Blind Estimation of Room Volume from Reverberant Speech Based on the Modulation Transfer Function

Nutchanon Siripool[1], Suradej Duangpummet[2], Jessada Karnjana[2], Waree Kongprawechanon[1], and Masashi Unoki[3]
[1] Sirindhorn International Institute of Technology, Thammasat University, Thailand
E-mail: d6622300249@g.siit.tu.ac.th, waree@siit.tu.ac.th
[2] National Science and Technology Development Agency, Thailand
E-mail: {suradej.dua, jessada.kar}@nectec.or.th)
[3] School of Information Science, Japan Advanced Institute of Science and Technology, Japan
E-mail: unoki@jaist.ac.jp

***Abstract*—Estimating room volume is crucial in acoustic engineering, architectural acoustics, and emergency response. Traditional methods relying on the Sabine equation face challenges due to the difficulty in accurately measuring absorption coefficients. This paper introduces a method for blindly estimating room volume using reverberant speech on the basis of the modulation transfer function (MTF). Our approach processes reverberant speech through a seven-octave band filterbank, extracting temporal amplitude envelopes to form MTF features and input the MTF feature into a convolutional neural network to estimate room volume. Experimental results on various real-world and simulated room impulse responses demonstrate our method's effectiveness, achieving a mean square error of 0.14 in volume estimation. The mean absolute logarithm of the ratio between the estimated volume and the ground truth is 1.15. This approach significantly improves accuracy and robustness over the baseline method, offering a practical solution for scenarios where direct measurements are impractical. Potential applications include security monitoring, space management, and emergency evacuation planning.**

I. Introduction

Estimating room volume is essential in various fields, including acoustic engineering, where it is used to design sound systems [1], optimize room acoustics [2], and create soundproofing for any room. It also plays a critical role in emergency response scenarios, such as rescue operations during fires. Knowing the room volume (V) helps assess the capacity of the target room, estimate the oxygen requirements for firefighters, and determine the number of people potentially trapped inside a smoke-filled room. Additionally, room volume needs to be accurately estimated for architectural acoustics [2], influencing the design of concert halls [1], lecture rooms, and recording studios to enhance sound quality and speech intelligibility.

Traditional methods for calculating room volume often rely on the Sabine equation as defined as

$$V \approx \left(\frac{RT_{60} \times 6.4\bar{\alpha}}{0.161} \right)^3, \tag{1}$$

which necessitates knowing the reverberation time (RT_{60}) and average absorption coefficient of room surface $(\bar{\alpha})$ [3]. However, measuring the absorption coefficient accurately in real-world situations can be challenging. Obtaining precise absorption coefficients involves finding out details of the materials and their properties within the room [3], which is often impractical and time-consuming [4]. Consequently, a more efficient and reliable approach is needed to estimate room volume.

Estimating room volume from reverberant speech is crucial in acoustics and speech processing. However, direct measurements or manual calibration are often impractical. Hence, many researchers have proposed methods for estimating room volume [5]–[9].

Recently, there have been many interesting developments in convolutional neural networks (CNNs) in the acoustic field. Genovese *et al.* [7] employed CNNs for blindly estimating room volume from single-channel noisy speech, achieving accurate results across various room sizes. Their approach focuses on critical parameters such as RT_{60} and direct-to-reverberant ratio (DRR). However, this model's performance decreased significantly on a separate, measured data corpus of unseen rooms, speech, and ambient noise. Vaswani *et al.* [9] introduced transformers, highlighting attention mechanisms for sequence modeling that enhance the analysis of reverberant speech characteristics, thereby improving room volume estimation accuracy. The paper employs data augmentation techniques such as SpecAugment [6]. This indicates that the base model may not perform optimally without these enhancements. Additionally, Shabtai *et al.* [5] explored room-volume classification from reverberant speech. Although the method is promising, it is less accurate and reliable than standard measurement techniques, requiring further investigation to ensure robustness in practical applications.

