EURASIP Journal on Audio,
Speech, and Music Processing

EMPIRICAL RESEARCH **Open Access**

An end-to-end approach for blindly rendering a virtual sound source in an audio augmented reality environment

Shivam Saini[1,2*], Isaac Engel[1] and Jürgen Peissig[2]

Abstract

Audio augmented reality (AAR), a prominent topic in the field of audio, requires understanding the listening environment of the user for rendering an authentic virtual auditory object. Reverberation time (RT_{60}) is a predominant metric for the characterization of room acoustics and numerous approaches have been proposed to estimate it blindly from a reverberant speech signal. However, a single RT_{60} value may not be sufficient to correctly describe and render the acoustics of a room. This contribution presents a method for the estimation of multiple room acoustic parameters required to render close-to-accurate room acoustics in an unknown environment. It is shown how these parameters can be estimated blindly using an audio transformer that can be deployed on a mobile device. Furthermore, the paper also discusses the use of the estimated room acoustic parameters to find a similar room from a dataset of real BRIRs that can be further used for rendering the virtual audio source. Additionally, a novel binaural room impulse response (BRIR) augmentation technique to overcome the limitation of inadequate data is proposed. Finally, the proposed method is validated perceptually by means of a listening test.

Keywords BRIR augmentation, Reverberation time estimation, Binaural rendering

1 Introduction

To create a convincing illusion of a virtual auditory object in a given environment via headphones, two major requirements must be met. The first requirement involves placing the object within a space which considers the room acoustics and the second involves the anatomy of an individual which changes the characteristics of the sound before reaching the eardrum [1]. Placing the object in a space can be achieved by convolving a dry (or anechoic) signal with a room impulse response (RIR). This gives the listener a perception of sound originating from a specific location within a room, replicating the exact conditions under which the RIR was recorded. An RIR describes the behaviour of any sound source in a particular room and condition. It is typically divided into direct sound (DS), early reflections (ER), and late reverberations (LR) and it is well established that DS and ER have perceptually relevant directional properties [2]. However, a single omnidirectional microphone lacks the ability to capture the directional information, which makes it inadequate to produce an authentic auditory rendering. For a realistic rendering of a virtual auditory object, binaural room impulse responses (BRIRs) are captured which describe how the sound would travel from a sound source to the ears of a listener in a particular room [3]. An ideal BRIR measurement would require capturing the microphone signals at both the eardrums of the listener for every possible direction of view (DOV) resulting in an individual BRIR dataset. It includes the influences of both: the room's characteristics and the listener's physical presence. When convolved with a dry signal and played

*Correspondence:
Shivam Saini
shivam.saini@huawei.com
[1] Huawei Munich Research Center, Huawei Technologies, Munich 80992, Germany
[2] Institute of Communications Technology, Leibniz University Hannover, Hanover 10587, Germany

back over headphones, it should give that specific listener a perfect reconstruction of a virtual sound source. However, this is a very time-consuming task if it has to be performed for each listener in every listening condition and for all possible DOV.

Alternatively, to imitate the effect similar to individual BRIR, a generic BRIR can be recorded using a head-and-torso simulator (HATS). The direct sound from this BRIR can then be replaced with a head-related transfer function (HRTF) of the particular listener. It is well understood that performing binaural rendering with an individual HRTF rather than a generic one leads to higher plausibility and externalization [4–6]. Additionally, Werner et al. have demonstrated that the perceived externalization also depends on visual cues, due to acoustic divergence between the rendered and the listening rooms [7]. Hence, for the virtual auditory object to sound plausible and external in audio augmented reality (AAR), it is crucial to render a reverberation that is similar to that of the listening environment, since an incorrect rendering of room acoustics would destroy the illusion of realism [8].

However, capturing BRIRs in every possible environment would require an impractical amount of effort and prime apparatus. To overcome this challenge, data-driven approaches have been proposed that estimate room acoustics directly from noisy reverberant speech signals using machine learning (ML)-based techniques which can be seen dated as far back as 2001 [9]. In contrast to estimating room reverberation, other researchers propose estimating well-known room acoustic parameters such as reverberation time (RT) to get a rough understanding of the environment [10].

Apart from wide-band RT, frequency-dependent RT, energy decay curve (EDC), clarity (C_{50}), and direct-to-reverberant ratio (DRR) have been considered the most important parameters of a room that assist in the analysis of a given scenario [11]. Prior knowledge of these parameters could result in a more accurate representation of a virtual sound object in any given listening environment. But generally these parameters can only be calculated from high-quality measured RIRs and as mentioned earlier, measuring RIR is not practical at the user end of the applications due to the cost and effort involved. Hence, blind estimation of the parameters solely from reverberant speech signals has been of great interest to researchers in the field [10, 12–22].

RT_{60} is often considered as a key parameter to describe the acoustics of a space. It is a measure that defines the time taken for the sound energy to decrease by 60 dB after the source has stopped. As defined by ISO 3382-2 [23], based on an energy decay curve of an RIR, RT_{60} can be calculated by observing the time taken to reach 60 dB below the initial level used. But generally, from an RIR, RT_{60} is extrapolated based on a smaller dynamic range such as 30 dB (T_{30}), i.e. by taking twice of T_{30} value. Another well-known method to calculate RT_{60} of a space is given by Sabine's formula as:

$$RT_{60} = 0.161V/S\alpha \tag{1}$$

where RT_{60} is the time in seconds required for a sound to decay 60 dB, V is the volume of the room, S is the boundary surface area, and α is the average absorption coefficient. Numerous methods exist for its estimation blindly, i.e. without the use of an RIR, using audio signals including both signal processing and ML-based methods [10, 13–22]. A review of most of the given algorithms can be found in [24]. Another important parameter of room acoustic is the early-to-late reverberant ratio (ELR) or clarity (C_{50}). It is the ratio of the early sound energy (until 50 ms) and the residual energy in an RIR and is expressed in dB. As given by [25], C_{50} can mathematically be defined as:

$$C_{50} = 10\log\left(\frac{\int_0^{50} p^2(t)dt}{\int_{50}^{\infty} p^2(t)dt}\right) \tag{2}$$

Although C_{50} is often used as an indicator of speech clarity or intelligibility, it may assist in discriminating rooms with similar RT_{60} [11]. Several non-intrusive approaches exist for its estimation [26] or jointly with RT_{60} [13, 16].

In 2015, The ACE challenge [12, 24] created a benchmark for evaluation of RT_{60} and direct-to-reverberant ratio (DRR) estimation approaches. The benchmark allows the researchers to fairly evaluate their models against the state-of-the-art models. Eaton et al. [24] compared multiple estimation approaches in terms of mean squared error (MSE), estimation bias, and Pearson correlation coefficient (ρ). MSE is defined as the average of squared differences between predicted and real values and can be mathematically understood as:

$$MSE = \frac{1}{N}\sum(x_i - y_i)^2 \tag{3}$$

where, x_i is the ith observed value, y_i is the corresponding predicted value and N is the number of observations. Bias is the mean error in the results and is given by:

$$Bias = \frac{1}{N}\sum(x_i - y_i) \tag{4}$$

Finally, ρ is the Pearson coefficient correlation between the estimated and the ground truth results and is defined as:

$$\rho = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{5}$$

where, $\bar{x}$ and $\bar{y}$ are the means of real and predicted values respectively. By considering the MSE, bias, and ρ together, it is possible to determine how well an estimator performs [12]. Estimation results with a low bias and MSE might give an estimate close to the median for every speech file. But by examining ρ, it is possible to distinguish between such an algorithm (ρ close to 0). An algorithm which is more accurately estimating the parameter will have ρ closer to 1 with low MSE and bias close to 0. The results from ACE challenge showed that signal processing and ML methods had similar performance for RT estimation, while machine learning methods were better for DRR estimation at the time of the challenge.

2 Related works

Since the ACE challenge, a handful of publications have demonstrated the use of ML to estimate RT_{60} from noisy reverberant speech signals [14, 15, 18–20]. In 2018, Gamper and Tashev used convolutional neural networks (CNN) to predict the average RT_{60} of a reverberant signal using Gammatone filtered time-frequency spectrum [15] which outperformed the best method from ACE challenge. In 2020, Looney and Gaubitch [17] showed promising results in the joint blind estimation of RT60, DRR, and signal-to-noise ratio (SNR). On the other hand, Bryan [18] proposes a method to generate augmented training datasets from real RIRs which showed improvement in RT_{60} and DRR predictions. Ick et al. [21] introduced a series of phase-relate features and demonstrated clear improvements in the context of reverberation fingerprint estimation on unseen real-world rooms. However, one common limitation of most of the aforementioned works is that they only provide broadband parameter values, rather than frequency-dependent ones. This is an oversimplification of the room acoustics model which can potentially limit the rendering realism in the context of AAR.

Instead of using speech inputs, a few researchers have tried estimating acoustic parameters from wide-band inputs such as music signals [27]. The results demonstrate lower estimation accuracy in signals with music input than those with speech only. According to the authors, one of the reasons for this is the additional reverb used in the music content during the mixing process, which adds up in the input signal when convolving with the room reverb. In the study by Götz et al. [28], authors extended the work from [19] to estimate sub-band RT_{60} and C_{50} in dynamic conditions using convolutional recurrent neural networks (CRNN). When considering estimation results for music input, the model improves greatly in estimating RT_{60} under dynamic conditions over static conditions. This shows the ability of a model to differentiate between room reverberation and reverb from music when more than one condition is presented. The model trained with dynamic data shows improved performance even under static conditions. Although the results show estimation error being well under the just noticeable difference (JND) for RT_{60}, the model fails to predict C_{50} as accurately for $C_{50} > 15$ dB. The training and evaluation are however only performed on noiseless data. Furthermore, as pointed out by authors, the model only considers single-channel signals which lack spatial information and could be an aspect of future works.

The parameters estimated by ML methods can also be used to drive artificial reverberators to reconstruct the RIR properties of the space. For example, the method from Srivastava et al. [22] predicts not only RT_{60} but also room volume and total surface area, which could be useful parameters for the configuration of artificial reverberations. Alternatively, some methods also aim at the selection of an RIR from a database [11, 29].

On the other hand, few researchers have focused on directly estimating the RIR leveraging the audio-visual cues [30, 31]. Although it is a valuable approach to match the room acoustics of the current room for the case of augmented reality (AR), relying on visual cues might not be practical in scenarios like headphone listening where visual input is not available. Additionally, 10 s required to optimize the material [30] is unsuitable for real-time rendering where the listener moves between different rooms.

Steinmetz et al. proposed to estimate time-domain RIR solely from reverberant speech using filtered noise shaping in [32]. The approach shows a nearly perfect reconstruction of RIR from speech. However, the estimation of single-channel RIRs demonstrated in the publications may not be sufficient for auralization of spatial audio. Ratnarajah et al. [33] proposed M^3-AUDIODEC, a neural network-based multi-channel multi-speaker audio codec which overcomes the limitation of generating single-channel RIR. The approach compresses multi-channel audio while retaining spatial information. The approach can also aid in solving acoustic scene-matching problems. The method not only allows the reconstruction of reverberant speech but also provides the separation of clean speech and the BRIR that can be utilized to auralize other signals. The approach is however only tested on simulated data and the spatial coherence of other BRIRs in the same room has not been yet evaluated.

