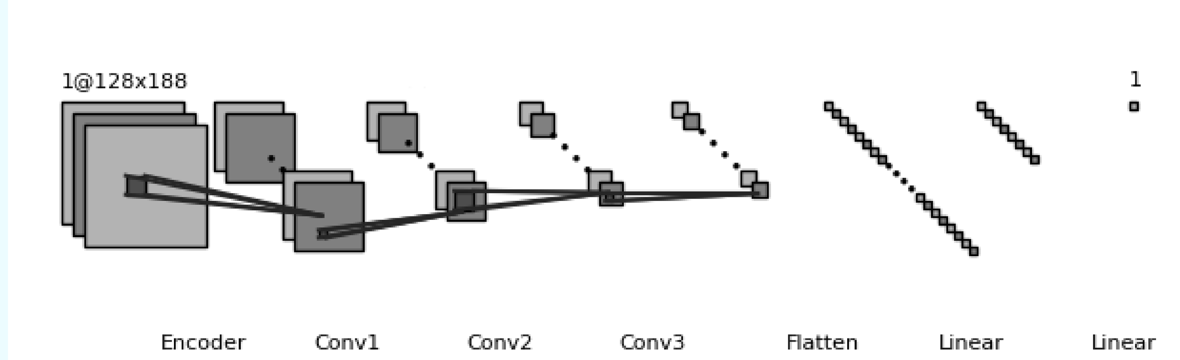

Figure 1: Encoder-CNN

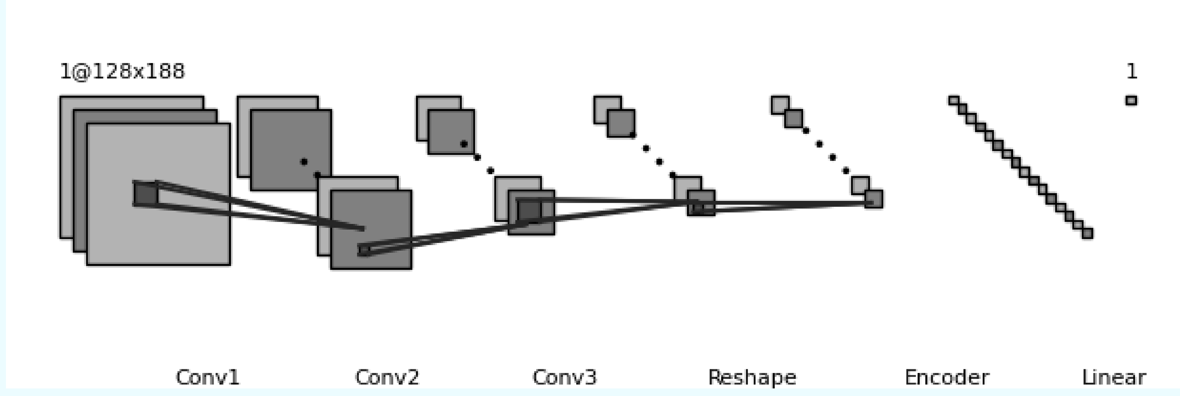

Figure 2: CNN-Encoder

- **Encoder-CNN** shown in Figure. 1 is based on **Assumption 1** which first extracts global interrelationship from the encoder and then feeds to the CNN to get estimation.
- **CNN-Encoder** shown in Figure. 2 takes the inverse connection order based on **Assumption 2** that first convolved input through CNN layers to enhance local information and then lets the encoder further extract and map global information to estimation result.

Results

- The results are given in two parts: one for the model performance on clean and noised RIRs shown in TABLE 1, and the other in TABLE 2 is evaluated based on audio recordings.
- Clean CNN in TABLE 1 and 2 serves as a baseline model, essentially the same CNN part of both Encoder-CNN and CNN-Encoder.
- Mean squared error (MSE), mean absolute error (MAE) and the Pearson coefficient (ρ) are used as evaluation metrics. Better performance is usually represented as lower MSE, MAE and higher ρ.
- CNN-Encoder performs best under all evaluations followed by Encoder-CNN and clean CNN. It also achieves similar accuracy in audio recordings compared to the state-of-the-art model AudMobNet L.
- An explanation for the superiority of CNN-Encoder compared to Encoder-CNN is that it avoids alternated global information in later CNN stages.
- Figure. 3 illustrates the inconsistent performance among the high RT60 range which may caused by the truncation of the input signal that loses important information for estimating longer RT60s.