Previously, the use of modulation transfer function (MTF) based CNNs for blindly estimating room acoustic parameters [10] was demonstrated by simultaneously predicting various parameters, such as RT_{60} and speech transmission index (STI) [11], [12]. Although this method can estimate various room acoustic parameters, it still needs to catch up in determining the room volume, which is a crucial area of interest.

Therefore, we extend our previous methodologies by fo-

cusing on blind estimation of room volume from reverberant speech using MTF features extracted via a seven-octave band filterbank. The MTF feature extracted from reverberant speech signals can be explained by the concept of the MTF [13]. We aim to provide a method that blindly estimate room volume robustly and accurately.

II. Background

A. Modulation Transfer Function

MTF, a transfer function of a linear system, characterizes a transmission channel on the basis of modulation frequency and the reduction in modulation depth [14]. It was introduced in room acoustics to evaluate how an enclosure affects speech intelligibility. In room acoustics, MTF quantifies the impact of reverberation; increased reverberation results in a lower modulation depth for modulated signals passing through the room. The ratios of modulation distortion between the input envelopes and their corresponding outputs are referred to as modulation indices. The MTF is defined as

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

where $h(t)$ denotes the room impulse response.

B. Temporal Amplitude Envelopes

The temporal amplitude envelope (TAE) can reveal information about the reverberation characteristics of a room, such as the RT_{60} [15], which are influenced by the room's volume. The TAE is a smoothed version of the input signal in reverberant environments. It provides essential information regarding the room's acoustic parameter [16]. In this work, we extracted the TAE using Eq. (3), which involves the Hilbert transform and a low-pass filter (LPF). The Hilbert transform is a mathematical operator used to analyze a real-valued signal, which is useful for extracting the amplitude and phase of the signal [17]. The LPF is a sixth-order Butterworth filter with a cut-off frequency of 30 Hz, and the signal is downsampled to 60 Hz to reduce complexity.

$$e_y(t) = \mathrm{LPF}\left[|\mathrm{y(t)} + j\mathrm{Hilbert(y)(t)}|\right]. \tag{3}$$

We conduct a preliminary study to establish a foundational understanding and verify the feasibility of estimating room volume using RT_{60} through the Sabine Eq. (1) and Eq. (4) from Kuster *et al.* [18], without relying on α or any specific coefficients. The goal was to determine whether using only RT_{60} could accurately map to room volume.

$$RT_{60} = 0.34 \ln(V) - 1. \tag{4}$$

This experiment uses publicly available data of 11 rooms (RT_{60} and room volume) [19][20], as shown in Table I.

The GT is defined as ground truth, and TableI shows the error between GT and the values calculated from Eqs. (1) - (4). Volume* represents the volume m^3 calculated from Eq. (1), while Volume** represents the volume calculated from Eq. (4).

TABLE I: Preliminary experiment result.

RT_{60} (GT)	Volume (GT)	Volume**	Volume*	Square Error**	Square Error*
0.08	11.88	23.96	0.57	128.02	145.98
0.21	15.00	35.12	10.22	22.81	404.89
0.33	47.30	50.28	40.40	47.59	8.89
0.37	12.00	56.23	55.92	1929.20	1956.10
0.37	246.00	56.39	56.38	35956.78	35950.65
0.39	48.30	59.63	65.49	295.50	128.47
0.47	99.60	76.12	116.83	296.98	551.16
0.64	202.00	123.67	286.71	7175.96	6135.00
0.65	72.90	126.62	297.63	50504.55	2885.65
0.70	18.00	148.41	378.68	130092.63	17007.59
1.22	370.00	685.00	2004.76	2672431.83	99221.95
			RMSE	154.78	36.86

The results show both the square error and root mean square error (RMSE). As shown in Table I, both Eqs. (1) - (4) yield very large errors. Therefore, we can conclude that we cannot accurately define room volume using the Sabine-based equation alone. This led us to utilize another method to estimate room volume without relying on the Sabine equation and α.