Most of the neural network approaches mentioned earlier make use of traditional CNN models with time-frequency representations (STFT, mel, etc.). In acoustic characterization, CNN is usually applied to solve the

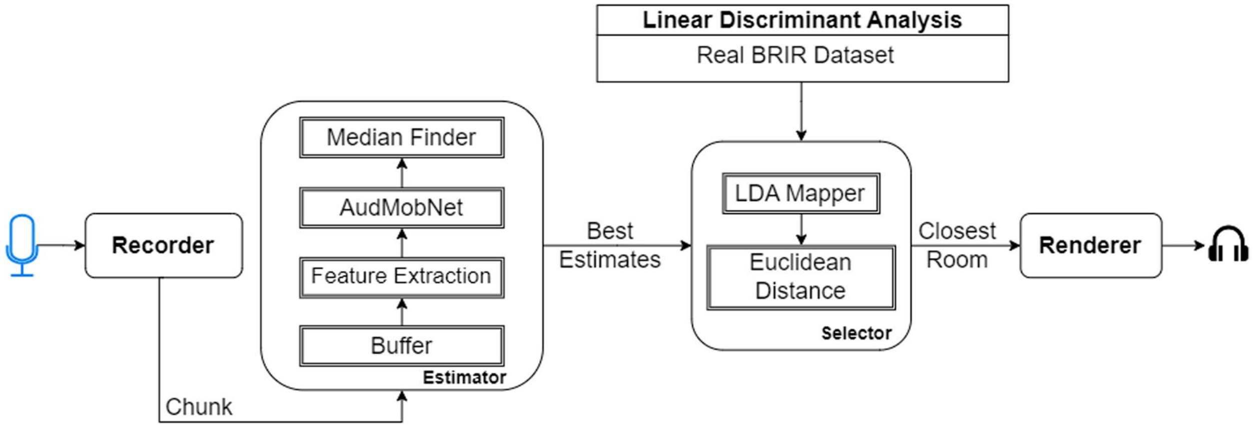


Fig. 1 The end-to-end blind rendering setup of the proposed method

blind acoustic parameter estimation problem as a regression task. CNNs are suitable for learning two-dimensional time-frequency signal patterns for end-to-end modelling, which is why they are widely used in aforementioned approaches. To capture long-range global context better, hybrid models that combine CNN with a self-attention mechanism have been proposed. These models have achieved state-of-the-art results for various tasks such as acoustic event classification and other audio pattern recognition topics [34].

Gong et al. [35] recently proposed a purely attention-based model for audio classification, called audio spectrogram transformer (AST). They evaluated AST against different audio classification benchmarks to achieve state-of-the-art results, which shows that CNNs are not always necessary in this context. However, transformers such as AST require a huge amount of computation power due to their complexity. To tackle this, in [13], a mobile audio transformer (AudMobNet) was proposed, which not only is independent of the length of the sequence but is also more robust against noise and computationally less expensive. The approach is only verified for broadband RT_{60} and C_{50} values. Upon further inspection, it was found that instead of only estimating single broadband RT_{60} value, sub-band RT_{60} and C_{50} can be beneficial in understanding the tonal characteristics of the room.

In this work, we aim to focus on estimating the acoustic parameters that correlate more with perceived plausibility. We extend our model (AudMobNet) presented in [13] to jointly estimate broadband and sub-band RT_{60} and C_{50} from noisy reverberant binaural speech signals. We also propose to improve the existing model architecture and use additional features such as phase and continuity differences to improve the performance of the model. Additionally, a novel multi-channel data augmentation technique to enhance the generalization capability of the network in sub-band RT estimation is presented. Lastly, leveraging on prior works [11, 13], an end-to-end blind spatial audio rendering setup is developed (Fig. 1) that takes a noisy speech recording as input to output a plausible binaural rendering of an arbitrary signal by means of BRIR selection. We use the estimated parameters to find closely related rooms from a dataset of high-quality BRIRs using linear discriminant analysis (LDA). The selected rooms are further used for rendering a virtual auditory object. Contrary to existing approaches mentioned earlier, our method solely relies on reverberant speech to binauralize the listening room and selection of 360 degrees BRIR set allows head rotation without any audible artefacts.

3 Proposed method

Our proposed blind end-to-end spatial audio rendering setup consists of 4 modules, namely recorder, estimator, selector and renderer (see Fig. 1). The estimator operates in a frame-wise manner in parallel with the recorder, and the input buffer length can be adjusted as needed. A longer buffer will lead to increased latency in the room parameter estimation (when in real-time), whereas a too-short buffer might produce inaccurate or unstable predictions. The estimated parameters are then used to select a set of 360 degrees BRIRs recorded in a real environment using the technique proposed in [11]. The selected set is then used for rendering the virtual object over the headphones through convolution. This end-to-end setup is used to evaluate the perceptual relevance of the proposed method (Section 4.4).

The most important module in the end-to-end setup is the Estimator which makes use of earlier presented AudMobNet [13] and extends it to the application of AAR rendering. Exploiting the binaural recording feature of new-age consumer end microphones, we can blindly estimate the room acoustics parameters that can be used to find a similar high-quality BRIR set from the database. This high-quality measured BRIR set ensures that the intra-set spatial coherence is preserved. We believe if the parameters of rendered BRIR are close enough to that of the listening room, the room divergence effect will be mitigated, resulting in a plausible virtual sound source [36]. Furthermore, it is believed that the head-tracking may also lead to a more authentic virtual sound source [37].

3.1 Model and feature input

The effectiveness of mobile transformers for estimating room acoustic parameters has already been demonstrated by the AudMobNet model proposed in [13]. Here, we propose using AudMobNet with additional changes for estimating room acoustic parameters from binaural signals. Instead of using mel-spectrograms as the only input to the network, we propose using inter-channel differences, exploiting the binaural nature of the problem, in addition to the logarithmic mel-spectrogram. It has been well established that for localizing a sound source, inter-channel (aural) phase differences (IPD) are the main contributors in low frequencies (< 1500 Hz) while the Inter-channel Level Differences (ILD) contributes in frequencies above 1500 Hz [38]. Furthermore, Ick et al. demonstrates that using phase and continuity features assist in improving RT_{60} estimation in low frequencies [21]. Similarly, results from Srivastava et al. [22] showed how using inter-channel features such as ILD and IPD lead to better parameter estimation over the networks where only single-channel features (STFT) are utilized. Hence, we believe that using IPD would help the network in understanding the low-frequency components. Furthermore, continuity features are used to track the phase variations across time which might help in understanding the overall context of the spectrum when estimating sub-band parameters. For boosting the generalization ability of the network and achieving full potential for estimating sub-band parameters, a BRIR augmentation technique is also presented which is discussed later in the section.

For the input data, the 2-channel raw audio is transformed into spectrograms using STFT with a sampling rate of 16,000 Hz and in frames with a 50% overlap using a Hann window. The STFT is further filtered with mel filterbanks generating a mel-spectrogram of the shape $M \times L$, where M is the number of mel bins and L is the length of the resultant spectrogram. Here, L depends on the frame size F, used for calculating STFT. For faster training and ease of evaluation, we keep M fixed to 64 bins but two different frame sizes are studied, 256 and 512. The mel-spectrograms are further used for generating phase and continuity features as in [21]. The mel features are then transformed to a logarithmic scale. The sine and cosine phase features from the left and right channels are then utilized to generate IPD as:

$$sinIPD(t,f) = sin(\theta_{t,f}) \tag{6}$$

$$cosIPD(t,f) = cos(\theta_{t,f}) \tag{7}$$

where, $\theta_{t,f} = \angle x_{t,f,l} - \angle x_{t,f,r}$ is the inter-channel phase difference between the mel-spectrogram x_l and x_r at time t and frequency f of the signals at microphones l and r. The second-order derivatives of IPDs are then calculated which we call Inter-channel Continuity Difference (ICD) and are given by:

$$sinICD(t,f) = \lim_{i \to 0} \lim_{j \to 0} \frac{\sin(\theta_{t,f}) - \sin(\theta_{t+i,f+j})}{\theta_{t,f} - \theta_{t+i,f+j}} \tag{8}$$

$$cosICD(t,f) = \lim_{i \to 0} \lim_{j \to 0} \frac{\cos(\theta_{t,f}) - \cos(\theta_{t+i,f+j})}{\theta_{t,f} - \theta_{t+i,f+j}} \tag{9}$$

As shown in Fig. 2, the features (sine and cosine IPDs; and sine and cosine ICDs) are stacked with logarithmic mel-spectrograms to generate 6-channel inputs. The 6-channel inputs are then masked with a time and frequency mask of size 64 and 16 respectively. During training, the mask is applied to the input randomly masking 64 timesteps and 16 mel sub-bands. This allows the model not to rely upon specific regions in the audio. It can be regarded as the usual and widespread dropout, but applied to the input and an example can be seen in Fig. 2a.

Also, since the input already has more than 1 channel, the spectrum could not be utilized to shorten the sequence into patches. Hence, to generate 16 patch embeddings for the model input as in [13], a 3 × 3 convolution layer is used instead. The convolution provides us with 16 representations of the 6-channel masked input as can be noted in Fig. 2. These patch embeddings are used as inputs for AudMobNet. The output linear layer produces 27 embeddings which include a single full-band RT_{60}, 8 sub-band RT_{60}s, and similarly 9 C_{50} values for each channel.

