


METHODOLOGY

Open Access



# Exploring the power of pure attention mechanisms in blind room parameter estimation

Chunxi Wang<sup>1</sup>, Maoshen Jia<sup>1\*</sup> , Meiran Li<sup>1</sup>, Changchun Bao<sup>1</sup> and Wenyu Jin<sup>2</sup>

## Abstract

Dynamic parameterization of acoustic environments has drawn widespread attention in the field of audio processing. Precise representation of local room acoustic characteristics is crucial when designing audio filters for various audio rendering applications. Key parameters in this context include reverberation time ( $RT_{60}$ ) and geometric room volume. In recent years, neural networks have been extensively applied in the task of blind room parameter estimation. However, there remains a question of whether pure attention mechanisms can achieve superior performance in this task. To address this issue, this study employs blind room parameter estimation based on monaural noisy speech signals. Various model architectures are investigated, including a proposed attention-based model. This model is a convolution-free Audio Spectrogram Transformer, utilizing patch splitting, attention mechanisms, and cross-modality transfer learning from a pretrained Vision Transformer. Experimental results suggest that the proposed attention mechanism-based model, relying purely on attention mechanisms without using convolution, exhibits significantly improved performance across various room parameter estimation tasks, especially with the help of dedicated pretraining and data augmentation schemes. Additionally, the model demonstrates more advantageous adaptability and robustness when handling variable-length audio inputs compared to existing methods.

**Keywords** Acoustic environments, Blind room parameter estimation, Pure attention mechanisms

## 1 Introduction

In recent years, there has been a growing focus on the dynamic parameterization of evolving acoustic environments. The parameters that describe local rooms or other acoustic spaces hold significance as they can be harnessed in the modeling and design of audio filters for a diverse range of applications. Understanding the specific acoustic properties of the surrounding room can be applied to improve speech signals and support dereverberation algorithms, ultimately improving word error

rate for automatic speech recognition (ASR) and the clarity of voice communication systems [1–3]. Additionally, spatial sound reproduction systems could leverage this data to enhance their performance in tasks related to acoustic room equalization either using predefined filters [4, 5] or in an adaptive manner [6].

Furthermore, the successful realization of audio augmented reality (AAR) necessitates the seamless integration of virtual acoustic objects into the physical environment. This integration underscores the importance of achieving a harmonious alignment between the acoustical properties of virtual elements and the characteristics of the local space [7]. In pursuit of this goal, a significant challenge lies in the accurate estimation of related acoustical parameters of a room to enhance the realism of immersive audio. Notably, Jot et al. [8] introduced the concept of a “reverberation fingerprint,”

\*Correspondence:

Maoshen Jia  
jjamaoshen@bjut.edu.cn

<sup>1</sup> Beijing Key Laboratory of Computational Intelligence and Intelligent System, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

<sup>2</sup> AcousticDSP Consulting LLC, St Paul, MN, USA



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

comprising the room's volume and its frequency-dependent diffuse reverberation decay time. This innovative concept was proposed to characterize rooms specifically for the realistic binaural rendering achievable with audio augmented reality headphones. It is worth noting that this fingerprint primarily focuses on the part of reverberation that is independent of the position, treating a room's acoustic characteristics in isolation from the orientation and directivity of sound sources and receivers.

Conventionally, room parameters like reverberation time ( $RT_{60}$ ) and direct-to-reverberant ratio (DRR) are typically obtained through a direct analysis of measured room impulse responses (RIRs). Meanwhile, room volume is closely linked to the determination of a concept known as the "critical distance." This critical distance is defined as the distance at which the direct and reverberant power components of a sound source become equal, effectively making the DRR reach 0 dB. In cases where we assume an ideal diffuse soundfield, the relationship between these parameters is mathematically described by Sabine's well-known equation [9]:

$$RT_{60}(b) \approx 0.16 \frac{V}{\alpha(b) \cdot S}, \quad (1)$$

where  $S$  denotes the total area of the room's surfaces, and  $\alpha(b)$  is the area-weighted mean absorption coefficient in octave band  $b$ .

In practical applications, conducting in situ measurements of RIRs and determining the volumes of users' local acoustic spaces can often present significant challenges. A compelling alternative involves blind estimation of room acoustic parameters from audio recordings obtained using microphones, even when the sound sources are unknown and in the presence of background noise. The 2015 ACE challenge [10] established a benchmark for blind estimation of  $RT_{60}$  and DRR from noisy speech recordings. The leading systems in this challenge primarily relied on signal modeling-based approaches [11, 12]. Meanwhile, room volume estimation has long been formulated as a classification problem [13, 14]. Audio forensics systems described in [13, 14] make use of Mel-frequency cepstral coefficients (MFCC)-based features to identify the specific room associated with an environmental sound or speech recording, typically within a predefined closed set of possibilities.

Due to the recent advancements in deep neural networks (DNNs), there is a growing trend to reframe the challenge of blind room acoustic parameter estimation as a regression problem. This approach leverages convolutional neural network (CNN) models in combination with time-frequency representations, offering an increasingly relevant and effective solution. Gamper et al. [15] introduced a CNN designed to directly estimate  $RT_{60}$

from a four-second recording of reverberant speech. The experimental results demonstrate that this CNN outperforms other methods in the ACE challenge, offering both superior performance and higher computational efficiency. The same approach was also applied to blind volume estimation in [16], and results show that it can estimate a broad range of volumes from real-measured data (with average estimated errors typically ranging from half to twice the actual values). CNN-based systems with similar methodologies have been put forward for the blind estimation of room acoustic parameters, utilizing either single-channel [17–19] or multi-channel speech signals [20]. These systems have showcased promising outcomes in terms of both accurate parameter estimation and resilience to temporal variations in dynamic acoustic environments. Notably, in contrast to the conventional approach of log-energy calculations for spectro-temporal features used in prior studies, Ick et al. [21] introduced a set of phase-related features. Their research demonstrated clear improvements in the context of estimating reverberation fingerprints for real-world rooms that had not been previously seen, highlighting the enhanced efficacy of this method.

CNNs are widely considered in the fore-mentioned approaches due to their suitability for learning two-dimensional time-frequency signal patterns for end-to-end modeling. CNNs can be extended by a recurrent layer to form convolutional recurrent neural networks (CRNN) that exploit sequential dependencies in the data [22] and improve the capability of processing input sequences of variable length [23]. To further enhance the capture of long-range global context, hybrid models combining convolutional neural networks (CNNs) with self-attention mechanisms have yielded state-of-the-art results in a range of tasks, including acoustic event classification [24, 25] and various audio pattern recognition endeavors [26, 27]. Gong et al. [28] pushed the boundaries even further by introducing purely attention-based models for audio classification. Their creation, the Audio Spectrogram Transformer (AST), was evaluated on several audio classification benchmarks, achieving new state-of-the-art results. This underscores that CNNs may not always be essential in this particular context.

Building on the inspiration derived from the research presented in [28], our study introduces a convolution-free, purely attention-based model for the blind estimation of acoustic room parameters by extending our previous work in [29]. To the best of our knowledge, this marks the inaugural application of an attention-based system in the field of blind acoustic room parameter estimation. The proposed system utilizes Gammatone magnitude spectral coefficients as well as the low-frequency phase spectrogram as inputs and captures long-range

global context, even in the lower layers of the model. Furthermore, to enhance system performance, we apply transfer learning through the use of a pretrained transformer model from ImageNet. For the evaluation of the proposed method, we curate a RIR corpus that includes publicly available RIRs, synthesized RIRs, and RIRs obtained through in-house measurements of real-world rooms. Experimental results clearly demonstrate the superiority of our proposed model when compared to CNN-based blind acoustic parameter estimation systems, particularly when dealing with previously unseen real-world rooms using single-channel recordings of variable length.

The remainder of the article is organized as follows.

Section 2 introduces the construction of RIR datasets, including real-world and simulated datasets. Section 3 demonstrates the generation of audio data with reverberation and noise using constructed RIR datasets, followed by data preprocessing, augmentation, and feature extraction schemes for neural network training. Section 4 details the model structures of a CNN-based model, a CRNN-based model, the proposed attention-based systems. Section 5 conducts a comprehensive evaluation of the proposed system against state-of-the-art methods in various room parameter estimation tasks and its performance under variable-length inputs. Section 6 draws the conclusion.

## 2 Data generation pipeline

Applying neural network methods to address blind room parameter estimation is a challenging task, as it generally requires a substantial amount of data. Since this task necessitates the need of having audio samples from rooms with various acoustic characteristics, manually creating a suitably diverse dataset would incur exorbitant costs and time. In this work, audio samples are created from public real-world RIR datasets, the BJUT Reverb dataset, and a room-simulation-based RIR dataset.

### 2.1 Public real-world RIR datasets

In this study, six publicly available real-world RIR datasets that include 44 authentic rooms are considered, with the aim of encompassing a wide range of acoustic room parameters.

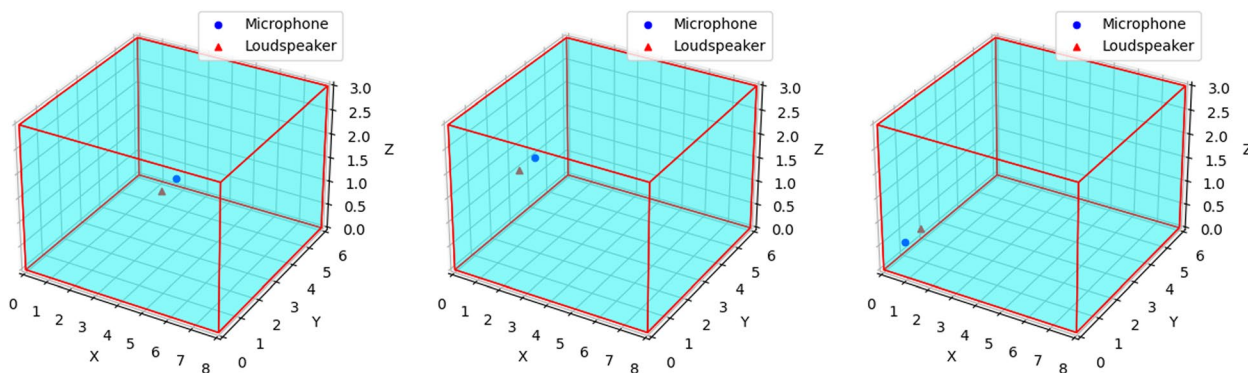
The majority of the data targets at geometrically regular rooms, including spaces like offices, classrooms, and auditoriums/lecture halls. These datasets include the ACE Challenge dataset [10], the Aachen Impulse Response (AIR) dataset [30], the Brno University of Technology Reverb Database (BUT ReverbDB) [31], the C4DM dataset [32], and the dEchorate dataset [33]. Additionally, the OpenAIR dataset [34] primarily covers larger acoustic spaces, such as churches, nuclear reactor halls, and other substantial structures. As a result, a large variety of real-world room configurations with different volume parameters are incorporated.