III. Proposed Method

In this section, we will explain our proposed framework in detail, including the input features and the architecture of the CNN model.

A. Proposed Framework

In our paper, we present a method for estimating room volume through a blind experiment. The proposed framework uses reverberant speech to assess the room volume, leveraging advanced signal processing and machine learning techniques. Our proposed framework is shown in Fig. 1.

The reverberant speech is input into a seven-octave band filterbank, which covers frequencies of 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz. Each frequency band is processed to extract the TAEs. TAE 1 corresponds to the 125 Hz band, TAE 2 to the 250 Hz band, and so on, up to TAE 7 for the 8 kHz band.

Next, the seven TAEs are combined to form the MTF feature, which has dimensions of 7×240. This MTF feature encapsulates the essential characteristics of the reverberant speech across the different frequency bands, comprehensively representing the acoustic environment.

The MTF feature is fed into a CNN, which is trained to process this high-dimensional input. Then, the CNN estimates the room volume. The deep learning model leverages the rich feature set provided by the MTF, enabling it to learn complex patterns and relationships that correlate with the room's volume.

The proposed framework offers a sophisticated approach to estimate room volume, moving beyond traditional methods that rely heavily on specific coefficients or prior knowledge of the room's acoustic properties. By utilizing reverberant speech and advanced signal processing techniques, our method provides a robust solution for blind estimation of room volume, with potential applications in various real-world scenarios.

For instance, this methodology can be applied to predict a room's current occupancy. By calculating the difference between the actual room volume and the estimated volume

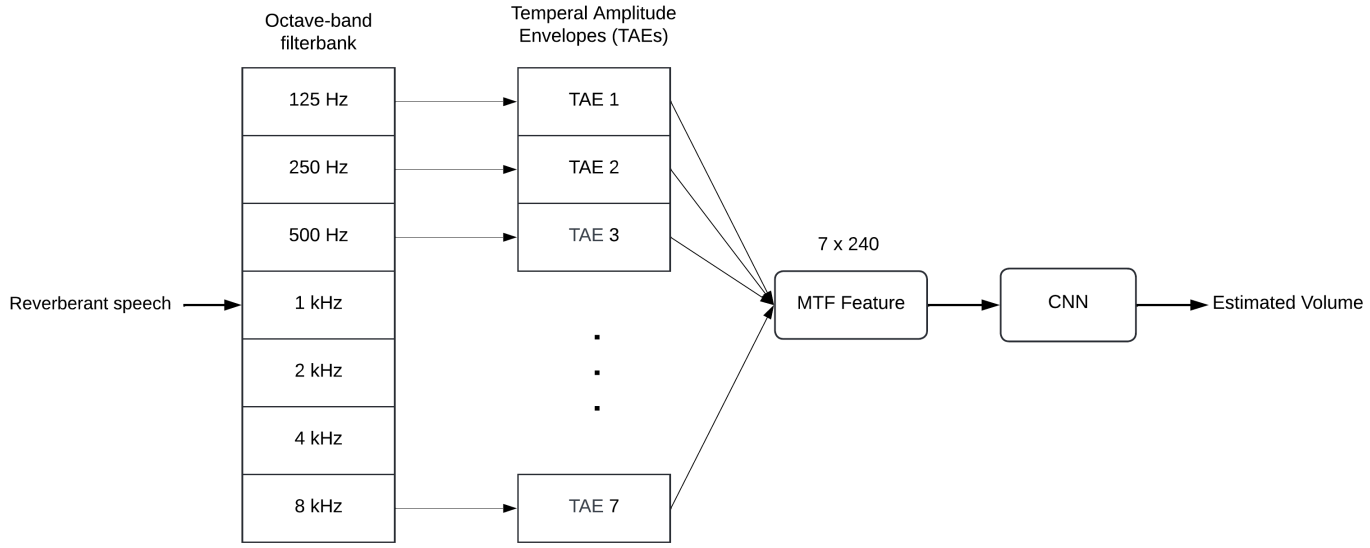

Fig. 1: Proposed framework for blind room volume estimation.