3.2 Datasets

Multiple publicly available datasets [39, 40] of measured BRIRs were used, resulting in an overall of 571 real BRIRs across 45 rooms. In addition, a highly detailed internal dataset of 6 rooms presented in [11] with 1440 real BRIRs was utilized. To balance the dataset, only 200 BRIRs were chosen from the latter, resulting in a total of 771 BRIRs. Different speech corpus [41, 42] and anechoic noise

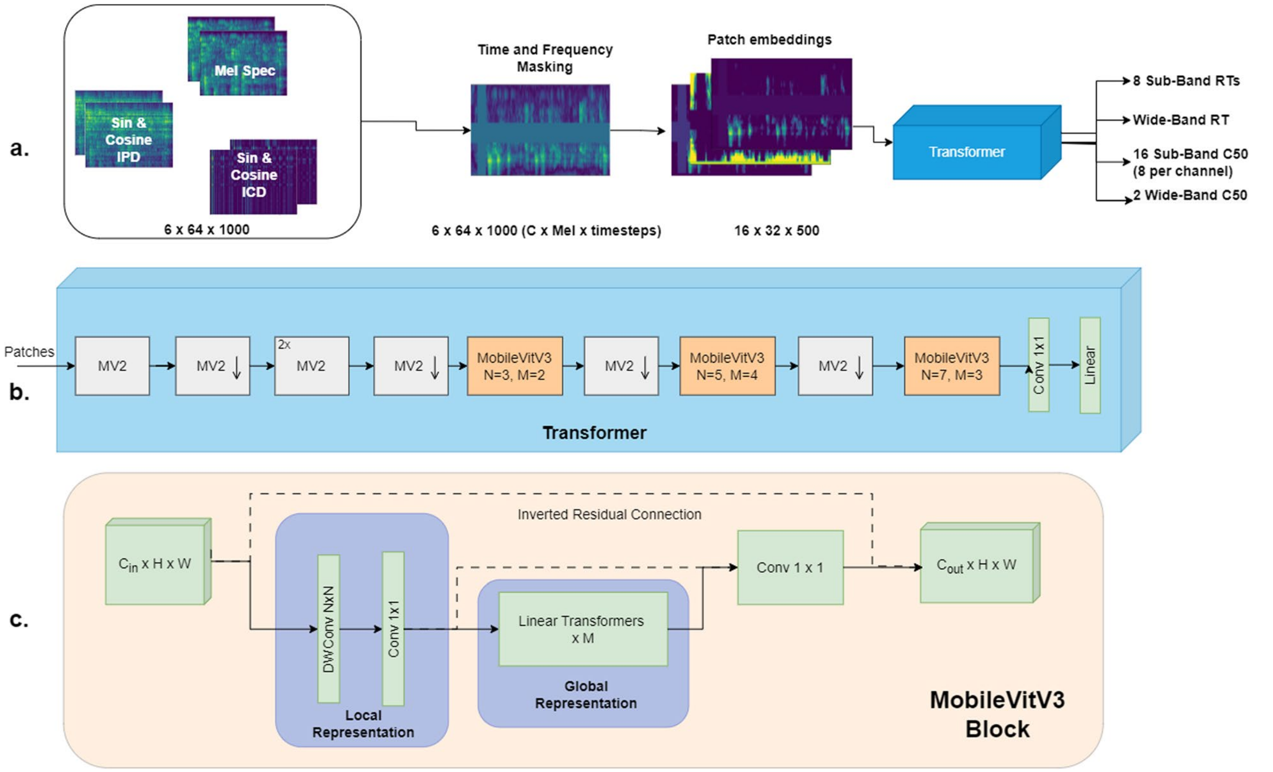


Fig. 2 Overview of the proposed model architecture. The top row (**a**) illustrates the pre-processing step where features are extracted from the binaural speech signal and stacked with the logarithmic mel-spectrogram resulting in the input shape of 6 × 64 × 1000 (C × H × W). Further, a time and frequency mask is applied to the input spectrum and the resultant is convolved with 16 kernels of size 3 producing 16 representations of size 32 × 500 which is treated as input to the transformer. The middle row (**b**) describes the architecture of the transformer where MV2 means MobileNetV2 block and ↓ means a reduction in the input size. *N* is the size of the kernel and *M* is the number of the linear transformers in the MobileViTV3 block. The bottom row (**c**) shows the architecture of each MobileViTV3 block where N and M are dependent upon the position of the block

dataset [43] were used to generate final samples. Babble noise was also simulated using a different speech dataset [44]. The BRIR, Speech and Noise datasets were split into training and evaluation sets to avoid any overlapping. For example, all BRIRs from 5 different rooms were taken in the evaluation set, and the speech samples were taken from the ACE evaluation set [12]. Further, in order to expand the size of the dataset and to tackle over-fitting, a novel BRIR augmentation technique was also incorporated. It shall be noted that the augmentation technique was only applied to the training set.

3.2.1 BRIR augmentation

We augment our BRIR dataset by parametrically modifying the diffuse tail (after mixing time) of the original BRIRs in different frequency bands, which allows us to mimic various tonal absorption patterns of rooms that are not present in the dataset. In [18], Bryan proposes replacing the later reverberation tail with a synthetic version generated using a Gaussian noise. Our technique is based on [18] with a few key differences. Firstly, our artificially generated Gaussian reverberation tail length varies in different sub-bands mimicking various tonal absorption patterns of sound in real rooms, in contrast to [18] where no frequency domain adaptation is incorporated. Secondly, due to the convolution with the original reverberation, our augmentation method also works for multi-channel RIRs. Finally, we also incorporated different decay types such as linear or logarithmic, taking into consideration more absorptive or reflective rooms.

In our augmentation approach, a circular set of BRIR is augmented by changing the reverberation tail. This circular set consists of $360/n$ BRIRs, where n is the spatial resolution in degrees which varies from room to room depending on the dataset. As shown in Fig. 3, to generate a reverberation tail, the original BRIR is filtered with mel-filterbank and RT is calculated for each mel sub-band. Then, a Gaussian noise with a similar length to that of the BRIR is generated. The noise is then convolved with the diffused part of the original BRIR to achieve a similar decorrelation between the left and right channels. The resulting binaural noise is

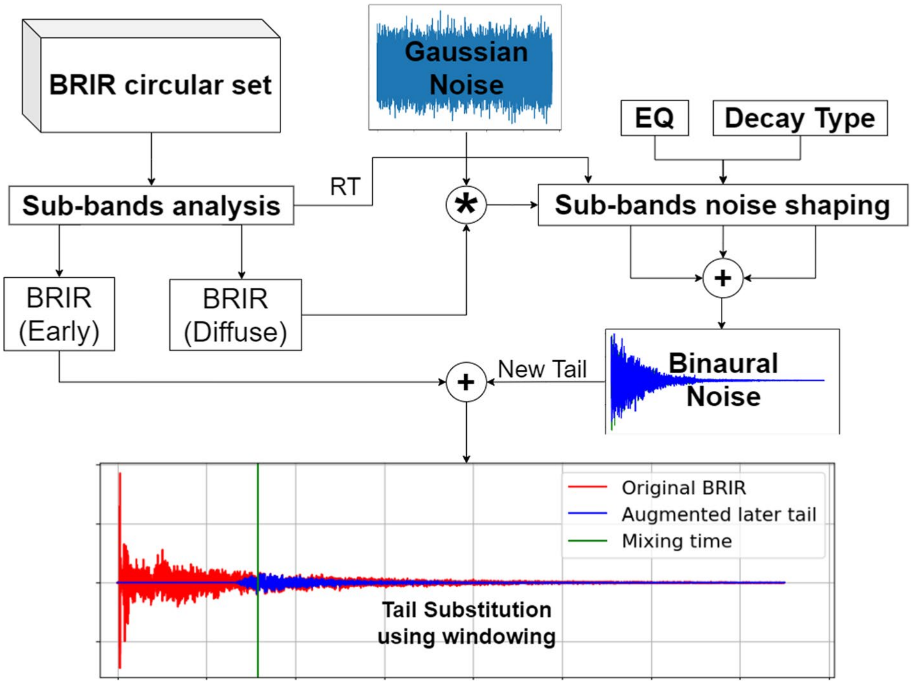


Fig. 3 Brief illustration of BRIR augmentation process

filtered with the same mel filterbank and the sub-band noise shaping is performed using different EQ gains for each sub-band. Back in the time domain, a decay filter is applied for each band using either a linear or a logarithmic decay with ±500 ms RT of that of the original sub-band band RT. All the sub-band tails are then added up to generate the full-band BRIR tail. The old tails are then replaced in all the BRIRs from the same room after the mixing time. The augmented BRIR $h_r(t)$ is generated from the real BRIR $h(t)$ by replacing the later tail with $h_{aug}(t)$ using a crossfade that can be interpreted as:

$$h_r(t) = \begin{cases} h(t), & t_0 \le t \le t_m - t_w \\ h(t).w_e(t) + h_{aug}(t).w_n(t), & t_m - t_w < t \le t_m + t_w \\ h_{aug}(t), & t_m + t_w < t \end{cases} \tag{10}$$

where, t_m is the approximate mixing time as given by [45] and calculated as $t_m = 80 \cdot RT_{500Hz}$. $w_n(t)$ is the first half of a Hanning window of 0.2 $(2 \cdot t_w)$ s and $w_e(t)$ is the later half of the window. Note that this augmentation technique can be applied to RIRs with any number of channels but due to the scope of research, only BRIR augmentation is demonstrated. Approximately 20,000 BRIRs were generated out of 671 real BRIRs using a single position from 40 rooms and the generated responses can be seen in Fig. 4.

3.2.2 Data pre-processing

It is worth mentioning that the convolution of each augmented BRIR (20,000) with each speech sample (6000) requires an impractical amount of disk space. Hence, during the generation of the training sets, the BRIRs were convolved with 50 random speech signals. For the real BRIR training dataset, 671 BRIRs were convolved with random speech signals resulting in 35,000 real samples and for evaluation set consisted of 1000 real samples generated from measurements from 7 separate rooms. The number of total generated samples are given in Table 1. The ground truth values were computed as defined by the ISO [23]. The RT_{60} is extrapolated by taking twice the T_{30} value as suggested by the ISO. To obtain sub-band $RT_{60}s$, the BRIR was first filtered with a mel-frequency filterbank and a RT_{60} value for each filtered BRIR was calculated similarly. Sub-band and wide-band C_{50} were calculated according to Eq. 2.

To simulate acoustic signals, the BRIR was convolved to each dry speech signal, s(t), and then noise n(t) was added as,

$$x(t) = h(t) * s(t) + a.n(t) \tag{11}$$

where, $*$ shows the convolution operator. For generating noise realistically, all available BRIRs in one room are selected. Afterwards, 30 random speech samples are

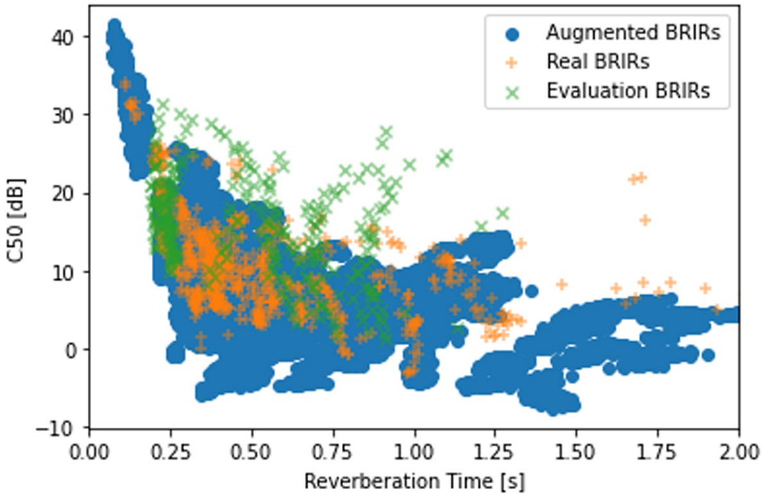


Fig. 4 Six hundred seventy-one training, 100 evaluation and 20,000 augmented BRIRs

Table 1 Summary of data used during training and evaluation

Data split	Number of rooms	Number of BRIRs	Total generated samples
Real training set	40	671	35,000
Augmented training set	-	20,000	200,000
Evaluation set	7	100	1000

picked from the noise-generation set mentioned in Section 3.2. Random EQ gains and overall gains are applied to the samples. The resultant samples are convolved with randomly chosen BRIR from the room. The convolution is done with only the diffuse tail of the real BRIR allowing us to generate diffused and spatial noise. The silent parts are removed from the signals to have continuous noise. Other noises, such as static ambient noise, babble noise, or office noises are also created similarly. The generated spatial noises ($n(t)$) are then multiplied with a gain constant (a) and added to the convolved speech signal at random SNR ranging from 6 to 30 dB. Although the network is able to adapt to variable-length sequences, to allow batch training, the input signals were trimmed/zero-padded to the length of 4 s. Another reason for choosing this length is the 4-s duration of most speech samples. Further, BRIRs with $RT_{60} < 2$ s were considered which involves most of the real scenarios and the 4-s signal is enough to contain all the necessary information.