Furthermore,  $RT_{60}$  values vary widely, ranging from less than half a second to over ten seconds, and these values are calculated using the Schroeder method [35].

In addition to the public datasets described above, RIRs from 11 distinct rooms at the campus of Beijing University of Technology were measured, including elevator shafts, classrooms, auditoriums, seminar rooms, and more. The parameters of these selected rooms were recorded. The aim of this endeavor is to bridge the natural gap in available real-world acoustic spaces within the volume range of  $12 \text{ m}^3$  to  $7000 \text{ m}^3$ . Three RIR measurements were conducted at different positions within the selected rooms. Specifically, measurements were taken at the geometric center of the room, a location near the wall, and a position near the corner, to capture the RIR with a sequence length of 4 s. The microphone and loudspeaker positions are illustrated in Fig. 1.

### 2.2 Simulated RIR dataset

The real-world data is supplemented by introducing 30 simulated RIRs derived from virtual rooms with various



**Fig. 1** Room measurement layout diagram

geometries. This aims to enhance the dataset's representation of less frequently encountered room volumes, thereby achieving a normal distribution of total volume.

The specific approach involves simulating a single sound source positioned near the center of each virtual room and evenly distributing five-point receivers throughout the volume of each room. To create this synthetic dataset, the *pyroomacoustics* [36] software package is deployed, which utilizes the image-source model to simulate RIRs for rooms with specific volumes. Although this geometric model does not account for phenomena like diffraction and scattering, empirical evidence demonstrates that the utilization of simulated data contributes to enhancing the model's performance, enabling it to effectively generalize to real-world data [16].

### 3 Preprocessing

In this section, we provide a detailed explanation of how audio data with reverberation and noise is generated. We started with convolving acoustic response with audio signals and adding various types of noises for subsequent neural network comparisons.

To ensure the quality and consistency of the dataset, we performed a series of data preprocessing. Firstly, we partitioned the audio signals into training, validation, and test sets. Only real-world RIRs were selected in the test set to assess system performance on unseen non-simulated rooms.

Furthermore, we employed a data augmentation technique called SpecAugment that aims to enhance the neural network's ability to generalize in unknown rooms and noisy environments.

Lastly, we discussed the method for audio feature extraction. Gammatone ERB filterbank was used to generate time-frequency representations. After processing, these features resulted in a two-dimensional feature block used as input to the neural network, allowing it to handle various datasets and provide accurate blind room parameter estimation performance.

#### 3.1 Audio generation

In the acquired RIR dataset, a total of 55 real-world rooms and 30 simulated rooms are included, comprising a total of 570 RIRs. The volume labels span from 11.88 m<sup>3</sup> to 21,000 m<sup>3</sup>, while the range of RT<sub>60</sub> varies from 0.41 s to 19.68 s. Due to the significant differences in volume labels spanning multiple orders of magnitude, we chose to represent them using a logarithmic base 10 scale. Additionally, to ensure consistency across all datasets, all RIRs were downsampled to 16 kHz. The distribution of volume and RT<sub>60</sub> in different datasets is shown in Fig. 2.

From a given RIR dataset with room parameter labels, we generated audio data with reverberation and noise for the purpose of feeding it into different neural networks for comparison. To achieve this, we mapped the acoustic response  $r(t)$  of different types of rooms in the RIR dataset onto the audio signal  $y(t)$ .

We used source speech signals  $x(t)$  recorded in anechoic chambers without reverberation and convolve them with  $r(t)$  in the time domain. The source speech signals  $x(t)$  are obtained from the ACE dataset [10], where samples are recorded without reverberation, and include both male and female speakers. In the RIR dataset, some rooms have more RIR measurements than others. To ensure a uniform representation, each room is equally sampled, so that the distribution of audio samples in our dataset matches the volume distribution in our RIR dataset.

Additional noise signals  $n(t)$  were added to simulate recordings at four different signal-to-noise ratio (SNR) levels, including [+ 30, + 20, + 10, + 0] decibels. The noise  $n(t)$  comprises two types of noise, namely white noise and babble noise [37].

In summary, the audio signal  $y(t)$  is constructed by convolving source speech signals  $x(t)$  with room impulse response  $r(t)$  and adding additional noise  $n(t)$ , represented as:

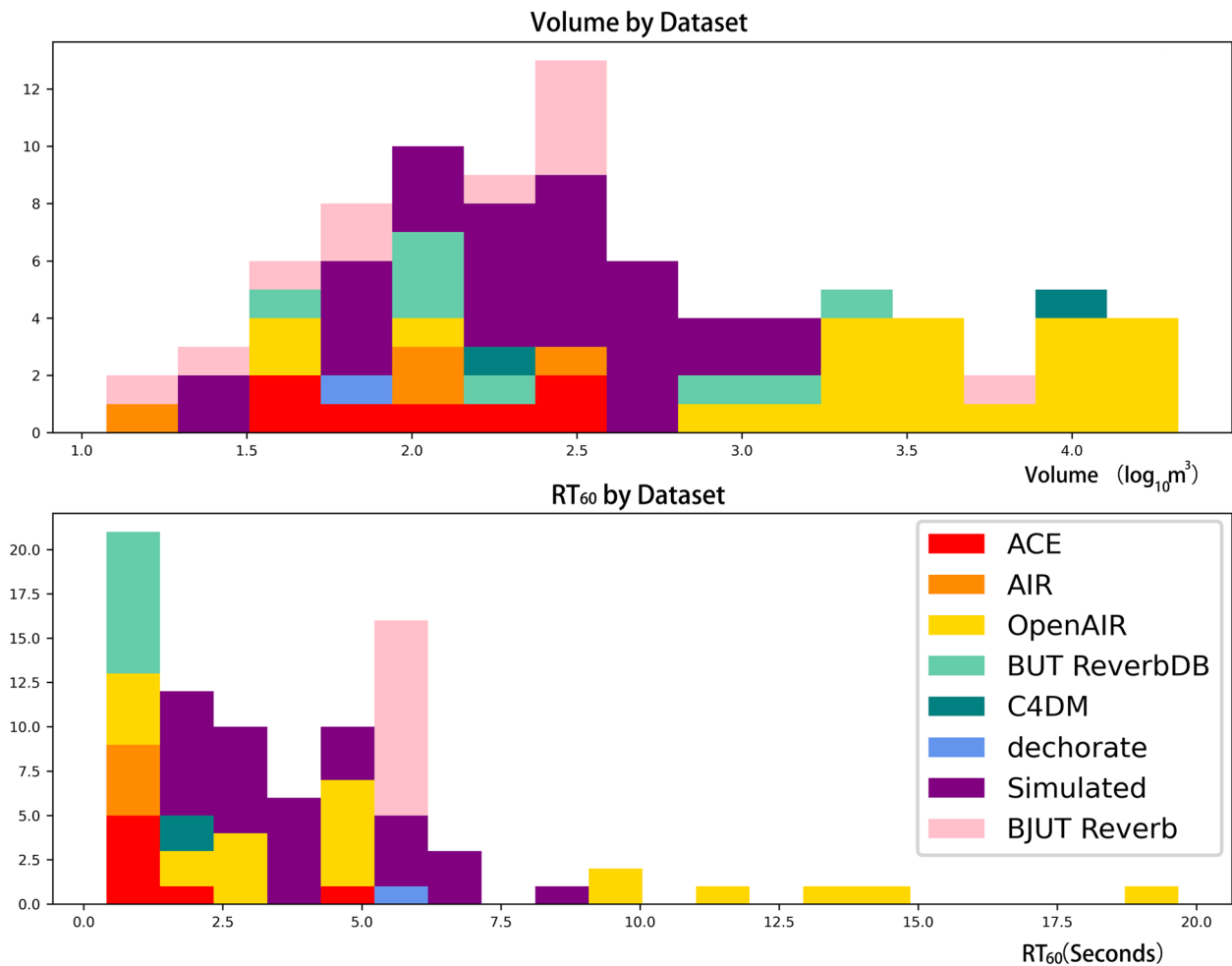
$$\begin{aligned} y(t) &= x(t) * r(t) + n(t) \\ &= s(t) + n(t) \end{aligned} \quad (2)$$

Here,  $t$  represents the discrete time index.  $s(t)$  represents the reverberation audio signal without noise.

We split 32,000 audio signals  $y(t)$  into training, validation, and test sets in a 6-2-2 ratio. During the training process, we randomly sampled a subset from both real and simulated rooms for room validation. A subset was also extracted from real-world unseen rooms for room testing. Rooms with these specific parameters were not included in the training set. The purpose of this step is to assess whether the model, when confronted with room parameters not encountered during the training process, can still demonstrate robust predictive performance under noisy and reverberant conditions. Overall, *Dataset I* is formulated as listed in Table 1.

#### 3.2 Audio augmentation

To enhance the generalizability of neural networks in unknown rooms and noisy environments, we employed the widely-used data augmentation technique known as SpecAugment [38]. This method enhances the model's robustness to unknown conditions through modifications and augmentations to the available training data. Specifically, we selected reverberation signals without noise  $s(t)$  as described in Eq. 2. Subsequently, these audio signals



**Fig. 2** The distribution graphs of volume and RT<sub>60</sub> in different datasets

**Table 1** Summary of data splits for datasets I and II

Data Split	# of Dataset I	# of Dataset II	Real Rooms	Simulated Rooms
Train	19200	24000	34	18
Validation	6400	6400	21	12
Test	6400	6400	21	0

were transformed into log Mel-frequency spectrograms and subjected to time warping, frequency masking, and time masking, as shown in Fig. 3.