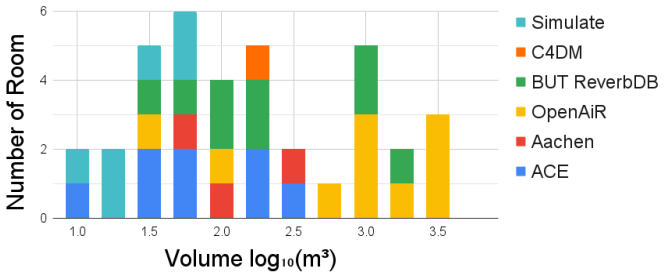

Fig. 2: Numbers of rooms for different room volumes.

(Which is influenced by the presence of people), the number of individuals present can be inferred. This application is particularly beneficial for security monitoring, space management, and emergency evacuation planning, showcasing the practical utility of our proposed framework.

B. MTF Feature

The critical input feature in our framework is the MTF feature, a sophisticated representation derived from the temporal characteristics of reverberant speech. The MTF captures the modulation properties of the speech signal across different frequency bands, providing rich, multidimensional input for the subsequent processing stages.

The construction of the MTF feature begins with the extraction of TAEs from the reverberant speech signal. The reverberant speech is processed through a seven-octave band filterbank covering 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz. Each frequency band's TAE is computed, resulting in seven distinct TAEs.

These TAEs capture the amplitude fluctuations over time within their respective frequency bands, reflecting the dynamic characteristics of the reverberant speech signal. Combining these seven TAEs forms the basis of the MTF feature, encapsulating the modulation properties across a broad frequency spectrum.

The MTF feature combines the seven TAEs into a single multidimensional representation. Specifically, each TAE is represented as a vector of length 240, capturing the temporal modulation details at a high resolution. The concatenation of these seven TAE vectors results in the MTF feature with dimensions of 7×240. This high-dimensional feature vector comprehensively represents the temporal modulation characteristics of the reverberant speech signal across the different frequency bands.

The MTF feature is the input to our CNN, designed to process this rich, multidimensional input and estimate the room volume accurately. The CNN leverages the detailed modulation information encapsulated in the MTF feature, enabling it to learn complex patterns and relationships that correlate with the room's volume.

C. CNN Model Architecture

The CNN architecture has been utilized in this work because it can produce two-dimensional input signals. Figure 3 shows the CNN architecture with MTF input and one single output. The model has a total of four convolutional layers followed by an average pooling layer. The first two convolutional layers will be followed by batch normalization. Batch normalization helps accelerate the training process by allowing the network to use higher learning rates. Also, batch normalization makes the training process more stable [21]. For more details on the CNN architecture model, visit our [1]GitHub repository.

IV. Experimental Setup

A. Dataset

As shown in Fig. 2, five publicly available real-world RIR datasets, using 31 actual rooms, are selected to cover a wide