3.3 Training setup

After generating the input signals mentioned in the previous section, the 6-channel features are extracted (as shown in Section 3.1) which are used as inputs for the neural network. As seen in Table 5, 4 different configurations are presented to evaluate the effectiveness of the proposed method. In the first configuration, a single AudMobNet is used as in [13] to generate 29 embeddings (8 sub-band and a wide-band RT_{60} and 8 sub-band and a wide-band C_{50} for each channel) but instead of using the single channel mel-spectrogram input, we used 2-channel logarithmic mel-spectrogram inputs. In the second and third configurations, we use additional mel Phase and Continuity features for each channel similar to Ick et. al. [21], along with logarithmic mel-spectrograms producing 4-channel and 6-channel inputs respectively. In the final configuration, we propose using sine and cosine IPDs along with ICDs together with logarithmic mel-spectrograms of both channels as explained earlier to produce 6-channel inputs. Rather than looking at the features as two separate channels as in the third configuration, in the proposed method the difference of the sine and cosine phase components gives the network a better understanding of the behaviour of low frequencies. By stacking these features together, we can compare the performance differences of these features while keeping the same model complexity.

3.4 BRIR selection and rendering

In order to evaluate the proposed method, an end-to-end system was built as explained at the end of Section 1, which takes a noisy speech recording as input and is able to output a plausible binaural rendering of an arbitrary signal through BRIR selection. As mentioned in Section 3.2.2, to allow batch training, the training samples were trimmed/padded to the length of 4 s, allowing the network to generalize better for this specific length. Furthermore, the results suggest higher accuracy for longer input samples when compared against signals shorter

Table 2 RT_{60} predictions from AudMobNet [6] at 1 kHz for different input durations samples from ACE challenge

Input duration [s]	MSE	ρ	Bias
Shorter than 2	0.015	0.86	– 0.019
3	0.018	0.85	– 0.021
4	0.015	**0.94**	0.018
Longer Than 4	**0.012**	0.91	**0.016**

than 3 s (see Table 2). However, it is not efficient to use 30-s-long input sequences for making a single prediction. Hence, the samples are chopped into shorter lengths such as 4-s-long segments to get more stable estimations using multiple predictions in the Estimator module (Fig. 1). A hop size of 0.5 s was found to be a good tradeoff between latency and accuracy, hence this is chosen as the hop size for the buffer for further evaluation. Finally, the median values of predictions for the full input signal length are chosen to be the best estimates.

After the Estimator module (Fig. 1), the Selector module selects 2 best-matching room based on the parameters. Similar to the technique presented by Treybig et al. [11], linear discriminant analysis (LDA) is performed on the dataset of all real BRIRs. This separates all the rooms in the latent space based on parameters provided. The latent space is then stored in the disk along with eigenvectors. During the runtime, the predicted parameters are plotted in the same space using the saved eigenvectors. The closest measurements to the predicted parameters are selected as the best matching rooms using the nearest neighbour technique which are further used in the Renderer.

Since the dataset consists of BRIR measurements that have irregular spatial resolution, we employ a BRIR interpolation technique to obtain 1° resolution for all the BRIR circular sets. We follow the dynamic time warping (DTW)-based interpolation technique as presented by [46, 47]. The Renderer module performs binaural rendering by convolving dry audio signals with the selected BRIR pair utilizing partition convolution [48]. The BRIR pair is swapped in real-time according to the listener yaw orientation, which is tracked with a MotionNode inertial-measurement-unit head tracker (5-ms latency).

4 Evaluation

We present a concise evaluation of the room parameter estimation and the BRIR augmentation techniques separately to understand their contribution.

4.1 Model selection and input features

4.1.1 Preliminary model evaluation

In Saini and Peissig [13], we presented a mobile audio transformer to estimate wide-band RT_{60} and C_{50} from single-channel noisy reverberant speech signals. We evaluated the model against the relevant baselines using the benchmark evaluation criteria and dataset provided by the ACE Challenge [12]. The results (Table 3) demonstrate that using the proposed mobile audio transformer effectively improves the estimation accuracy of both the parameters. It is to be noted that these models only estimate wide-band parameters from single-channel signals. Apart from high accuracy, another advantage of this method is its ability to adapt to variable length input due to its unique hybrid transformer architecture. The model achieves higher accuracy even in low SNR levels (Fig. 5) while keeping the model complexity low (see Table 4). estimation compared to Baseline

4.1.2 Input features

In this work, we extend the AudMobNet L model to estimate sub-band parameters from noisy reverberant binaural speech signals (see Section 3.1). To evaluate our feature extraction method, we compared it against two approaches. The first approach involves the previously employed technique that involves mel-spectrogram as the only input feature proposed in [13] and in the second approach we use an input feature extraction technique similar to [21]. The proposed feature extraction technique is compared against both of the methods mentioned above in Table 5. All the methods are compared for 2 different frame sizes (256 and 512 samples) resulting in a total of 8 configurations to be evaluated.

The models round up to 1.2 million trainable parameters for each configuration with approximately 500 MFLOPS, making it suitable for deploying on mobile devices. All configurations were trained on the same 35,000 samples (with only real BRIRs) so they could be fairly assessed.

Table 3 Evaluation results from [13] on the ACE challenge evaluation set [12] for wide-band RT_{60} predictions from single-channel noisy speech signals. On the left is a comparison against Baseline [17] and best-performing method [49] from the ACE challenge and on the right for wide-band C_{50} estimation compared to Baseline [26]

Model	ρ	MSE	Bias [s]	Model	ρ	RMSE [dB]
RT Baseline [17]	0.84	–	–	C_{50} Baseline [26]	0.77	3.05
QA Reverb [49]	0.77	0.07	– 0.07	-	-	-
AudMobNet S	0.89	0.03	0.13	AudMobNet S	0.81	2.8
AudMobNet L	**0.90**	**0.02**	**– 0.05**	AudMobNet L	**0.85**	**2.3**

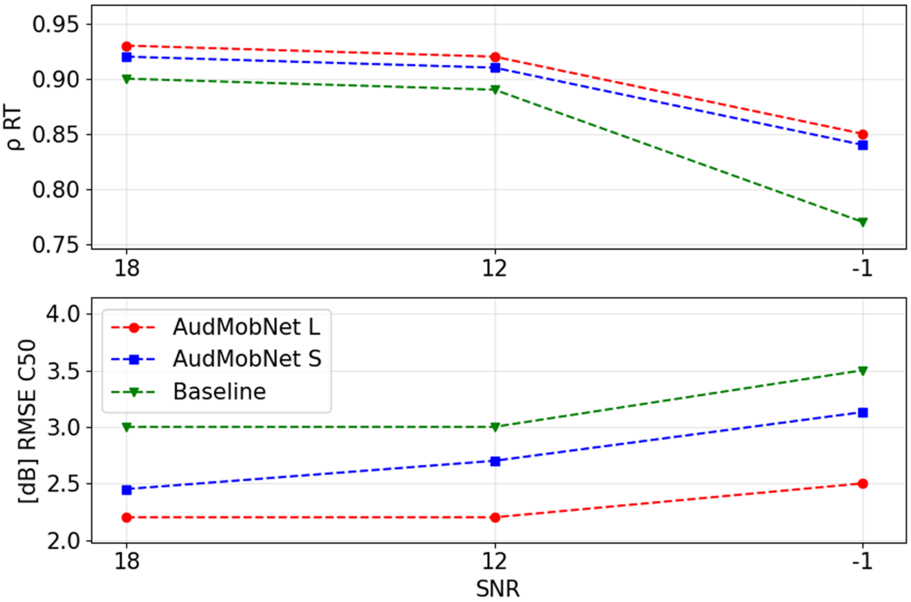


Fig. 5 Effects of SNR levels [dB] on the performance of the AudMobNet variations and baseline models evaluated on the ACE evaluation set. The top graph displays ρ for RT_{60} estimates in comparison to the RT Baseline [17] while the lower graph illustrates the RMSE [dB] for C_{50} estimates compared to the Baseline [26]. The comparison metrics were selected based on those provided in the respective baseline

Table 4 Number of trainable parameters and floating point operations per second (FLOPS) in millions per sample

Model description	Params	MFLOPS
AudMobNet S	**61 K**	**70**
AudMobNet L	1.2 M	501
Baseline RT [17]	4 M	1950
Baseline C_{50} [26]	74 K	135

Table 5 Evaluation results for wide-band RT_{60} and C_{50} (left channel only) for the models trained only with real training data. The input features describe the type of inputs and the size of the window

Model and input feature	ρ_{RT}	MSE_{RT}[s]	$\rho_{C_{50}}$	$RMSE_{C_{50}}$[dB]
Mel_{256} [13]	0.74	0.04	0.82	3.1
Mel_{512} [13]	0.67	0.03	0.80	3.3
+ $Phase_{256}$ [21]	0.76	0.04	0.88	2.7
+ $Phase_{512}$ [21]	0.75	0.04	0.79	3.1
+ Phase and $continuity_{256}$ [21]	0.78	0.03	0.85	**1.8**
+ Phase and $continuity_{512}$ [21]	0.79	0.03	0.88	2.4
+ IPD and ICD$_{256}$	**0.87**	**0.01**	**0.93**	1.9
+ IPD and ICD$_{512}$	0.85	0.01	0.84	2.9

The evaluation set consisted of 1000 real samples from 7 unique rooms generated in a similar manner as the training set but with the ACE evaluation speech dataset [12]. The distribution of wide-band parameters and an LDA using wide-band and sub-band parameters is given in Fig. 15 (Appendix). The training uses stochastic gradient descent on the mean-squared loss with an initial learning rate of 0.001 using an Adam optimizer. Since the parameters differ in units (s and dB), the training set was first standardized using min-max normalization. This gives each parameter equal weight when calculating loss. Batch size was selected manually to get the best out of each model and the available resources. Models were trained until convergence and the best-performing epoch was selected for each configuration.