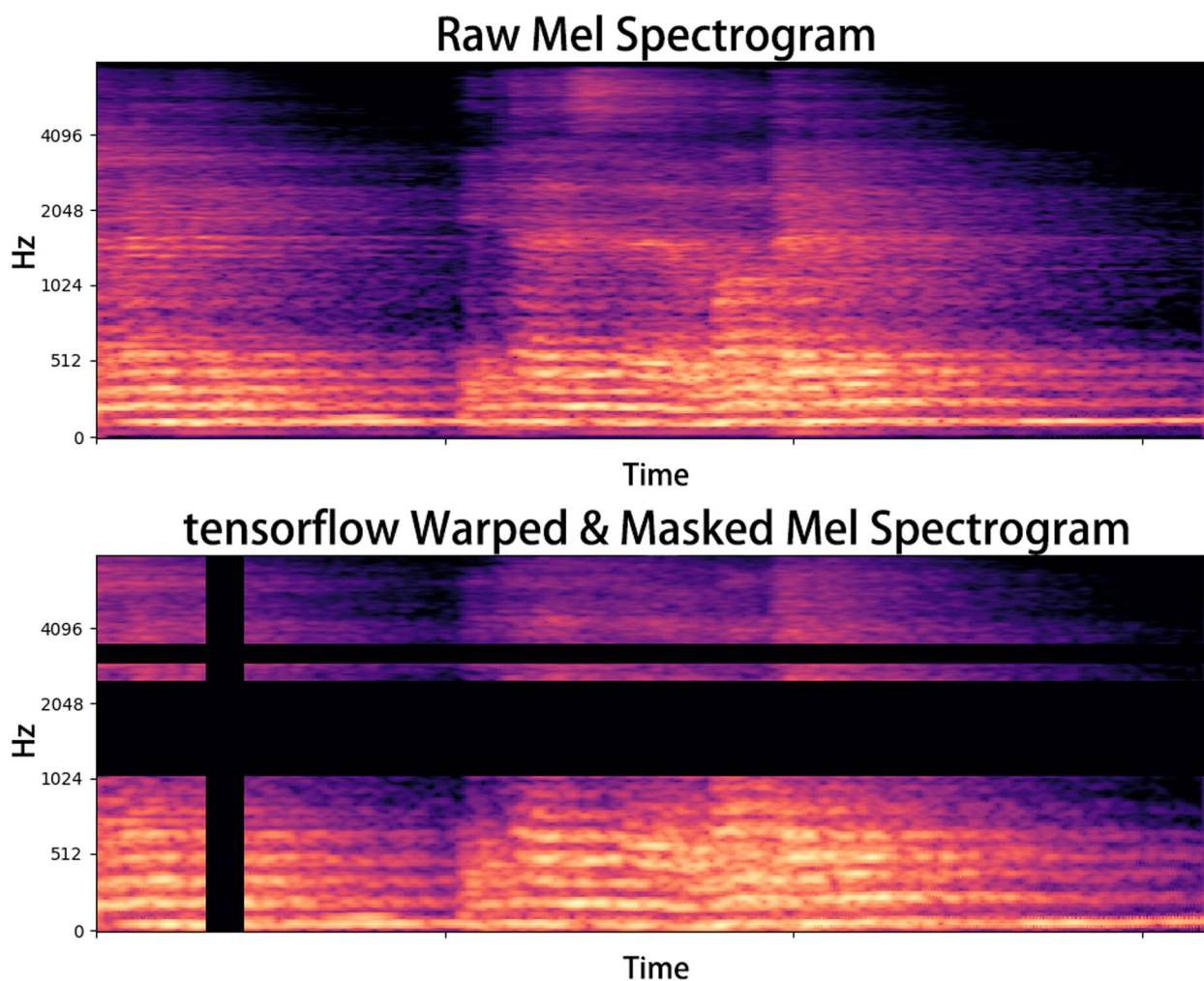
Time warping was implemented using TensorFlow's sparse image warping function. For a log Mel-frequency spectrogram with  $\tau$  time steps, we treated it as an image with the time axis horizontal and the frequency axis vertical. In the image, we randomly selected points within the interval  $[W_m, \tau - W_m]$

located between time steps and applied random warping. The warping distance,  $w$ , was chosen uniformly from the range  $[0, W_m]$ , where  $W_m$  is the time warp parameter. Six anchor points were fixed at the boundaries of the image, including the four corners and the midpoints of the vertical edges.

Frequency masking was applied as follows: a continuous set of Mel frequency channels  $[f_0, f_0 + f_1)$  is masked, where  $f_1$  is initially chosen from a uniform distribution  $[0, F_m]$ , and  $f_0$  is chosen from  $[0, \nu - f_1)$ , where  $F_m$  is the frequency mask parameter, and  $\nu$  is the number of Mel frequency channels.

Time masking was applied in a similar manner: a continuous set of time steps  $[t_0, t_0 + t_1)$  is masked, with  $t_1$  being initially chosen from a uniform distribution  $[0, T_m]$ , and  $t_0$  chosen from  $[0, \tau - t_1)$ , where  $T_m$  is the time mask parameter, and  $\tau$  is the total number of time steps. The Mel-frequency spectrogram with masking





**Fig. 3** Augmentation schemes applied to reverberation signals without noise

applied is then converted back to the time-domain signal.

Finally, 4800 speech sequences with these masking effects were added to the original training dataset for neural network training, and this dataset is labeled as *Dataset II*, as shown in Table 1. Constrained by computational resources, it is important to note that SpecAugment is not applied on-the-fly during each epoch. Instead, it undergoes offline processing on the data and is subsequently integrated directly into the training set. This approach aims to strike a balance between computational costs and the effectiveness of data augmentation. This comprehensive data augmentation strategy aims to help the neural network better adapt to various environments and conditions, ultimately improving its generalization performance.

### 3.3 Featurization

Audio feature extraction is crucial in convolutional neural networks, as it directly influences the model's performance. However, combining multiple feature extraction methods into one model led to complex models and requires a substantial amount of data and expensive training costs. Therefore, it is necessary to balance the addition of feature extraction methods while retaining key acoustic information to ensure that the model can handle a variety of datasets and provide general and accurate blind room parameter estimation performance.

Prior works in [15, 16, 39] emphasize the importance of low-frequency information for room acoustic parameter estimation. Consequently, feature representation is restricted to the relatively low-frequency range (< 2000 kHz). The Gammatone ERB filterbank

is used to generate time-frequency representations, comprising 20 frequency bands covering the frequency range from 50 to 2000 Hz. The audio is computed using a 64-sample Hann window with a 32-sample hop size, resulting in a  $20 \times 1997$  complex Gammatone spectrogram.

Furthermore, the phase information extracted from the audio is also retained following the work in [21]. Phase angles are computed for each time-frequency bin to generate phase features. These features are then truncated to include only the frequency bands associated with frequencies below 500 Hz (i.e.,  $5 \times 1997$ ) since lower-frequency components generally carry more information related to room volume. Additionally, the first-order derivatives of the phase coefficients along the frequency axis are calculated (i.e.,  $5 \times 1997$ ). This feature configuration aligns with the “+phase” model described in [21], which has been proven to outperform methods based solely on amplitude spectrogram features.

By combining spectral features, phase features, and first-order derivatives of phase coefficients, a two-dimensional feature block is obtained. The dimension of the feature block is  $30 \times 1997$ , where 30 represents the feature dimension ( $F$ ), and 1997 represents the time dimension ( $T$ ).

### 4 Model architecture

In this section, different architectures for audio data processing models are described for blind room parameter estimation tasks. These models include a CNN-based model, a CRNN-based model, and proposed attention-based systems.

Firstly, the CNN-based model utilizes multiple convolution and pooling layers to capture features of the audio data through convolution operations, followed by reducing the parameter count using pooling operations. Secondly, the CRNN-based model combines CNN and LSTM networks, designed to handle time series data better, capturing both time and frequency features of the audio data. Finally, the proposed model employs a completely different approach, relying solely on attention mechanisms without using convolution. This model has a unique structure, breaking

inputs into patches, processing data through embedding layers and positional encoding layers, and ultimately extracting features and producing results using Transformers. Additionally, the study also employs transfer learning by utilizing pretrained image models to process audio data, improving performance and efficiency.

#### 4.1 Convolutional neural network

In this section, a model based on a CNN following the “+phase” model in [21] is introduced, for processing two-dimensional feature blocks extracted from audio data. The model comprises six convolutional layers with corresponding average pooling layers, and each convolutional layer is followed by a rectified linear unit (ReLU) activation function. To prevent overfitting, dropout layers, which discard 50% of the connections, are introduced within the network structure. Taking the estimation of room volume parameters as an example, the final output layer is a fully connected layer, mapping the output dimension to downstream tasks. In particular, the structure of its last layer is dynamically adjusted according to the requirements of the blind room parameter estimation task to meet the performance needs of different tasks. Its system architecture is illustrated in Fig. 4.

#### 4.2 Convolutional recurrent neural network

CRNN is designed to capture both temporal and frequency features in audio data while also having a memory to handle time series data. CRNN efficiently extracts features from data and models sequences, better accommodating variable-length inputs, making it highly suitable for practical blind room parameter estimation problems.

A CRNN-based model is introduced in this section as it combines the parameteric efficiency of CNNs with the capability of sequential modeling from gated RNNs. The system architecture of CRNN is illustrated in Fig. 5.

Overall, the model consists of six convolutional layers with ReLU activation, followed by an LSTM layer, a max-pooling layer, a dropout layer, and a time-distributed fully connected layer.