[1]https://github.com/winnerrock/APSIPA2024

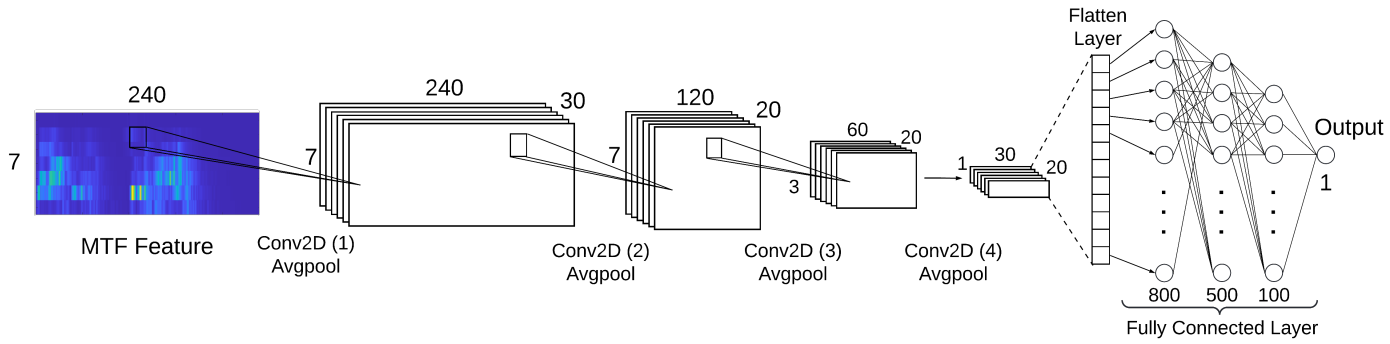

Fig. 3: Proposed CNN architecture model with the MTF feature.

range of realistic acoustic room parameters. The RIR data collection locations include offices, meeting rooms, lecture rooms, churches, and more. These datasets contain the ACE challenge dataset [19], the Aachen Impulse Response (AIR) dataset [20], the Brno University of Technology Reverb database (BUT ReverbDB) [22], the OpenAIR dataset [23], and the C4DM dataset (C4DM) [24], which cover room volumes between 12 and 4500 cubic meters. All RIRs are resampled to have a sampling rate of 16 kHz. Moreover, we simulate six additional rooms and RIRs on the basis of the Image-Source Method and HRTF Interpolation technique [25] to compensate for the lack of data range.

V. Experimental Setup and Results

A. Evaluation Metrics

The room volume estimation problem was formulated as a regression task on the $\log_{10}$ of the room volume, ensuring that the estimation error was proportional to the magnitude of the room size. Due to the significant variability in room sizes, a logarithmic approach was deemed more appropriate than a linear one. The performance evaluation involved three widely used metrics: mean squared error (MSE), mean error (bias), and Pearson's correlation coefficient (ρ). Additionally, this study utilized a specific metric that is based on the mean absolute logarithm of the ratio between the estimated volume $\hat{V}$ in m^3 and the ground truth volume V. The MSE is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{5}$$

where N is the number of observations in the dataset, y_i is the actual value of the observation, and $\hat{y}_i$ is the predicted value of the i-th observation [26].

The mean absolute error (MAE) is determined by:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{6}$$

where $|y_i - \hat{y}_i|$ represents the absolute error.

Pearson's correlation coefficient (ρ) is computed using the formula:

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{7}$$

where ρ is the correlation coefficient, x_i are the values of the ground truth in a sample, $\bar{x}$ is the mean of the ground truth, y_i are the values of the estimate value in a sample, and $\bar{y}$ is the mean of the estimate value values [27].

The mean absolute logarithm of the ratio between the estimated volume and the ground truth is (MM), which is determined by:

$$\text{MM} = e^{\frac{1}{N}\sum_{i=1}^{N}\left|\ln\left(\frac{\hat{V}_i}{V_i}\right)\right|}, \tag{8}$$

where N denotes the number of test samples and $\ln$ is the natural logarithm. This metric summarizes the error in terms of the average multiple of the estimated volume compared to the true volume [7].

B. Results

In this section, we present a detailed analysis of the performance of our proposed method for blindly estimating room volume from reverberant speech. The results are compared with those of the method proposed by Genovese *et al.* [7] to highlight the improvements and efficacy of our approach.

Table II summarizes the performance metrics, including MSE, MAE, ρ, and MM. Our proposed method significantly reduced MSE, achieving a value of 0.139, compared to 0.190 by Genovese *et al.* This indicates that our method more accurately estimates volume with minor deviations from the actual values.