The evaluation results presented in Table 5 demonstrate that using extra features improves the overall estimation compared to where only mel-spectrograms are utilized as input to the network. Although calculating the phase and continuity features as [21] is straightforward and improves the wide-band parameter estimation, it may not be the best choice for RT_{60} estimation, be it wide-band or low-frequency sub-bands (Fig. 6). The strongest correlation can be observed between 500 Hz and 2 kHz across all the models, aligning with the findings from the ACE challenge [24]. The reason for this behaviour can be partially attributed to the spectral distribution of energy present in speech signals. Further, the results from Fig. 6 agree with our hypothesis that the use of sine and cosine IPD and ICD assist the network in understanding low-frequency reflection better, resulting in higher ρ value for parameter estimation in the low frequencies especially below 1500 Hz. Using the

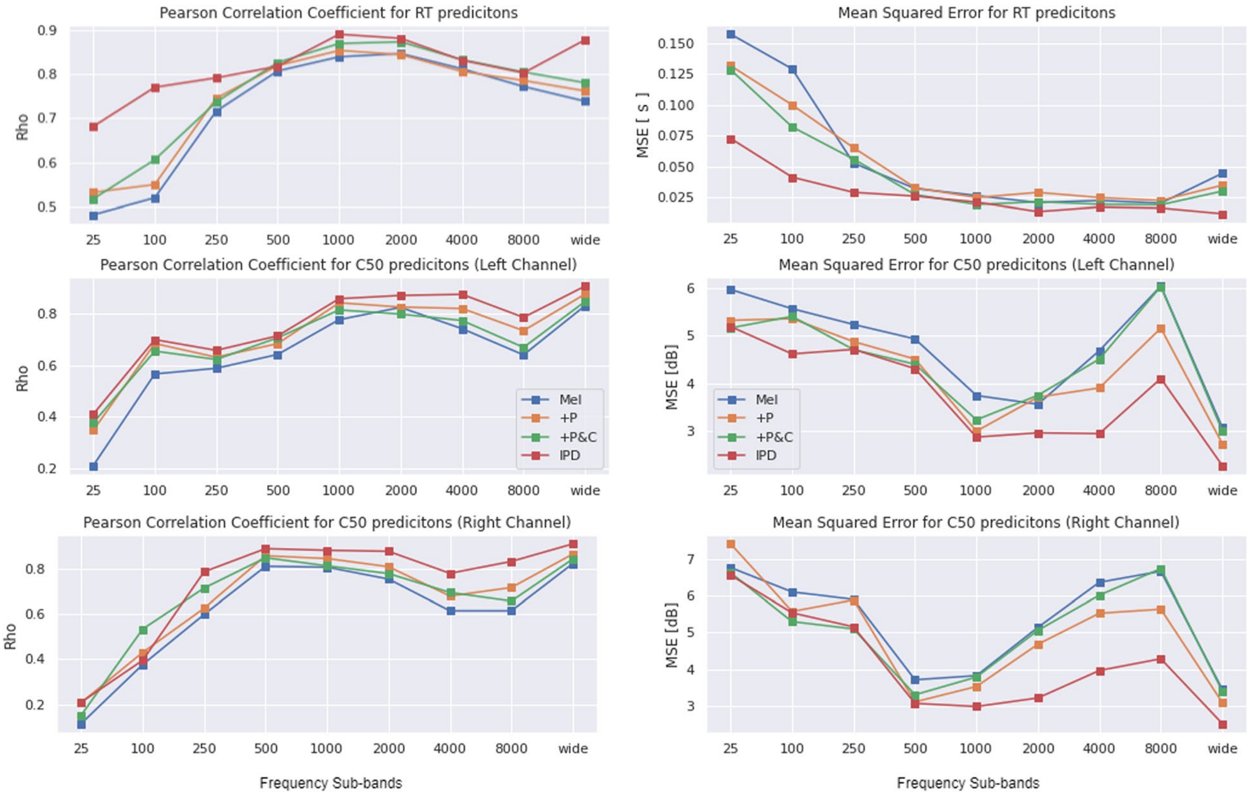


Fig. 6 Sub-band RT_{60} and C_{50} estimation evaluation only for the models with real training data (window size of 256 samples)

inter-channel differences not only improves the low-frequency prediction but also tends to improve the wide-band estimation accuracy. The lower value of MSE for the sub-band as well as wide-band parameters further confirms our hypothesis. One reason for this could be higher inter-channel decorrelation in rooms where more energy in late reverberation is present leading to a longer RT_{60} and a smaller C_{50}. However, this shall be investigated in independent research.

4.1.3 Model output

During the course of this research, we also compared if estimating the sub-band parameters has any advantage over estimating only wide-band parameters. The results from Table 6 show that even if only mel-spectrograms are utilized as feature inputs to estimate the parameters, the model tends to have a better understanding of the input and wide-band parameters estimations to be more accurate. A similar trend can be seen for estimations made with additional inter-channel differences.

4.1.4 Input sample size

In many scenarios, such as in real-time, it may not be possible to pre-process and predict having a small window size (256 samples) resulting in slower computation. Hence, using the window size of 512 samples when generating STFTs could reduce the computational complexity of the network by almost a factor of 2 when compared to the STFTs generated using 256 samples as well as the overall prediction time by at least a factor of 1.5. Using a longer window also drastically brings down the computation time since the matrix multiplication with the filterbank when computing the mel-spectrogram also reduces. Furthermore, the IPD and ICD calculation of smaller samples (spectrograms with bigger window size) is considerably faster due to fewer samples in the time axis. As a drawback, predictions for C_{50} are not as accurate because of the larger window

Table 6 Wide-band parameter (RT_{60} and C_{50}) estimation results for the proposed model (AudMobNet + IPD for sub-band parameter estimation) compared to the same model for wide-band parameter estimation

Model	ρ	MSE	ρ	RMSE [dB]
Mel wide-band [13]	0.75	0.02	0.82	3.1
Mel sub-band [13]	0.81	0.05	0.84	3.0
+IPD wide-band	0.83	0.01	0.89	2.2
+IPD sub-band	**0.87**	**0.01**	**0.93**	**1.9**

Table 7 Evaluation results for the proposed model trained with real data and augmented data compared to the baseline wide-band RT_{60}

Training data type	MSE	ρ	Bias
Mel_{256} [13]	0.044	0.74	0.014
Mel_{256} [13] + augmented data	0.013	0.85	− 0.038
+ IPD and ICD_{256}	0.014	0.87	− 0.04
+ Augmented data	**0.006**	**0.93**	**− 0.001**

size used due to energy binning involved (see Table 5). Smaller windows would have more time information which is useful for the calculation of C_{50}. This energy is binned together when larger windows are used for the calculation of STFT resulting in lesser time information available for the network. Overall, the performance of the proposed method with a window size of 512 samples is comparable to the model presented in [13] with a window size of 256. Furthermore, the proposed network outperforms the existing model by accuracy in broadband as well as sub-band parameters but at a cost of 1.2 times slower inference time that is required to calculate IPD and ICD features.

4.2 Data augmentation

To show the effectiveness of the proposed data augmentation technique, the best network, i.e. the configuration with IPD and ICD and with a frame size of 256, was additionally trained with 200,000 samples generated using the technique presented in Section 3.2.1. The model from [13] was trained on the same data as a baseline for comparison.

Figure 7 demonstrates the MSE between ground truth and predicted values for each sub-band from 20 to 8000 Hz. The effectiveness of incorporating noise shaping in frequency sub-bands within our augmentation technique is clearly evident, as indicated by the low mean squared error (MSE) value. Furthermore, the results presented in Table 7 suggest an improvement in full band RT_{60} prediction as well when compared to the models trained with real data only. Finally, employing different decay types benefits the model in predicting full-band C_{50} bringing the ρ value as high as 0.94.

4.3 BRIR selection

In Section 3.4, we propose a BRIR selection method based on the output from the proposed neural network model that allows us to select a set of 360-degree BRIRs from a database of 45 rooms which are closest to the estimated parameters utilizing LDA of the dataset. The selected BRIR set is further used for rendering a virtual

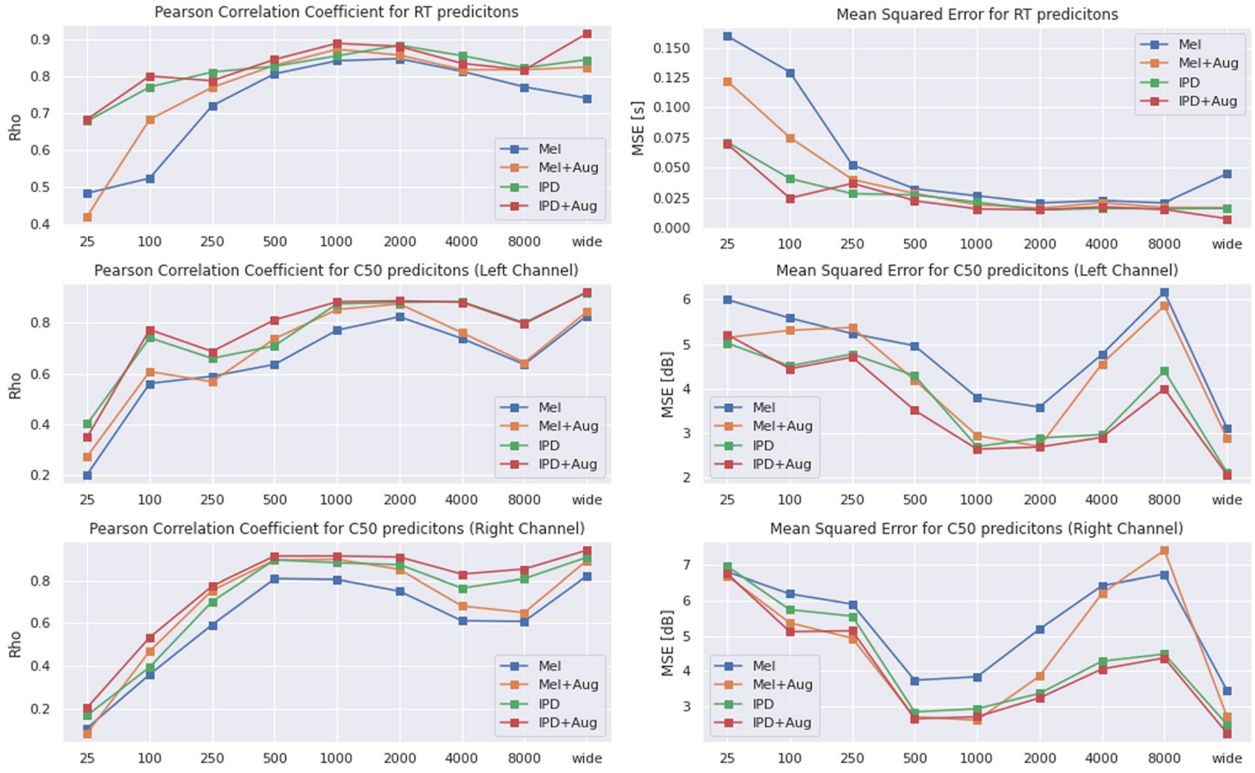


Fig. 7 Evaluation results of models trained with only real data compared to the models additionally trained with augmented data

Table 8 MAE for wide-band RT_{60} and C_{50} estimation (left channel) of each mentioned scenario and ground truth

$M^3 - AUDIODEC$ [33]	RT_{60}	C_{50}
BRIR selection (Selector only)	23.8	1.36
Estimator only (AudMobNet + IPD)	**15.6**	**0.67**
Estimator + selector	26.5	1.5
	22.7	0.79

Table 9 Comparision of baseline and proposed method on different grounds

-	$M^3 - AUDIODEC$ [33]	Proposed
Real-time factor (RTF)	0.046	**<0.01**
Parameters	500 M	**1.1 M**
Input type	Clean reverberant speech	**Noisy reverberant speech**
BRIR output	Noisy BRIR	**HQ 360-degree BRIR set**

sound source. The closest approach to extracting a BRIR solely from speech signal was recently presented by Ratnarajah et al. ($M^3 - AUDIODEC$) [33]. Although, as mentioned in Section 2, the method is designed for neural compression of binaural signals, it may also be applied to extract BRIR from the signal. Hence, we use it as the baseline for evaluation of our BRIR selection.