Convolutional blocks gradually reduce the size and dimensions of the feature maps, allowing more sequences

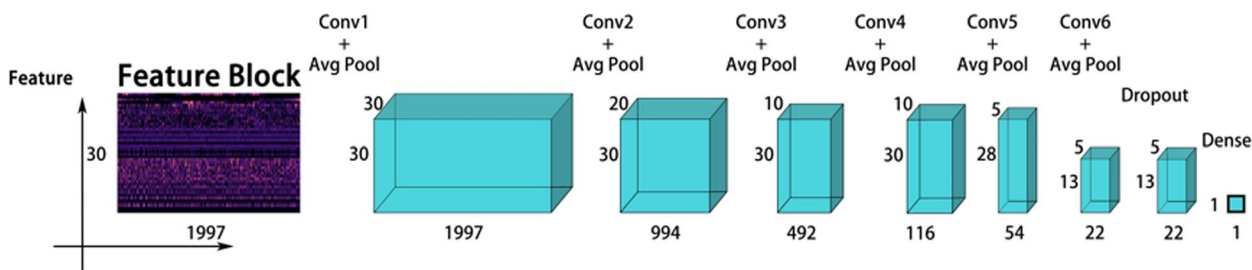
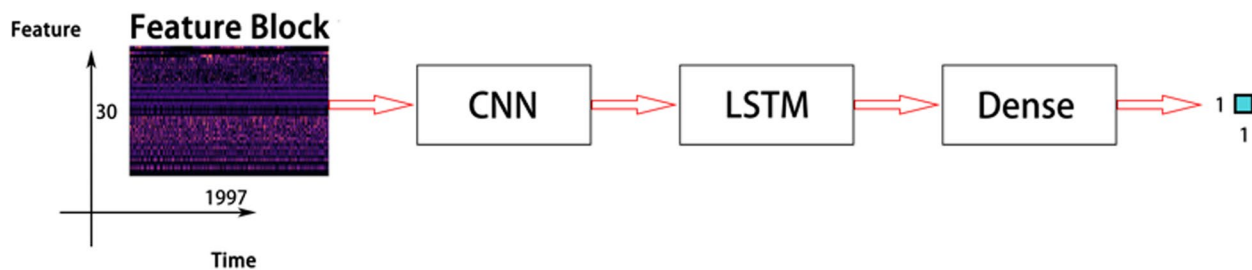


Fig. 4 The system architecture of the CNN-based model



**Fig. 5** The system architecture of the CRNN-based model

to enter the network. Simultaneously, the max-pooling layer is applied to extract useful features from the raw audio data.

The LSTM layer, which serves as a key component of the CRNN, is used for processing time series data, capturing the temporal relationships and sequence information in the input data. The hidden layer size of the LSTM is set to 64 and can be adjusted based on the size of the input fed into the LSTM. Prior to the dense layer, a max-pooling layer is employed to reduce the parameter count, with a pooling size set to 2, similar to [40].

Subsequently, the model includes a dropout layer with rate of 0.5 before the dense layer. The data is then flattened and passed to the fully connected layer, whose output size can be adjusted as needed. Considering that the estimated room parameters are positive values, an additional ReLU activation function is added to the final output layer.

Finally, the model outputs the estimated room parameters from the last time step. In this example, blind indoor volume estimation is used as the task, which is a regression task with an output size of 1. The structure can be adapted to specific application scenarios and datasets.

### 4.3 In-depth: convolution-free audio spectrogram transformer

#### 4.3.1 Audio spectrogram transformer

In this section, we introduce a model based purely on attention mechanisms without convolution for blind room parameter estimation. The design of this model is inspired by the workings of the Audio Spectrogram Transformer described in [28], which has shown remarkable performance in end-to-end audio classification tasks. However, it is noted that this purely attention-based approach has not been extensively explored in other domains, especially in the realm of blind room parameter estimation.

The primary goal of this section is to apply the proposed model that is purely attention-based to the blind room parameter estimation problem and compare its performance with traditional CNN and CRNN models.

In this work, two-dimensional feature block with dimensions of  $30 \times 1997$  as input for the proposed model is used. To better capture local information in the audio, the two-dimensional feature block is divided into  $P$  patches, each sized  $16 \times 16$ . The goal of patch split is to ensure a better capture of local features within the audio signal. Additionally, to maintain consistency in both feature and time dimensions, each patch has a 6-feature dimension and 6-time dimension overlap with its surrounding blocks. As a result, the number of patches  $P$  is determined to be 398, shown as:

$$P = \lceil \frac{F - 16}{10} \rceil \lceil \frac{T - 16}{10} \rceil, \quad (3)$$

where  $F$  represents the feature dimension, and  $T$  represents the time dimension.

To further process these patches, we introduced a linear projection layer. This layer's role is to flatten each  $16 \times 16$  patch into a one-dimensional patch embedding with a dimension of 768, referred to as the patch embedding layer. This embedding process helps reduce the data's dimensionality, making it more suitable for subsequent processing in the model.

Since these patches are not arranged in chronological order, and traditional Transformer architectures do not directly handle input sequences, we introduced trainable positional embeddings of dimension 768 in each patch. By introducing these trainable positional embeddings, the model is better able to understand the spatial structure of the audio spectrogram and grasp the positional information between patches.

Furthermore, the feature sequence is fed into the Transformer. Similar to [28], each feature sequence begins with a [CLS] token. In this model, the encoding and feature extraction of the input sequence only utilizes the encoder part of the original Transformer architecture [41]. The advantage of using the original Transformer structure is that it is a standard architecture already available in PyTorch and TensorFlow, making it easy to reproduce. Secondly, we plan to apply transfer learning to this task, and the standard



architecture facilitates transfer learning. Specifically, the embedding size of the Transformer encoder we use is 768, with 12 layers and 12 heads, which are the same as those in [42, 43].

We adjusted the output of the encoder based on the type of room parameters being estimated. Taking room volume estimation as an example, the input consists of a sequence formed by a feature block with the dimensions of  $30 \times 1997$ , and the output is a single label used for volume prediction. The entire output of the Transformer serves as the feature representation for the two-dimensional audio feature block, which is subsequently mapped to labels for volume estimation using a linear layer with a Sigmoid activation function. The system architecture of the proposed model is depicted in Fig. 6.

In summary, traditional convolutional neural networks typically have multiple layers, small kernels, and stride sizes, in contrast to the proposed model, which includes patch embeddings (viewed as a single convolutional layer with large kernels and strides) and projection layers within Transformer blocks (viewed as  $1 \times 1$  convolution). The proposed model used in this study can be referred to as a convolution-free model, distinguishing it from CNN and CRNN [42, 43].

### 4.3.2 ImageNet pretraining

Many researchers have noted that Transformers lack some of the inductive biases inherent to CNNs, such as translation equivariance and locality [44]. Consequently,

they exhibit poorer generalization capabilities under conditions of limited training data, compared to simpler models like CNNs and CRNNs [42].

To achieve accurate blind room parameter estimation, a substantial amount of publicly available data with correctly labeled room parameters is required to train the network. Therefore, two approaches are adopted:

- 1) As introduced in Sect. 2.2, the use of an image-source model to synthesize RIR datasets.
- 2) Transfer learning.

Transfer learning has been widely explored in previous research, particularly in transferring from visual tasks to audio tasks. This transfer often focuses on using CNN-based models [25, 45–47], where ImageNet-pretrained CNN weights are used as initial weights for audio classification tasks. However, the computational cost of training state-of-the-art visual models could be relatively high. Fortunately, some common architectures like ResNet [47] and EfficientNet [48] offer readily available ImageNet-pretrained models for TensorFlow and PyTorch, making transfer learning more convenient.

For image classification tasks, research indicates that Transformer models start to outperform traditional models in performance when the dataset size exceeds 14 million [42]. However, audio datasets for blind room parameter estimation tasks typically cannot provide such massive amounts of data, posing a challenge. Therefore,

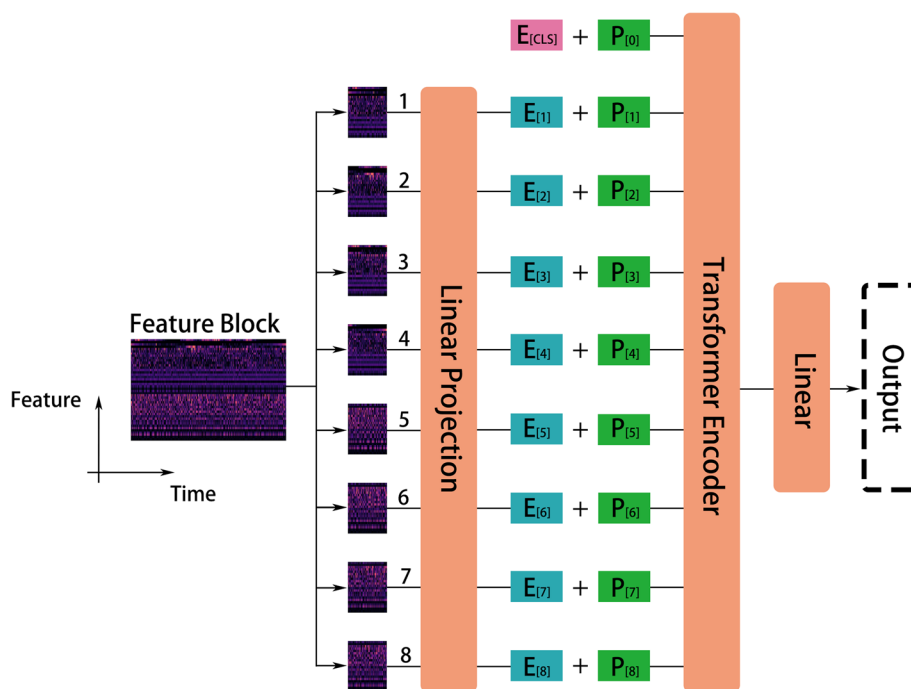


Fig. 6 The system architecture of the proposed model

we have decided to explore cross-modality transfer learning for the task of audio spectrogram processing, leveraging the similar format between image and audio data.

In this study, we use a pretrained off-the-shelf vision transformer (ViT) model from ImageNet [43] to simplify the transfer learning process. Afterward, we make appropriate adjustments to adapt it to the blind room parameter estimation task. Although both ViT and the proposed model employ the standard Transformer with the same patch and embedding sizes, their architectural similarities require adjustments to the structure before migration to ensure compatibility with the blind room parameter estimation task. Three adjustments are implemented:

- 1) One challenge is that the proposed model's input is a single-channel feature block, whereas ViT's input is a three-channel image. To overcome this issue, we adopted an approach to calculate the average weights corresponding to each of the three input channels of the ViT patch embedding layer. This averaging method helps integrate the information from the three channels into a single channel. We then used these averaged weights for the patch embedding layer. This essentially extends the single-channel spectrogram to a three-channel image with the same content but higher computational efficiency. This approach helps us better adapt to the differences in model input, thus improving the efficiency and performance of the research.
- 2) Another issue encountered in this study is that the input dimensions of ViT are fixed, whereas in practical tasks, the model needs to adapt to variable-length audio inputs. As the audio length changes, the dimensions of the feature block also change. While the Transformer naturally supports variable input lengths and can be directly transferred from ViT to the proposed model, special handling of positional encodings was required. This is because the ViT model learns to encode spatial information during ImageNet training. Therefore, we used the "Cut and bi-linear interpolate" method [25] to adjust the input size and manage positional encodings. This way, even with different input shapes, we can pass on the two-dimensional spatial knowledge obtained from the pretrained ViT to the proposed method, allowing the model to adapt to audio inputs of varying lengths. This method helps us better handle data under different input conditions.
- 3) To adapt to different sub-tasks in blind room parameter estimation, we take the example of blind room volume estimation. We reinitialize the final classification layer of ViT to output corresponding volume labels in the proposed model.