Additionally, our method achieves a correlation coefficient ρ of 0.792, which is considerably higher than the 0.39 obtained by Genovese *et al.*, suggesting a stronger linear relationship between the estimated and actual volumes. The MM value of 1.148 further corroborates the accuracy of our estimates, being closer to 1, which implies less deviation from the actual room volumes.

Figure 5 presents a scatter plot of the predicted room volumes versus the ground truth volumes. The distribution

TABLE II: Error comparison between the baseline and our proposed method

Methods	MSE	MAE	ρ	MM
Genovese *et al.*[7]	0.190	-	0.390	2.280
Proposed Method	0.139	0.322	0.792	1.148

of points along the diagonal line indicates a high degree of correlation, reinforcing the quantitative findings. Most predictions are closely aligned with the actual values, showing the reliability of our method.

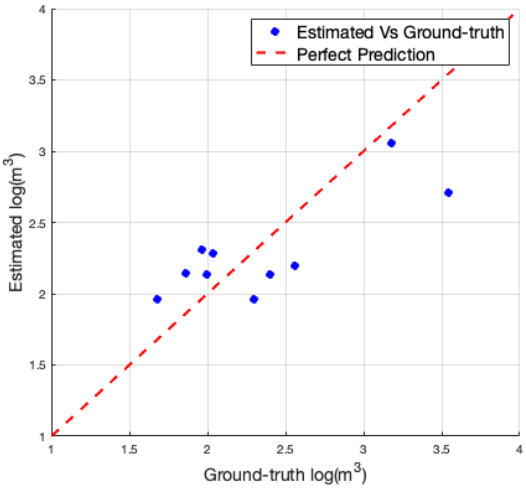

Fig. 4: Results of estimate volumes and ground truth volumes.

Figure 6 provides a bar chart comparing the logarithmic estimates of room volumes with the ground truth across various room sizes. The close match between the estimated and actual values across different volume ranges demonstrates the robustness of our method. This consistency across different room sizes indicates that our approach can generalize well to diverse acoustic environments.

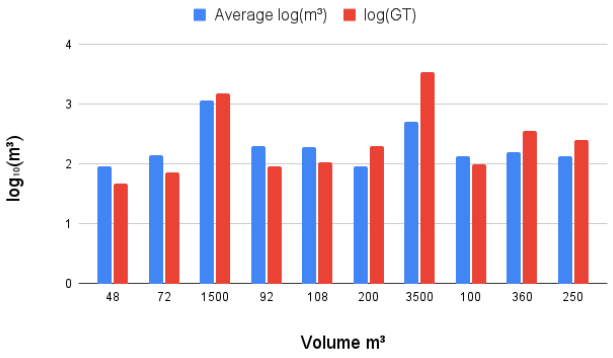

Fig. 5: Comparison between estimate and ground truth room volume.

VI. Discussion

The results demonstrate that our proposed method can effectively estimate room volume blindly from reverberant speech. The performance metrics, including MSE, ρ, and MM, indicate good results of 0.139, 0.792, and 1.148, respectively. Our approach outperforms the baseline method by Genovese *et al.*[7], although it is important to note that our work is currently based on reverberant speech only, not noisy speech. We plan to extend our method to handle noisy speech in the following research phase. We might also use other room acoustic parameters for the feature.

VII. Conclusion

This paper presented a method for blindly estimating room volume from reverberant speech. By leveraging the modulation transfer function (MTF) feature in conjunction with a CNN, our method significantly improved accuracy and reliability compared to the baseline method according to the experiment results. This is evident from notable reductions in MSE, MM, and MAE and a higher ρ. Qualitative evaluations further reinforced these findings, showing a solid alignment between predicted and actual room volumes.

However, our method is based on reverberant speech rather than noisy reverberant speech. In future work, we will focus on noisy reverberant speech. Also, additional acoustic features could be added to improve the robustness and accuracy of room volume estimation. Expanding the dataset to include more diverse rooms could help generalize and improve the model. Integrating occupancy estimation techniques to assess the number of people in a room on the basis of deviations from the known empty room volume could provide valuable insights for building management and security systems.