To evaluate our selection method against [28], we generated 30,000 new samples with the script provided in the GitHub repository[1] and fine-tuned our model on these samples. One major reason for this being unable to train $M^3 - AUDIODEC$ on our data due to its huge size. The test set consisted of 752 samples generated using VCTK Speech Corpus [50] and BRIRs simulated with Pygsound[2].

Table 8 reports the mean of absolute errors (MAE) between the real and estimated parameters for 4 cases. The first case is selector only, i.e. when a BRIR is selected from the dataset directly using ground truth values in the LDA space. The second case considers the output of our neural network model. In this case, the model was fine-tuned and modified to output 4 embeddings, i.e. each parameter for each channel. The third case selects a room using LDA based on the output of the proposed model as used in our end-to-end approach. Principal component analysis (PCA) is also performed in this case, similar to [11] to find the best-matching BRIR in the selected room so that the C_{50} error is minimum. To be noted, the room selected (in cases 1 and 3) is from the dataset of real BRIRs and hence might not be as accurate as generating a BRIR resulting in larger errors (see bottom image in Fig. 8). This does not mean that our method performs worse than [33] but is a limitation of the size of the dataset to select a BRIR from. On the other hand, the BRIR predicted by [33] shows neural network noise (top image in Fig. 8 after 40,000 samples) and hence to calculate the parameters, only the first 0.25 s of the whole BRIR is used.

Table 9 presents feature comparisons of the baseline [33] and the proposed method. We can see that our model being approximately 500^{th} of the baseline in terms of parameters, shows faster estimation resulting in a very small RTF value. This shows us that the model can be deployed on a mobile device while providing accurate real-time parameter estimation. Our model is also robust against noise as can be seen in Fig. 5 as opposed to the baseline, which was trained on simulated data with no additional noise added to the input. Furthermore, the BRIR set selected through our approach consists of high-quality 360-degree BRIR sets which were recorded in different rooms at multiple positions and hence can be used for spatial auralization of an object providing some degrees of freedom of head rotation without introducing any audible artefacts. This also leads to a more plausible illusion of a virtual object, as shown in the next section.

4.4 Perceptual evaluation

We measured 360-degree BRIR sets in two different rooms with non-identical acoustical characteristics to test the effectiveness of the proposed method perceptually. These measured sets were used as the hidden references in the listening test.

4.4.1 *Listening test measurement setup*

To measure the rooms, a dummy head was placed at a distance of 3 m from a Genelec 8020 speaker and circular sets of BRIRs were measured in the two rooms. Both rooms are different in shape, size and reverberation patterns. The rooms are shown in Fig. 9. Room A imitates a living room environment and is dry while Room B complements it as a reverberant meeting room. Room A has an average RT_{60} of 0.29 s with the room dimensions of 10.5 m x 5.25 m x 2.75 m (L x B x H) while Room B has an average RT_{60} of 0.65 s with the room dimensions of 12.5 m x 5.75 m x 2.75 m. The measurements from Room A have a C_{50} of 17 dB while that of Room B is 10 dB. A comparison of IRs and sub-band RT_{60} from both rooms is given in Fig. 10. A glance into the figure reveals how Room B has strong first two reflections and a long decay time. A meeting table in the middle of the measurement

[1] https://github.com/anton-jeran/MULTI-AUDIODEC

[2] https://github.com/GAMMA-UMD/pygsound

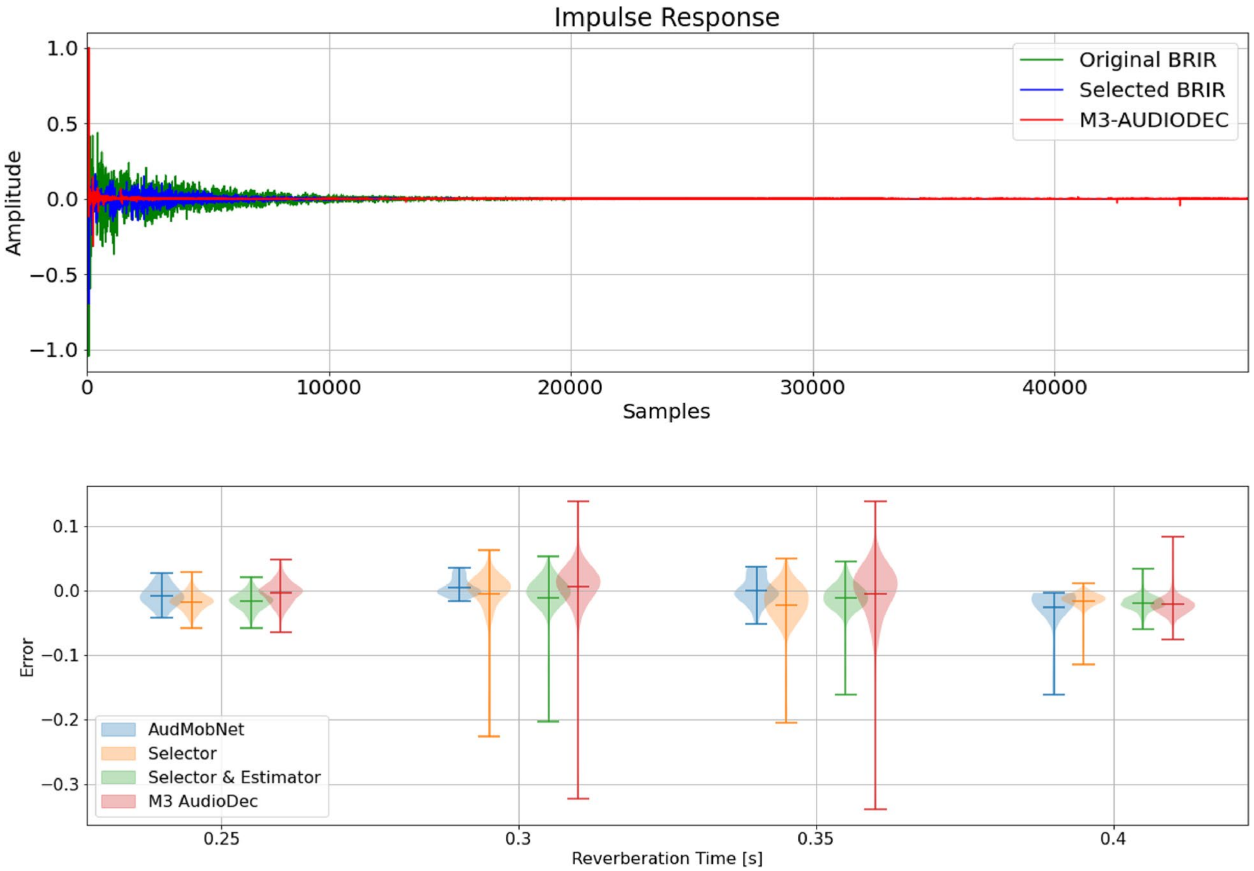


Fig. 8 Comparison of BRIR output (top) and MAE for RT_{60} estimates from proposed methods against $M^3 - AUDIODEC$ [33]. Noise in the output BRIR from baseline can be noted after 40,000 samples. The test data consists of 752 samples generated using the script provided in [33]

setup creating a strong first reflection can be noticed in the impulse response of Room B adjacent to the direct sound. On the other hand, Room A has prominent early reflections but a shorter decay time due to the type of material used in the room such as carpet and absorbing curtains.

We recorded 30 s of speech using Genelec 8030 CP loudspeakers in both rooms using the same measurement setup described above. For best results, the recorded signal is then chopped into multiple 4-s-long sequences with a hop size of 1 s. The distribution of all the full-band RT_{60} and C_{50} predictions made by the proposed model for both rooms can be seen in Fig. 11. Furthermore, the comparison in Table 10 demonstrates the differences in predicted median values and ground truth. Finally, the sub-band RT_{60} predictions can be noted in Fig. 12. All the median values are then used to find the two best matching rooms using the Selector described in Section 3.4.

For each scenario, the two best matching rooms were selected from the dataset of real BRIRs using the method proposed in Section 3.4 for the listening test. A comparison of sub-band RT_{60} of the best matching rooms are given in Fig. 12 and wide-band RT_{60} and C_{50} are given in Table 11. Although it is evident from Table 10 that the predicted wide-band parameters are well under the JND of 5% and 1.1 dB for RT_{60} [51] and C_{50} [52], the best-matched rooms (Table 11 and Fig. 12) may not be. At 1 kHz, the RT_{60} for 2nd best-matched room of room B is almost 24% longer which is slightly longer than the JND of 22% defined by [53]. Similarly, the wide-band RT_{60} and C_{50} values for best-matching rooms fall outside of JNDs.