These adjustments are crucial to ensuring that the pretrained ViT model can be effectively used for the specific task of blind room parameter estimation and achieve improved efficiency and performance.

## 5 Experiment

In this section, an in-depth exploration of blind room parameter estimation tasks is conducted, utilizing two different datasets (*Dataset I* and *Dataset II*) to assess the performance of various models. We employed log-scaled and normalized data to better handle the magnitude differences between room parameters and utilized multiple evaluation metrics to comprehensively assess the accuracy and robustness of the models.

We employed various model architectures (CNN-based model, CRNN-based model, and the purely attention-based method in Sect. 4.3) and conducted a detailed comparison of their performance on different tasks and datasets.

Overall, through these experiments, we examined the performance of different models in various blind room parameter estimation tasks and assessed their adaptability in handling variable-length audio inputs.

### 5.1 Datasets

In the task of blind room parameter estimation, *Dataset I* and *Dataset II* mentioned in Sect. 3.1 and 3.2 were utilized. In the preprocessing phase, room volume labels (in  $\text{m}^3$  units, in logarithmic scale) were exclusively read, and four models, CNN-based model, CRNN-based model, the proposed method, and the "proposed method w/ pre-train" model, were individually evaluated for their performance on *Dataset I* and *Dataset II*. For the blind room parameter estimation with variable-length audio input, *Dataset II* was employed. Similarly, in the preprocessing phase, only room volume labels (in  $\text{m}^3$  units, in log-scaled) were considered. However, a modification was made to the test set of *Dataset II*. Specifically, samples were extracted from 1 to 4 s with a step size of 0.5 s, and zero padding was applied to different lengths of audio samples to match the original length. This was done to assess the performance of different models in handling blind room parameter estimation under audio inputs with different length.

Finally, in the task of joint estimation of room parameters, *Dataset II* was used. In the preprocessing phase, the model simultaneously reads room  $\text{RT}_{60}$  (in seconds) labels and room volume labels (in  $\text{m}^3$  units). In order to overcome the significant scale differences between these two parameters, we adopted an approach where we mapped the values of  $\text{RT}_{60}$  to volume values and applied a logarithmic scaling to them. It is worth emphasizing that this data processing method is reversible, allowing us to revert all parameters to

standard units at any time. The advantage of mapping the parameter relationship rather than standard normalization is that it eliminates the need for frequent adjustment of hyperparameters when dealing with different blind room parameter estimation tasks, as it effectively addresses the differences in units and magnitudes among the parameters. This is done to evaluate the performance of different models in joint room parameters estimation.

## 5.2 Evaluation metrics and loss function

As shown in Fig. 2, due to large span of room volume and RT<sub>60</sub> ranges, the estimation error could be related to its order of magnitude. Therefore, a log-10 estimation is more suitable than a linear estimation. This way, larger acoustic spaces in training are not disproportionately affected due to the relatively high contribution of error estimation. Using a logarithmic estimation better handles estimation errors of different orders of magnitude.

Four evaluation metrics using a base-10 logarithm are considered. They are as follows: (1) mean squared error (MSE): MSE is the average of the squared differences between estimated room parameters and ground truth room parameters. It is used to measure the degree of dispersion between estimated values and ground truth values. The smaller the average of squared differences, the closer the estimated values are to the ground truth values. (2) Mean absolute error (MAE): MAE represents the average of the absolute differences between estimated values and ground truth values. It provides the average deviation between estimated values and ground truth values and is commonly used to measure the accuracy of estimated values. (3) Pearson correlation coefficient ( $\rho$ ): the Pearson correlation coefficient is used to measure the strength and direction of the linear relationship between two variables. It is used to describe the relationship between estimated room parameters and ground truth room parameters, with values ranging from  $-1$  to  $1$ . Negative values indicate a negative correlation, positive values indicate a positive correlation, and  $0$  indicates no correlation. (4) MeanMult MM: MM is the mean absolute logarithm of the ratio between the estimated room volume and the ground truth room volume. This metric provides an overview of the mean error in the ratio between estimated room parameters and ground truth room parameters. Taking the logarithm of the ratio helps reduce the impact of data points with significant differences. For example, for the estimated volume parameter, given the estimated volume  $\hat{V}_n$  and the ground truth volume  $V_n$ :

$$MM = e^{\frac{1}{N} \sum_{n=1}^N \left| \ln \left( \frac{\hat{V}_n}{V_n} \right) \right|}, \quad (4)$$

where “ $n$ ” represents the sample index, and “ $N$ ” represents the total number of samples.

During model training stages, MSE was used as the loss function to minimize the error between estimated room parameters and ground truth room parameters. In the “Estimation of room parameter” and “Room parameter estimation under variable-length audio input” tasks, the loss function  $L_1$  formula was as follows:

$$L_1 = \frac{1}{B} \sum_{n=1}^B (\hat{V}_n - V_n)^2, \quad (5)$$

where “ $n$ ” represents the sample index, and “ $B$ ” represents the batch size during training.

In contrast, for the task of “Joint estimation of room parameters,” which involves the simultaneous estimation of RT<sub>60</sub> and volume parameters. To avoid differences in units and orders of magnitude between different parameters, as well as the impact of parameter scaling methods, the normalized MSE was used instead of the MSE in Eq. 5. The loss function  $L_2$  was formulated as follows:

$$L_2 = \lambda_1 * \frac{\sum_{n=1}^B (\hat{U}_n - U_n)^2}{B \sum_{n=1}^B (U_n)^2}, + \lambda_2 * \frac{\sum_{n=1}^B (\hat{V}_n - V_n)^2}{B \sum_{n=1}^B (V_n)^2}, \quad (6)$$

where  $\hat{U}_n$  and  $U_n$  represent the estimated and the ground truth RT<sub>60</sub>, respectively.  $\lambda_1$  and  $\lambda_2$  are weights used to control the balance between the RT<sub>60</sub> and volume normalized MSE loss functions. These weights are employed to adjust the relative importance of these two functions during model training. Based on empirical evidences and experimental results,  $\lambda_1$  is set to  $1$ , and  $\lambda_2$  is set to  $2$ . This weight configuration can be adjusted according to the specific task and model performance to better meet the training requirements.

## 5.3 Experiment configurations

Different MSE loss functions were chosen based on the task’s requirements. Each model utilized the Adam optimizer from PyTorch. During the training process, L2 regularization was applied to prevent overfitting. Simultaneously, an adaptive learning rate strategy was employed to ensure the convergence of the model. If the model’s validation set did not improve for ten consecutive epochs, an early stopping criterion was triggered, leading to the cessation of the training process to prevent further overfitting. Furthermore, to select the optimal-performing model, we monitored the MSE values on the validation set during grid search and optimized hyperparameters, including initial learning rate as well as batch size. The hyperparameter configuration that demonstrated the best performance was chosen as the final model parameters.

For the “Estimation of room parameter” task, to facilitate comparative testing, we switched between *Dataset*

*I* and *Dataset II* as well as determined whether to use a pretrained model from ImageNet. To ensure consistency in model configurations, hyperparameters were kept constant. CNN-based and CRNN-based models were trained for 1000 epochs with an initial learning rate of  $5e-4$ , a batch size of 128.

The proposed attention-based method and the “proposed method w/ pretrain” model were trained for 150 epochs with an initial learning rate of  $5e-6$ , a batch size of 16. For the “Joint estimation of room parameters” task, CNN-based and CRNN-based models were trained for 2000 epochs with an initial learning rate of  $2e-4$ , a batch size of 128. The proposed method and the “proposed method w/ pretrain” model were trained for 300 epochs with an initial learning rate of  $2e-6$ , a batch size of 16.

To ensure fairness, all models were trained on devices equipped with an Intel Core i9 processor and an NVIDIA GeForce 4090 GPU.

## 6 Results and discussion

### 6.1 Estimation of room volume parameter

To investigate whether comparable performance similar to that of CNN and CRNN can be achieved by using a pure attention mechanism, we extracted audio data from *Dataset I* for the purpose of estimating room volume parameter. We transformed audio data into a feature block, as described in Sect. 3.3. Subsequently, we separately input these feature blocks into the CNN-based model, CRNN-based model, and the proposed method (the base version without ImageNet pretraining) for training. We then compared the predicted volume labels to the ground truth values. The results of these three models are presented in Table 2, with the evaluation metric of our proposed method (the base version without ImageNet pretraining) being highlighted in bold. Note that in this section we mainly focus on estimation of room volumes as this task has been shown to be more challenging than  $RT_{60}$  estimation in the literature [15, 16, 21].

In addition, the table includes information on the model’s memory consumption and computational complexity, such as the number of parameters (#Param) and multiply-accumulate operations (MACs). To ensure fairness, we conduct tests using the PyTorch profiler [49] in the same GPU environment. A comprehensive comparison of the data in the

table reveals that CNN models have fewer parameters and relatively low memory consumption but perform worse in various evaluation metrics. While the CRNN model shows an increase in parameter count compared to CNN and some improvement in evaluation metrics, it still falls short of our proposed method. In contrast, although our method has relatively higher parameter count and memory consumption, it demonstrates significant advantages in all evaluation metrics. Specifically, our method outperforms in terms of MSE, MAE,  $\rho$ , *MM*, and MACs. Despite the relatively higher memory consumption, considering the performance improvement, this increase can be deemed acceptable.

Based on experimental results above, we can see that the proposed method using the pure attention mechanism significantly outperforms both CNN and CRNN-based approaches, even with a lower-layer network configuration and relatively fewer training epochs. This suggests that the proposed method can accurately capture the acoustic characteristics in the audio data, thereby improving the accuracy and stability of room volume estimation.

Meanwhile, the four evaluation metrics show that the CRNN-based model performs better than the CNN-based model. This can be attributed to the advantages of CRNN, which combines CNN with LSTM. CRNN can better handle the time series audio data while capturing local features, which is crucial for blind room parameter estimation tasks.