In conclusion, by providing a framework for blindly estimating room volume from reverberant speech, our proposed method offers a significant step forward in room acoustic analysis. With continued research and development, this approach has the potential to estimate other room acoustic parameters simultaneously. This method will enhance the capabilities of acoustic scene analysis, sound field control, and architectural acoustics.

Acknowledgments

This work was supported by a grant from the SIIT-JAIST-NSTDA Dual Doctoral Degree Program.

References

[1] R. Michael, *Model-based sound design and audio signal processing using State-space frequency responses*. Springer, 2019.

[2] F. Jacobsen, *Elements of Acoustical Engineering*. Springer, 2013.

[3] F. A. Everest and K. C. Pohlmann, *Master Handbook of Acoustics*. McGraw-Hill Education, 2014.

[4] O. Robin, A. Berry, O. Doutres, and N. Atalla, "Measurement of the absorption coefficient of sound absorbing materials under a synthesized diffuse acoustic field," vol. 136, no. 1, EL13–EL19, 2014.

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

[6] D. S. Park, W. Chan, Y. Zhang, *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[7] A. F. Genovese, H. Gamper, V. Pulkki, N. Raghuvanshi, and I. J. Tashev, "Blind room volume estimation from single-channel noisy speech," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 231–235.

[8] P. Srivastava, A. Deleforge, and E. Vincent, "Blind room parameter estimation using multiple multichannel speech recordings," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 2021, pp. 226–230.

[9] C. Wang, M. Jia, M. Li, C. Bao, and W. Jin, "Attention is all you need for blind room volume estimation," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 1341–1345.

[10] S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Blind estimation of speech transmission index and room acoustic parameters based on the extended model of room impulse response," *Applied Acoustics*, vol. 185, p. 108 372, 2022.

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

[12] M. Unoki, A. Miyazaki, S. Morita, and M. Akagi, "Method of blindly estimating speech transmission index in noisy reverberant environments.," *J. Inf. Hiding Multim. Signal Process.*, vol. 8, no. 6, pp. 1430–1445, 2017.

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

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

[15] M. Unoki and Z. Zhu, "Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 233–244, 2020.

[16] L. Wang, S. Duangpummet, and M. Unoki, "Blind estimation of speech transmission index and room acoustic parameters by using extended model of room impulse response derived from speech signals," *IEEE Access*, vol. 11, pp. 49 431–49 444, 2023. DOI: 10.1109/ACCESS.2023.3276327.

[17] F. R. Kschischang, "The hilbert transform," *University of Toronto*, vol. 83, p. 277, 2006.

[18] M. Kuster, "Reliability of estimating the room volume from a single room impulse response," vol. 124, no. 2, pp. 982–993, 2008.

[19] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ace challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[20] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *2009 16th International Conference on Digital Signal Processing*, IEEE, 2009, pp. 1–5.

[21] W. Jung, D. Jung, B. Kim, S. Lee, W. Rhee, and J. H. Ahn, "Restructuring batch normalization to accelerate cnn training," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 14–26, 2019.

[22] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[23] D. T. Murphy and S. Shelley, "Openair: An interactive auralization web resource and database," in *Audio Engineering Society Convention 129*, Audio Engineering Society, 2010.

[24] R. Stewart and M. Sandler, "Database of omnidirectional and b-format room impulse responses," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 165–168.

[25] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," vol. 138, no. 2, pp. 708–730, 2015.

[26] GeeksforGeeks, *Mean squared error*, Accessed: 2024-07-13. [Online]. Available: https://www.geeksforgeeks.org/mean-squared-error/.

[27] N. University, *Strength of correlation*, Accessed: 2024-07-13. [Online]. Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/strength-of-correlation.html.