To validate the proposed approach, a subjective listening test similar to MUSHRA [54] was employed (see Fig. 13). The audio signal was convolved with the selected BRIR sets in real time to be able to provide head-tracked binaural output. Participants were asked to rate the plausibility of the sound source on a scale from 1 to 5. For naive listeners, the explanation of plausibility was the perceived size of the room compared to the reference where 1 was much smaller and 5 was much bigger. The listening setup had automatic switching, so when the listener would take off the headphones, the reference loudspeaker would play. Beyerdynamic's DT 990 pro headphones were utilized for the

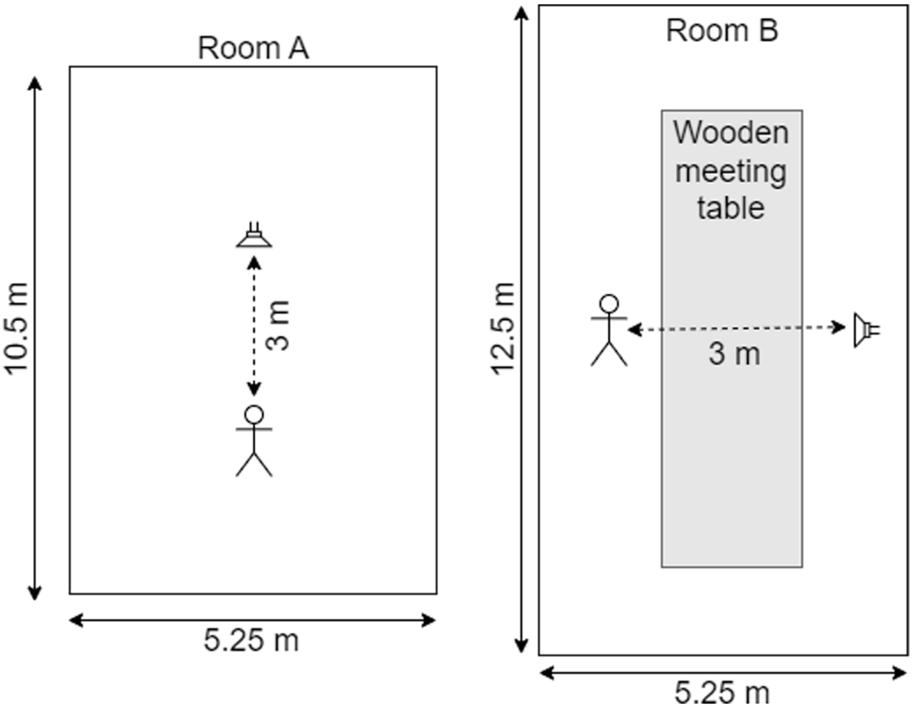


Fig. 9 Measurement setup for listening test in room A (left) and room B (right)

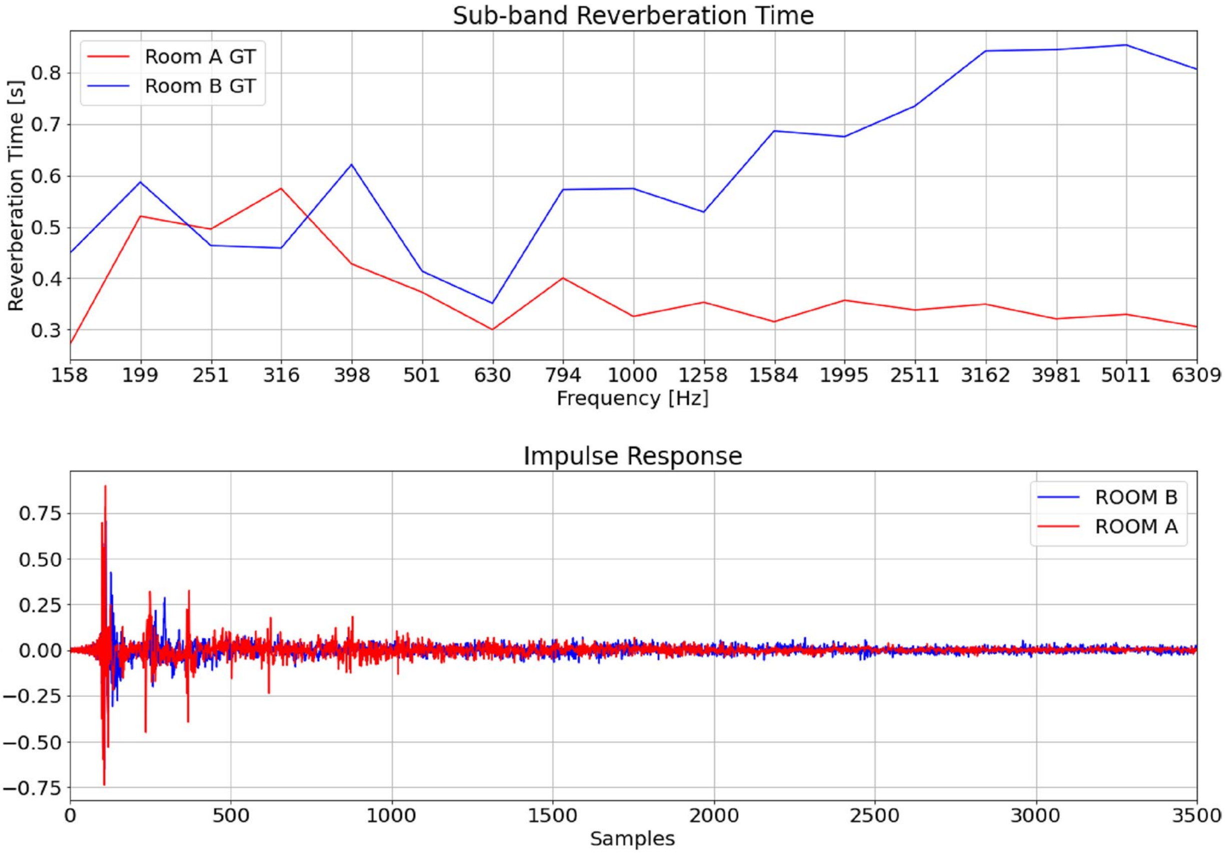


Fig. 10 Comparision of sub-band RT_{60} (top) and impulse responses (bottom) for room A and room B

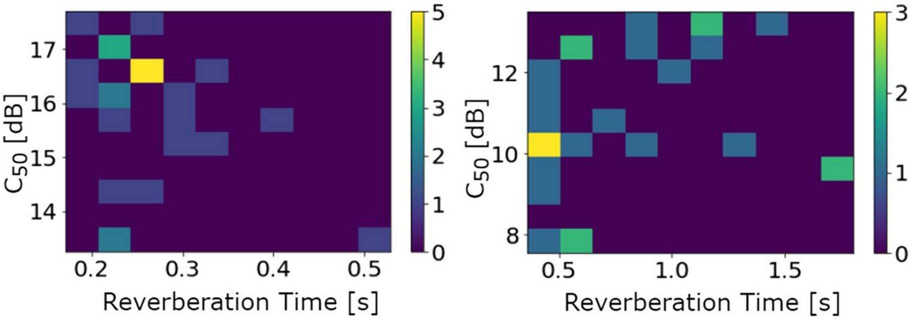


Fig. 11 Distribution of 30 RT_{60} and C_{50} predictions from the model for room A (left) and room B (right). Violet means no estimation was made for the value and yellow represents the most number of estimations

Table 10 Calculated and estimated parameters from room A and room B

Room	Ground truth RT_{60}	Prediction RT_{60}	Ground truth C_{50}	Prediction C_{50}
Room A	0.29	0.244	17	16.24
Room B	0.65	0.63	10	10.5

Table 11 Ground truth parameters of all the rooms used in the listening test

Reverberant condition	RT_{60} Room A	RT_{60} Room B	C_{50} Room A	C_{50} Room B
Reference	0.29	0.65	17	10
Dry anchor	0.1	0.1	29	29
Reverberant anchor	0.57	0.57	6	6
Best match 1	0.27	0.67	15	7
Best match 2	0.35	0.69	16	6

listening test due to their openness characteristics and flatter response. A headphone equalization was also applied before playing back the stimuli through the headphones.

The listening test consisted of 4 parts, i.e. listening in two different rooms: room A and room B; and with two different stimuli types: female speech and music. The female speech was taken from the ACE challenge dataset [12] and "Get Lucky" by Daft Punk was chosen as the pop music piece. Initially, in room A, a small training session was done where the listener was presented a speech signal convolved with BRIR sets from rooms with very different acoustics were used to demonstrate the difference in reverberation in rooms (see Fig. 13, left).

In the actual test, the listeners were asked to blindly compare 5 BRIR sets in each room including 2 best matches (of the listening room), 2 anchors (reverberant and dry) and a BRIR set of the listening room (see Table 11). The anchors remain the same in both the listening rooms. Twelve listeners participated in the listening test including 7 audio and acoustic experts, 2 with audio background and 3 naive listeners. To be noted, the BRIR sets used for training were different than the ones in the test and the order was also randomized.

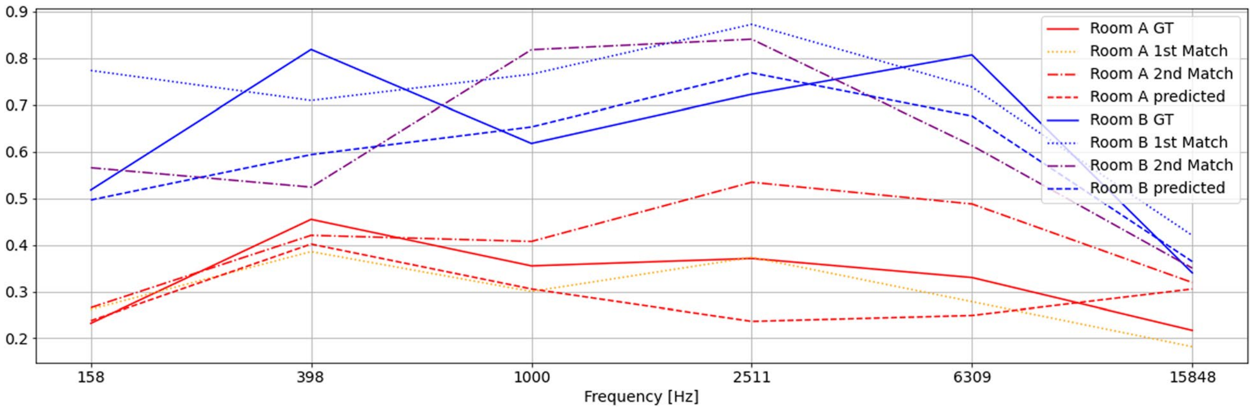
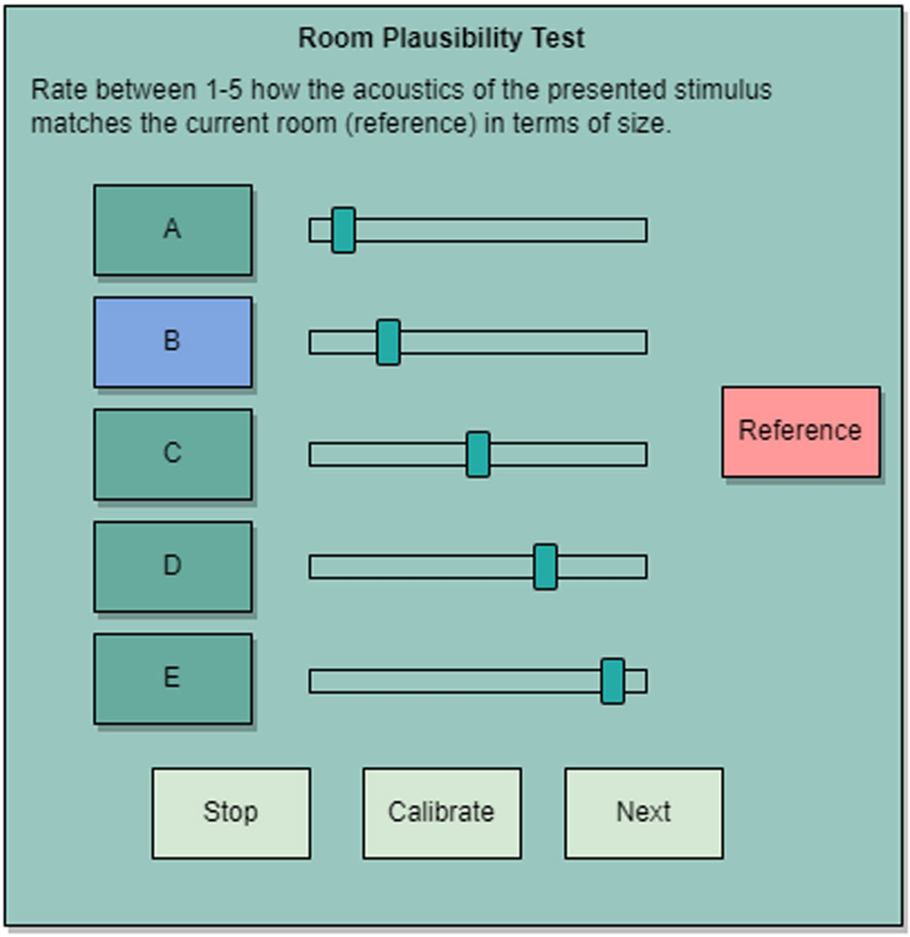