To further investigate the impact of ImageNet pretraining on the proposed method’s performance, the “proposed method w/ pretrain” model was trained on *Dataset I*. Simultaneously, to examine the effect of the SpecAugment data augmentation method on the performance of existing models, we retrained the existing four models on *Dataset II*. The results of the above experiments are shown in Table 3, where the performance of the “proposed method w/ pretrain” model on *Dataset II* is highlighted in bold

Based on the training results of different models on *Dataset I*, we can observe a significant improvement in the performance of the proposed method in the “Estimation of room parameter” task with the use of the ImageNet pretraining method. Furthermore, when the four models were retrained on *Dataset II*, the application of the SpecAugment method elevated the models’ performance to a new level. In particular, this method

**Table 2** The comparison between the CNN-based model [21], the CRNN-based model, and the base version of the proposed method

Method	# Params (M)	Evaluation metrics				Memory consumption (GB)	MACs (G)
		MSE	MAE	$\rho$	<i>MM</i>		
CNN [21]	0.013	0.3863	0.4837	0.6984	3.0532	1.81	0.237
CRNN	0.494	0.3572	0.4265	0.7262	2.6701	1.95	0.236
<b>Proposed method</b>	85.256	<b>0.2650</b>	<b>0.3432</b>	<b>0.8077</b>	<b>2.2039</b>	4.55	34.083



**Table 3** Performance comparison of different models with and without the application of SpecAugment

Method	Dataset I				Dataset II			
	MSE	MAE	$\rho$	MM	MSE	MAE	$\rho$	MM
CNN [21]	0.3863	0.4837	0.6984	3.0532	0.3154	0.4136	0.7678	2.5921
CRNN	0.3572	0.4265	0.7262	2.6701	0.2818	0.3684	0.7898	2.3471
Proposed method	0.2650	0.3432	0.8077	2.2039	0.1981	0.2884	0.8580	1.9427
<b>Proposed method w/ pretrain</b>	0.2157	0.3111	0.8529	2.0470	<b>0.1541</b>	<b>0.2423</b>	<b>0.8929</b>	<b>1.7470</b>

**Table 4** Comparison of median and mean absolute error for volume parameters among the best-performing models

Method	Median ( $m^3$ )	Mean absolute error ( $m^3$ )
CNN [21]	353	1919
CRNN	257	1644
<b>Proposed method w/ pretrain</b>	<b>155</b>	<b>1219</b>

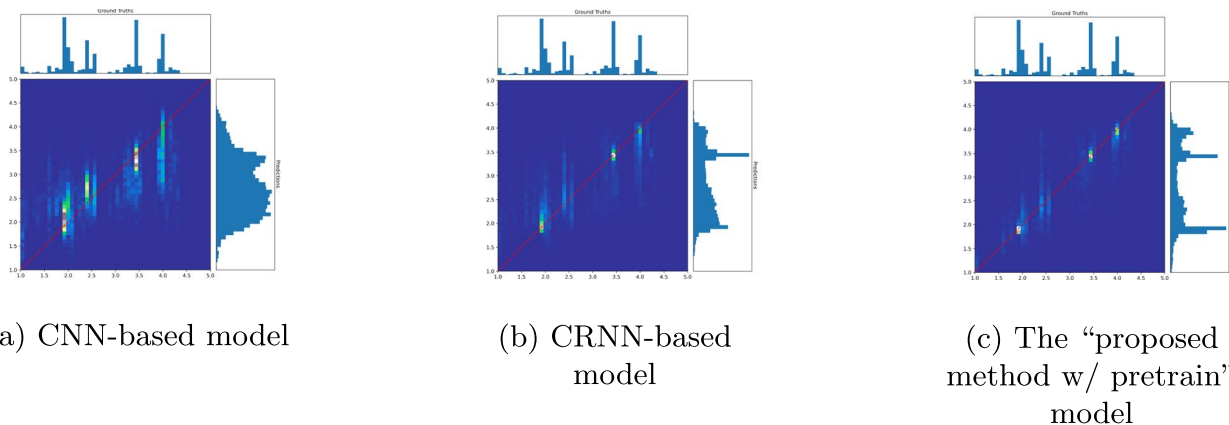
demonstrates a significant improvement in the performance of the “proposed method w/ pretrain” model. It confirms the effectiveness of SpecAugment in mitigating overfitting and enhancing model generalizability.

Meanwhile, in order to provide a more illustrative example, we rescaled the experimental results to a linear scale. In this experiment, the test set room volume ranges from 12 to 21,000  $m^3$ . We compared the performance of the best-performing models, namely the CNN-based model, CRNN-based model, and the “proposed method w/ pretrain” model. They were trained on *Dataset II*, and their model’s median as well as mean absolute error are shown in Table 4. In particular, the median and mean absolute error of the “proposed method w/ pretrain” model are highlighted in bold in the table.

From the table, the “proposed method w/ pretrain” model exhibits the best performance in terms of both median and mean absolute error, having the lowest error values.

The Fig. 7 displays the confusion matrices for these three best-performing models in the “Estimation of room volume parameter” task, with the  $x$ -axis and  $y$ -axis representing the log-10 exponent of volume size. From the visualization, it is evident that the “proposed method w/ pretrain” model exhibits excellent performance across the entire test range. Its distribution consistently closely surrounds the ground truth, clearly outperforming the CNN-based and CRNN-based models.

Results in this section indicate that the proposed purely attention-based model is capable of capturing relevant features and representations in the context of room volume regression efficiency. More importantly, it demonstrates remarkable generalization capabilities, effectively applying the patterns learned from the training data to real-world rooms, even for rooms the model has not encountered before, resulting in accurate volume estimates. This outcome provides a strong theoretical foundation for our approach and underscores its potential in more blind estimation practical problems, which will be addressed in the following section.

**Fig. 7** Confusion matrices for the best-performing models trained on *Dataset II* in the “Estimation of room volume parameter” task

### 6.2 Room parameter estimation under variable-length audio input

In this section, model performances under variable-length audio inputs are evaluated for the “Room parameter estimation” task. The selected models were tested with different lengths of audio inputs, and their performances were assessed using four objective evaluation metrics as shown in Fig. 8. It is evident from the figure that the accuracy of the models in predicting room volume parameter significantly depends on the length of the input audio. As the input audio length shortens, the estimation performance of all models inevitably experiences degradation.

By observing the curves of MSE and MAE metrics in Fig. 8, it can be noted that the CRNN-based model, the proposed model, and the “proposed method w/ pretrain” model exhibit smaller decay slopes. This suggests that, compared to CNN-based models, they can better handle time sequences of variable length. The smaller decay slopes of the evaluation metrics can be considered an

indication that the models better maintain performance stability, even when the input length decreases, maintaining relatively good performance.

Results in Fig. 8 also indicate that the “proposed method w/ pretrain” model performs the best at the same input length. For the shortest input sample, i.e., when the input length is 1 s, the MSE for the “proposed method w/ pretrain” model is 0.6458. In comparison, to achieve the same performance level, the CNN-based model, CRNN-based model, and the proposed model would require input audio lengths of approximately 2.8 s, 2.0 s, and 1.2 s, respectively. This advantage facilitates the proposed attention-based models to outperform both CNN and CRNN systems with significantly less temporal context, which can be a valuable merit when dealing with speech-based blind estimation problems in practice.

### 6.3 Joint estimation of room parameters

This section aims to address the “Joint estimation of room parameters” task, which involves training a single

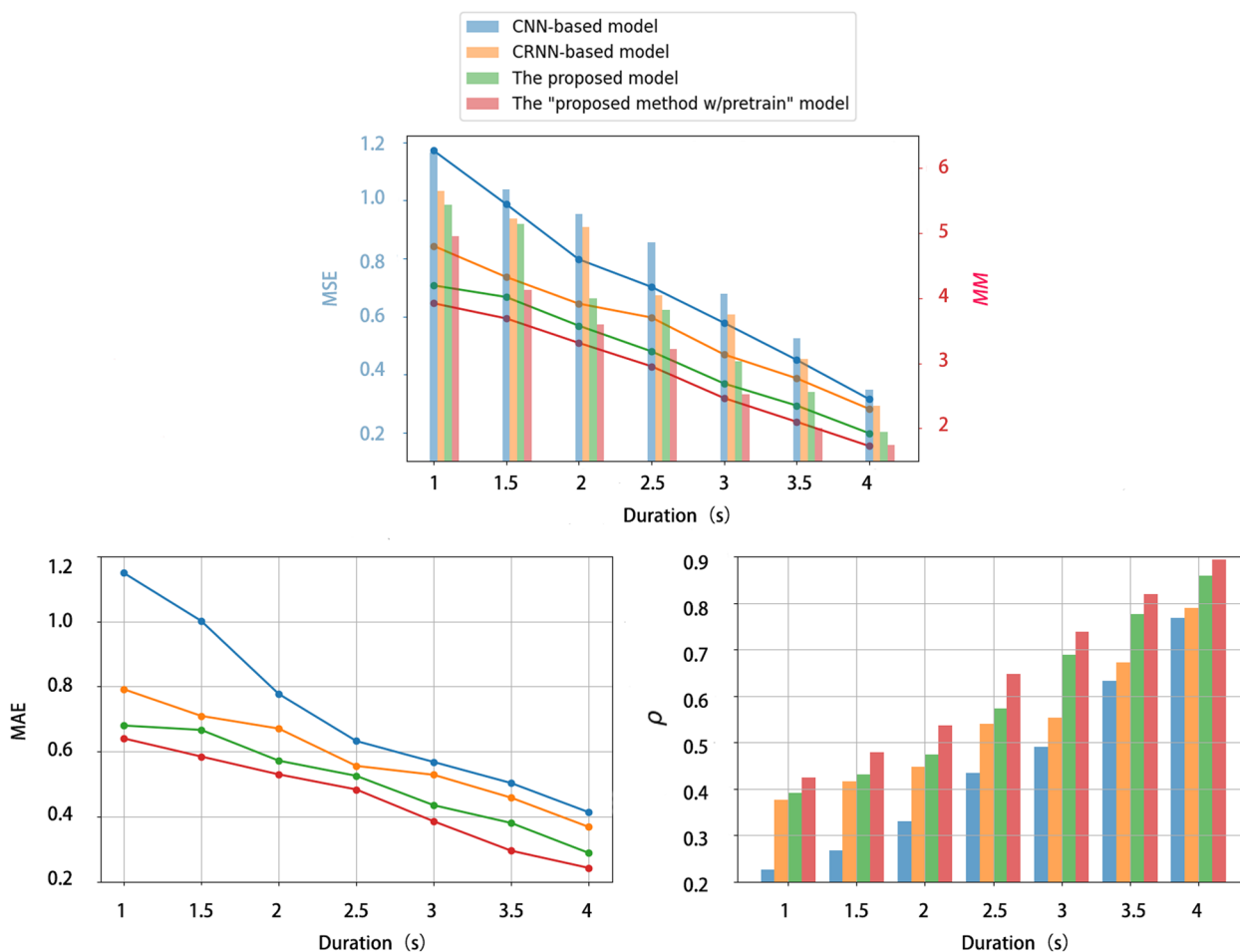


Fig. 8 Performance comparison of different models under the “Room parameter estimation under variable-length audio input” task

model to simultaneously estimate multiple room parameters. Specifically, due to the shared acoustic characteristics of room volume and  $RT_{60}$  parameters, it is possible to estimate them concurrently by extracting reverberation-related information. Considering difficulties in collecting groundtruth of real data for other room parameters, such as total surface area and average surface absorption coefficient in real-world room datasets, this experiment focuses on the joint estimation of room volume and  $RT_{60}$ .