TABLE 1: 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**

TABLE 2: 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	0.02	-	0.9

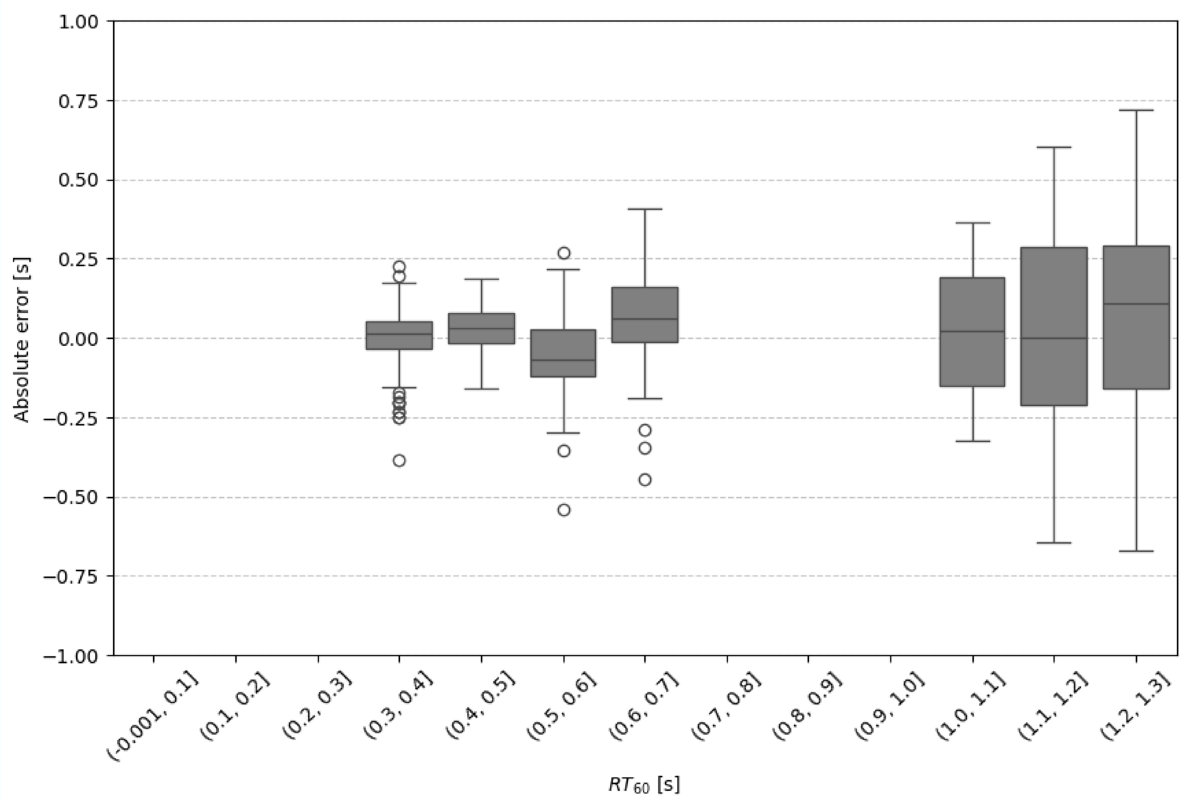

Figure 3: The distribution of absolute error for generalization performance on ACE audio recording datasets for RT60 Estimation. The X-axis represents the RT60 groundtruth range and the Y-axis represents the error between estimation and groundtruth.

Conclusion

- This study introduces a novel approach to blind reverberation time estimation.
- The approach integrates a convolutional neural network (CNN) with an encoder architecture based on the transformer mechanism.
- The CNN-Encoder model shows superior accuracy and generalization compared to alternative Encoder-CNN and standalone CNN models.
- The order of connecting the CNN and the encoder will influence the model performance and estimation accuracy.
- The proposed architecture effectively captures complex acoustic patterns, making it suitable for practical applications in diverse acoustic environments.

Future Work

- Optimizing the model for mobile applications.
- Reducing computational complexity while maintaining high accuracy.
- Expanding the dataset with real-world recordings.
- Assessing bias between target and generated RT60 results.
- Obtaining high-precision labels to validate and enhance the model's performance in more diverse scenarios.
- Exploring transfer learning techniques to adapt the model to different acoustic environments with minimal retraining.
- Investigating the integration of additional acoustic features to improve model accuracy.
- Evaluating the model's flexibility for joint estimation tasks.

References

1. Kuttruff, H. (2014). Room Acoustics, Fifth Edition. CRC Press.
2. Saini, S., & Peissig, J. (2023). 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, pp. 1–5.
3. Wang, C., Jia, M., Li, M., Bao, C., & Jin, W. (2024). Exploring the power of pure attention mechanisms in blind room parameter estimation. EURASIP Journal on Audio, Speech, and Music Processing, 2024(1), 23.
4. Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., & Girshick, R. (2021). Early convolutions help transformers see better. Advances in Neural Information Processing Systems, 34, 30392–30400.
5. Bryan, N. J. (2020). 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), pp. 1–5.
6. Gotz, P., Tuna, C., Walther, A., & Habets, E. A. P. (2022). Blind reverberation time estimation in dynamic acoustic conditions. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 581–585.
7. Deng, S., Mack, W., & Habets, E. A. (2020). Online blind reverberation time estimation using CRNNs. In Interspeech 2020, vol. 2020-October, pp. 5061–5065.
8. Gamper, H., & Tashev, I. J. (2018). Blind reverberation time estimation using a convolutional neural network. In 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 136–140.
9. Eaton, J., Gaubitch, N. D., Moore, A. H., & Naylor, P. A. (2016). Estimation of room acoustic parameters: The ACE challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(10), 1681–1693.
10. Schroeder, M. R. (1965). New method of measuring reverberation time. The Journal of the Acoustical Society of America, 37(3), 409–412.
11. Ratnam, R., Jones, D. L., Wheeler, B. C., O'Brien, W. D., Lansing, C. R., & Feng, A. S. (2003). Blind estimation of reverberation time. The Journal of the Acoustical Society of America, 114(5), 2877–92.
12. Wen, J. Y., Habets, E. A., & Naylor, P. A. (2008). Blind estimation of reverberation time based on the distribution of signal decay rates. In 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 329–332.