In this task, we selected three models, namely the CNN-based model, CRNN-based model, as well as the “proposed method w/ pretrain” model, and trained them on *Dataset II*. Their network architectures were fundamentally similar to those used for the “Estimation of room volume parameter” task, with minor modifications. In the “Joint estimation of room parameters” task, the three models are required to output two parameters, i.e., room volume and  $RT_{60}$ , instead of a single parameter. Consequently, the final output layers of the models were modified to include two fully connected layers for estimating different room parameters. During the training process, hyperparameters were fine-tuned (as described in Sect. 5.3), and the loss function was adjusted (as shown in Eq. 6).

It is worth noting that, in order to mitigate issues related to different units and scales among parameters, as well as the impact of parameter scaling during normalization, we chose to use only the  $\rho$  as the evaluation metric. This helps ensure consistency among the estimated parameters. The corresponding results are presented in the Table 5. In which, the evaluation metric values of the “proposed method w/ pretrain” model are highlighted in bold.

From these results, it can be clearly seen that the “proposed method w/ pretrain” model outperforms the other models, achieving the highest  $\rho$  for both room volume

and  $RT_{60}$ , indicating its effectiveness in jointly estimating these room parameters.

In this experiment, the test set room volume ranges from 12 to 21,000  $m^3$ , while the RT range from 0.41 to 19.68 s. We rescaled the experimental results to a linear scale. The median, as well as mean absolute error for the three models regarding volume and  $RT_{60}$ , is displayed in Table 6. Among them, the performance metrics of the “proposed method w/ pretrain” model are highlighted in bold in the table.

Furthermore, we conducted a comparative study between the volume estimation in the joint model and the estimation of volume results for the “Estimation of room volume parameter” task by comparing results in Table 3 and Table 5, as well as Table 4 and Table 6. Despite the fact that the joint estimation models aim to simultaneously handle multiple parameters, it is clear that their volume estimation results, while experiencing some degree of attenuation, are overall very similar to the results obtained from estimating only a single parameter. This suggests that the performance of the joint model is in par with that of models designed for estimating a single parameter.

Figure 9 shows confusion matrices for volume and  $RT_{60}$  parameter estimation in the “Joint estimation of room parameters” task, highlighting the best-performing models. The  $x$ -axis and  $y$ -axis represent the log-10 exponent of room parameters (volume and  $RT_{60}$ ). From the visualization results, it can be observed that estimation performances of the CNN-based model, CRNN-based model, and the “proposed method w/ pretrain” model gradually improves, and their fitting capabilities increase progressively.

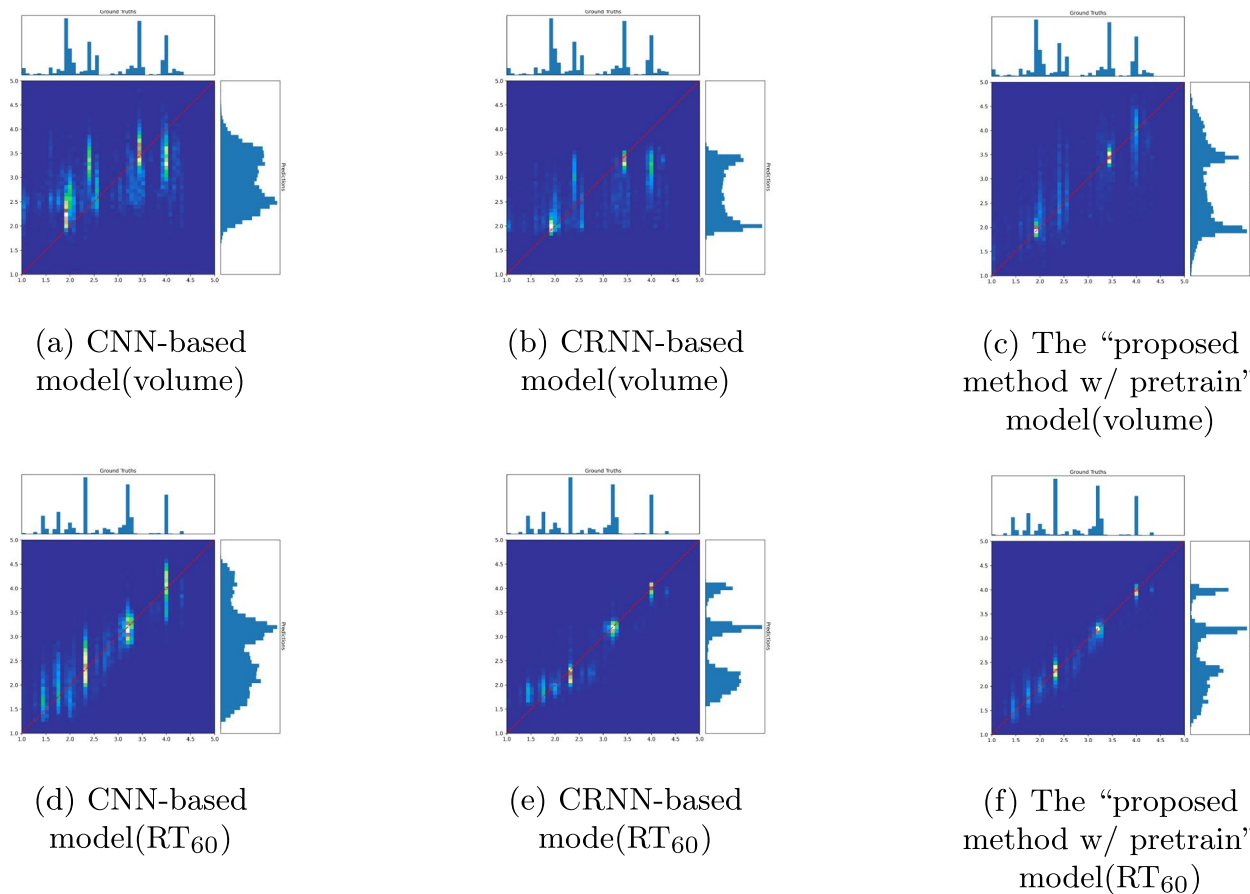
The comprehensive analysis of experimental results in this study demonstrates the effectiveness of the

**Table 5** Pearson correlation coefficients of best-performing models in “Joint estimation of room parameters” task

Method	CNN [21]		CRNN		Proposed method w/ pretrain	
	vol	$RT_{60}$	vol	$RT_{60}$	vol	$RT_{60}$
$\rho$	0.6187	0.9133	0.6584	0.9488	<b>0.8287</b>	<b>0.9681</b>

**Table 6** Comparison of median and mean absolute error for volume as well as  $RT_{60}$  parameters among the best-performing models

Method	Median		Mean absolute error	
	vol ( $m^3$ )	$RT_{60}$ (seconds)	vol ( $m^3$ )	$RT_{60}$ (seconds)
CNN [21]	728	0.64	2481	1.32
CRNN	329	0.39	2265	0.71
<b>Proposed method w/ pretrain</b>	<b>294</b>	<b>0.31</b>	<b>2208</b>	<b>0.61</b>



**Fig. 9** Confusion matrices for the best-performing models trained on *Dataset II* in the “Joint estimation of room parameters” task

joint estimation model for the blind room parameter estimation task. This method involves utilizing a single model to simultaneously estimate both room volume and RT<sub>60</sub> parameters, providing a more holistic understanding of acoustic environmental characteristics. Particularly, the “proposed method w/ pretrain” model achieves the highest  $\rho$  for both room volume and RT<sub>60</sub> parameters. This highlights the model’s capability of capturing the intricate characteristics of acoustic environments through the joint estimation of room parameters.

## 7 Conclusion and future work

In this study, we aim to explore the feasibility of using attention-based models to address audio processing tasks, specifically including the “Estimation of room volume parameter,” “Room parameter estimation under variable-length audio input,” and “Joint estimation of room parameters” tasks. We employ different training strategies to evaluate performances of a CNN-based model, a CRNN-based model, the proposed attention-based model, and the “proposed method w/ pretrain” model.

Experimental results based on unseen real-world rooms and realistic noise scenario indicate that our proposed method shows significant superiority in terms of accurately capturing the acoustic characteristics of audio data. This demonstrates that neural networks based on pure attention mechanisms can effectively handle regression problems related to audio and exhibit potential advantages in handling joint estimation tasks and variable-length inputs.

Future research directions will focus on optimizing and enhancing the performance of attention-based audio processing models in real-world applications. We plan to further improve the model structure, including considering more efficient variants, to better capture the complex features of audio data. Additionally, we will strive to collect more comprehensive and diverse room data to enhance the model’s generalization capabilities. We also aim to update robust and state-of-the-art RT<sub>60</sub> estimators [50–52] to obtain more accurate ground truth. These efforts will contribute to advancing the application of attention-based audio processing models in real-world scenarios.



### Acknowledgements

This work was supported by the Beijing Natural Science Foundation (No. L223033) and the National Natural Science Foundation of China under Grants (No. 61971015).

### Authors' contributions

Wang C. performed the whole research and wrote the paper. Jia M. and Jin W. provided support to the writing and experiments. All authors read and approved the final version of the paper.

### Funding

This work was supported by the Beijing Natural Science Foundation (No. L223033) and the National Natural Science Foundation of China under Grants (No. 61971015).

### Availability of data and materials

Not applicable.

### Declarations

#### Competing interests

The authors declare that they have no competing interests.

Received: 13 January 2024 Accepted: 10 April 2024

Published online: 24 April 2024

### References

- Z. Zhang, J. Geiger, J. Pohjalainen, A.E.D. Mousa, W. Jin, B. Schuller, Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Trans. Intell. Syst. Technol.* **9**(5) (2018)
- B. Wu, K. Li, F. Ge, Z. Huang, M. Yang, S.M. Siniscalchi, C.H. Lee, An end-to-end deep learning approach to simultaneous speech dereverberation and acoustic modeling for robust speech recognition. *IEEE J. Sel. Top. Signal Process.* **11**(8), 1289–1300 (2017). <https://doi.org/10.1109/JSTSP.2017.2756439>
- N. Mohammadiha, S. Doclo, Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(2), 276–289 (2016)
- Stefania Cecchi, Alberto Carini, and Sascha Spors. Room Response Equalization. *ATA Review*. Applied Sciences. **8**, 1 (2018). <https://doi.org/10.3390/app8010016>
- W. Jin, W.B. Kleijn, Theory and design of multizone soundfield reproduction using sparse methods. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(12), 2343–2355 (2015)
- W. Jin, "Adaptive reverberation cancelation for multizone soundfield reproduction using sparse methods," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, pp. 509–513 (2016) <https://doi.org/10.1109/ICASSP.2016.7471727>
- A. Neidhardt, C. Schneiderwind, F. Klein, Perceptual matching of room acoustics for auditory augmented reality in small rooms - Literature review and theoretical framework. *Trends Hear.* **26**, 23312165221092920 (2022)
- J.M. Jot, K.S. Lee, in *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*. Augmented reality headphone environment rendering (Audio Engineering Society). Los Angeles, (2016)
- Kuttruff, H. *Room Acoustics*. (CRC Press, Boca Raton, 2016)
- J. Eaton, N.D. Gaubitch, A.H. Moore, P.A. Naylor, Estimation of room acoustic parameters: The ACE challenge. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(10), 1681–1693 (2016). <https://doi.org/10.1109/TASLP.2016.2577502>
- T.M. Prego, A.A. Lima, R. Zambrano-López, S.L. Netto, in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition (2015), pp. 1–5. <https://doi.org/10.1109/WASPAA.2015.7336954>
- H.W.L. ollmann, A. Brendel, P. Vary, and W. Kellermann, "Single-channel maximum-likelihood T60 estimation exploiting subband information," in *Proc. ACE Challenge Workshop, Satellite IEEE*. New Paltz, (2015)
- Alastair H. Moore, Mike Brookes and Patrick A. Naylor, "Room identification using roomprints", *Audio Engineering Society Conference: 54th International Conference: Audio Forensics*, (2014)
- Nils Peters, Howard Lei, Gerald Friedland, Name that room: room identification using acoustic features in a recording. In *Proceedings of the 20th ACM international conference on Multimedia (MM '12)*. Association for Computing Machinery, New York, 841–844 (2012) <https://doi.org/10.1145/2393347.2396326>
- H. Gamper and I. J. Tashev, "Blind Reverberation Time Estimation Using a Convolutional Neural Network," *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Tokyo, pp. 136–140 (2018). <https://doi.org/10.1109/IWAENC.2018.8521241>
- A. F. Genovese, H. Gamper, V. Pulkki, N. Raghuvanshi and I. J. Tashev, "Blind Room Volume Estimation from Single-channel Noisy Speech," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, pp. 231–235 (2019) <https://doi.org/10.1109/ICASSP.2019.8682951>
- N. J. Bryan, "Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, pp. 1–5 (2020) <https://doi.org/10.1109/ICASSP40776.2020.9052970>
- P. Götz, C. Tuna, A. Walther and E. A. P. Habets, "Blind Reverberation Time Estimation in Dynamic Acoustic Conditions," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, pp. 581–585 (2022) <https://doi.org/10.1109/ICASSP43922.2022.9746457>
- S. Saini and J. Peissig, "Blind Room Acoustic Parameters Estimation Using Mobile Audio Transformer," *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, pp. 1–5 (2023) <https://doi.org/10.1109/WASPAA58266.2023.10248186>
- P. Srivastava, A. Deleforge and E. Vincent, "Blind Room Parameter Estimation Using Multiple Multichannel Speech Recordings," *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, pp. 226–230 (2021) <https://doi.org/10.1109/WASPAA52581.2021.9632778>
- I. Christopher, A. Mehrabi, W. Jin, in *Proc. ICASSP, Blind acoustic room parameter estimation using phase features (IEEE, Rhodes Island, 2023)*, pp. 1–5
- P. Callens, M. Cernak, Joint blind room acoustic characterization from speech and music signals using convolutional recurrent neural networks. (2020). arXiv preprint arXiv:2010.11167
- S. Deng, W. Mack, E.A.P. Habets: Online blind reverberation time estimation using CRNNs, in *INTERSPEECH*. (Incheon, 2020), pp. 5061–5065
- Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M.D. Plumley, Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 2880–2894 (2020)
- Y. Gong, Y.A. Chung, J. Glass, Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **29**, 3292–3306 (2021)
- P. Li, Y. Song, I.V. McLoughlin, W. Guo, L.-R. Dai, An Attention Pooling based Representation Learning Method for Speech Emotion Recognition, in *Proceedings of the Interspeech 2018*. (International Speech Communication Association, Hyderabad, 2018)
- Rybakov, Oleg, Natasha Kononenko, Niranjana Subrahmanya, Mirkó Vison-tai, and Stella Laurenzo. "Streaming keyword spotting on mobile devices." *arXiv preprint arXiv:2005.06720* (2020)
- Y. Gong, Y.A. Chung, J. Glass, in *Proc. Interspeech 2021*, Brno, Czech Republic. AST: Audio Spectrogram Transformer. pp. 571–575 (2021)
- C. Wang, M. Jia, M. Li, C. Bao and W. Jin, "Attention Is All You Need For Blind Room Volume Estimation," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, pp. 1341–1345 (2024) <https://doi.org/10.1109/ICASSP48485.2024.10447723>
- M. Jeub, M. Schafer and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *2009 16th International Conference on Digital Signal Processing*, Santorini, pp. 1–5 (2009) <https://doi.org/10.1109/ICDSP.2009.5201259>
- M. Jeub, M. Schafer and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *2009 16th International Conference on Digital Signal Processing*, Santorini, pp. 1–5 (2009) <https://doi.org/10.1109/ICDSP.2009.5201259>
- R. Stewart, M. Sandler, Database of omnidirectional and b-format room impulse responses. *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)* **13**(4), 165–168 (2010)

33. D.D. Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, S. Gannot, dechorate: A calibrated room impulse response dataset for echo-aware signal processing. *EURASIP J. Audio Speech Music Process.* Springer **2021**(1), 1–15 (2021)
34. D.T. Murphy, S. Shelley, Openair: An interactive auralization web resource and database. In *Audio Engineering Society Convention 129*. Audio Engineering Society, (2010)
35. M.R. Schroeder, New method of measuring reverberation time. *J. Acoust. Soc. Am.* **37**(3), 409–412 (1965). <https://doi.org/10.1121/1.1909343>
36. R. Scheibler, E. Bezzam, I. Dokmanić, Pyroomacoustics: A Python package for audio room simulations and array processing algorithms. arXiv e-prints arXiv:1710.04196 (2017). <https://doi.org/10.48550/arXiv.1710.04196>
37. N. Krishnamurthy, J.H.L. Hansen, Babble noise: Modeling, analysis, and applications. *IEEE Trans. Audio Speech Lang. Process.* **17**(7), 1394–1407 (2009). <https://doi.org/10.1109/TASL.2009.2015084>
38. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*. 2613–7 (2019)
39. Srivastava, A. Deleforge and E. Vincent, "Realistic Sources, Receivers and Walls Improve The Generalisability of Virtually-Supervised Blind Acoustic Parameter Estimators," *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. Bamberg, pp. 1–5 (2022) <https://doi.org/10.1109/IWAENC53105.2022.9914740>
40. F.R. Stöter, S. Chakrabarty, B. Edler, E.A. Habets, Countnet: Estimating the number of concurrent speakers using supervised learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**(2), 268–282 (2018)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I, Attention is all you need. *Advances in neural information processing systems*, CA, **30** (2017)
42. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *Proceedings of the 9th International Conference on Learning Representations*, (2021)
43. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, in *International conference on machine learning*. Training data-efficient image transformers & distillation through attention (PMLR), pp. 10347–10357 (2021)
44. Sun H, Liu X, Xu K, Miao J, Luo Q. Emergency vehicles audio detection and localization in autonomous driving. *arXiv preprint arXiv:2109.14797*. (2021)
45. G. Gwardys, D. Grzywczak, Deep image features in music information retrieval. *Int. J. Electron. Telecommun.* **60**, 321–326 (2014)
46. A. Guzhov, F. Raue, J. Hees and A. Dengel, "ESResNet: Environmental Sound Classification Based on Visual Domain Models," *2020 25th International Conference on Pattern Recognition (ICPR)*. Milan, pp. 4933–4940 (2021) <https://doi.org/10.1109/ICPR48806.2021.9413035>
47. K. He, R. Girshick and P. Dollar, "Rethinking ImageNet Pre-Training," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, pp. 4917–4926 (2019) <https://doi.org/10.1109/ICCV.2019.00502>
48. M. Tan, Q. Le, in *International conference on machine learning*. Efficientnet: Rethinking model scaling for convolutional neural networks (PMLR). California, pp. 6105–6114 (2019)
49. Pytorch profiler. (2020). [https://pytorch.org/tutorials/recipes/recipes/profiler\\_recipe.html](https://pytorch.org/tutorials/recipes/recipes/profiler_recipe.html). Accessed 21 Oct 2020
50. Karjalainen, M., Antsalo, P., Mäkivirta, A., Peltonen, T., Välimäki, V, Estimation of modal decay parameters from noisy response measurements. *J. Audio Eng. Soc.* **50**, 867–878 (2002)
51. T. Jasa, N. Xiang, Efficient estimation of decay parameters in acoustically coupled-spaces using slice sampling. *J. Acoust. Soc. Am.* **126**(3), 1269–1279 (2009)
52. G. Götz, R. Falcón Pérez, S.J. Schlecht, V. Pulkki, Neural network for multi-exponential sound energy decay analysis. *J. Acoust. Soc. Am.* **152**(2), 942–953 (2022)

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.