Fig. 12 Sub-Band RT_{60} for the listening room (ground truth), median-predicted (by the model) and selected rooms



Room Plausibility Test

Rate between 1-5 how the acoustics of the presented stimulus matches the current room (reference) in terms of size.

Fig. 13 Schematic of graphical user interface for the listening test

4.4.2 Results and discussion

A repeated measure analysis of variance (RM-ANOVA) test was performed on the data of each paired comparison to study the effect of the different variables and their interactions. An RM-ANOVA test gives us statistically significant differences in three or more dependent samples. In our case, three such cases can be seen: listening room, type of stimulus, and presented reverberant condition. Two values, i.e. the p-value and the F-value are presented from RM-ANOVA tests. The *F*-value in an RM-ANOVA represents the ratio of the variance between the groups to the variance within the groups. It tests the null hypothesis that there is no significant difference between the means of the groups, and a larger *F*-value indicates that the difference between the means is more likely to be significant. The *p*-value associated with the *F*-value represents the probability of observing such an extreme *F*-value by chance if the null hypothesis were true. Therefore, a smaller *p*-value indicates that the result is more statistically significant.

Not to be missed, the data from the listening did not pass the Mauchly sphericity test ($p < 0.001$) and the Greenhouse-Geisser epsilon was 0.69, which is smaller than 0.75, so the Greenhouse-Geisser correction was applied, following the ITU-R MUSHRA recommendation [55]. Afterwards, the data was grouped based on the results and a *t*-test was performed for each group. A significance value (α) of 0.05 was used.

Effect of the listening room and stimulus: The RM-ANOVA found no significant effect of the listening room on listeners' ratings [$F(1,11) = 0.6074$, $p = 0.4522$]. Further, no significant effect of the choice of stimulus was found [$F(1,11) = 2.8209$, $p = 0.1212$]. The interactions between the listening room and stimulus also yielded a higher *p*-value resulting in no significant difference [$F(1,11) = 0.2558$, $p = 0.6230$].

Effect of the reverberant condition: The RM-ANOVA found a significant effect of different reverberant conditions on listeners' ratings [$F(4,44) = 133.5074$, $p < 0.001$]. Furthermore, significant interaction with listening room [$F(4,44) = 109.3112$, $p < 0.001$], type of

Table 12 *p*-values of paired *t*-test performed on listener's ratings on speech for all reverberant conditions in room A

	Reverberant anchor	Dry anchor	Best match 1	Best match 2
Reference	< 0.001	< 0.001	**0.678**	**0.591**
Reverberant anchor	-	< 0.001	< 0.001	< 0.001
Dry anchor	-	-	< 0.001	0.003
Best match 1	-	-	-	**0.343**

Table 13 *p*-values of paired *t*-test performed on listener's ratings on music for all reverberant conditions in room A

	Reverberant anchor	Dry anchor	Best match 1	Best match 2
Reference	< 0.001	**0.014**	**0.244**	0.009
Reverberant anchor	-	< 0.001	< 0.001	0.001
Dry anchor	-	-	< 0.001	< 0.001
Best match 1	-	-	-	**0.508**

Table 14 *p*-values of paired *t*-test performed on listener's ratings on speech for all reverberant conditions in room B

	Reverberant anchor	Dry anchor	Best match 1	Best match 2
Reference	< 0.001	< 0.001	**0.555**	0.009
Reverberant anchor	-	< 0.001	< 0.001	< 0.001
Dry anchor	-	-	< 0.001	0.003
Best match 1	-	-	-	**0.095**

Table 15 *p*-values of paired *t*-test performed on listener's ratings on music for all reverberant conditions in room B

	Reverberant anchor	Dry anchor	Best match 1	Best match 2
Reference	< 0.001	< 0.001	0.003	**0.343**
Reverberant anchor	-	< 0.001	**0.269**	**0.088**
Dry anchor	-	-	< 0.001	< 0.001
Best match 1	-	-	-	**0.024**

stimulus [$F(4,44) = 5.6823$, $p < 0.001$] and all three variables [$F(4,44) = 5.9779$, $p = 0.0006$] were also found.

Post hoc pairwise t-tests were run on data separated by listening room and type of stimulus for each reverberant condition using a corrected significance level of $\alpha' = 0.01$. All the p-values are presented in Tables 12, 13, 14 and 15. Furthermore, for the ease of understanding, all pairs with no significant differences are marked with arrows in Fig. 14. The t-test found no significant difference in listeners' ratings for the reference and best match 1 in all the scenarios except for the case when listening to music in room B. It is apparent from Fig. 14 that many listeners found the room bigger than the reference in this case. One reason for this is a longer reverberation time in the low frequencies for best match 1 (0.8 s at 150 Hz) when compared to the ground truth (0.5 s) which is more prominent in the case of music with drums and bass rather than female speech as can be seen in Fig. 12.

The t-test between reference and best match 2 follows a similar trend showing no significant difference in listener's ratings. Although for music in room A and speech in room B, results show significant differences, it shall be noted the p-value (0.0095) is very close to α' in both cases. A general trend of best match 2 being slightly bigger than the reference can be noted. The reason for this could be longer RT_{60} (20% for best match 2 in room A) or lower C_{50} (4 dB for best match 2 in room B) than the JNDs in such cases. Furthermore, the results also show no significant difference between best match 1 and best match 2 under all circumstances. Similarities with respective anchors in each room can be noted in Fig. 14 but significant differences were found except for the dry anchor to the reference when listening to music in room

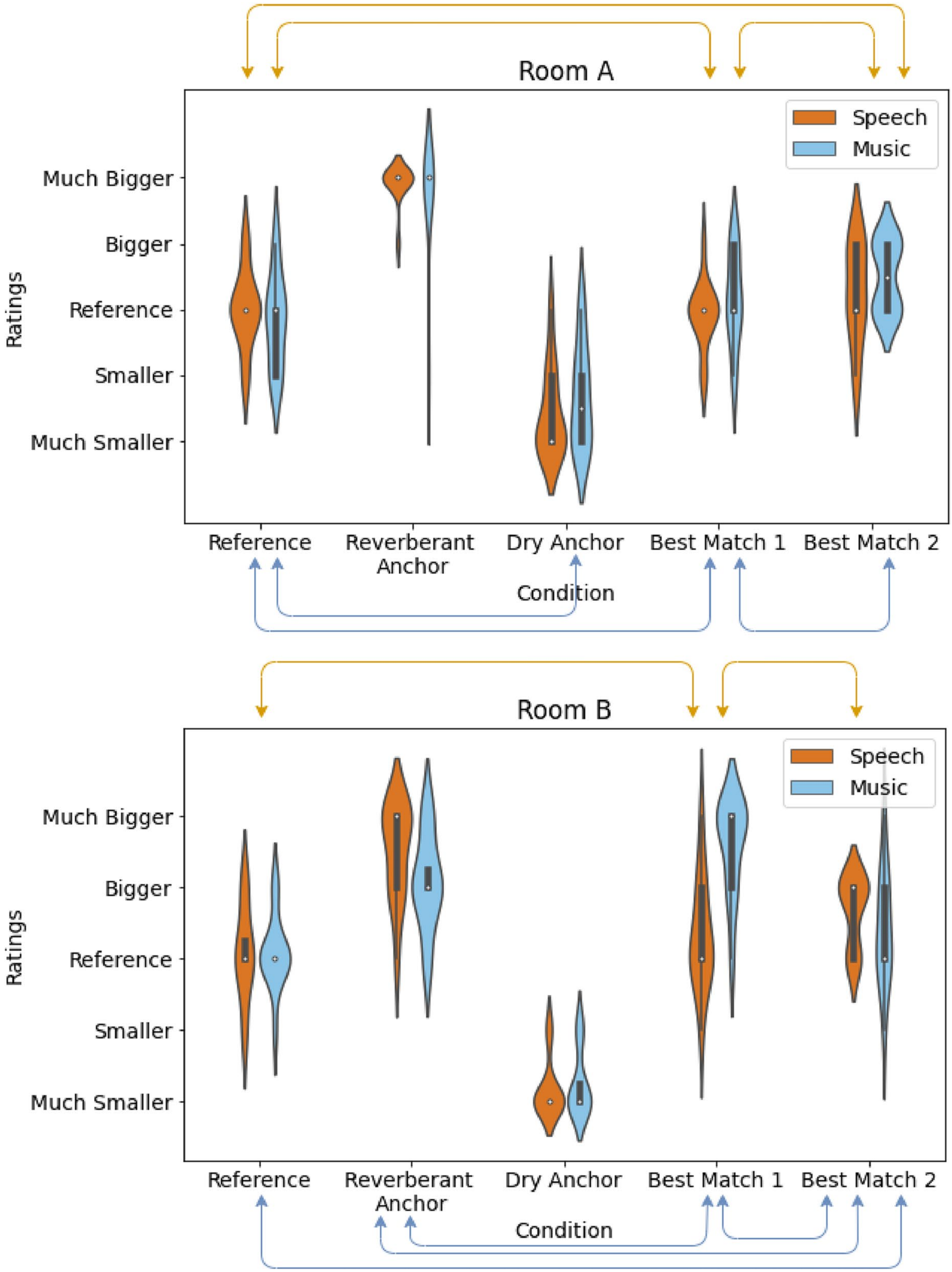


Fig. 14 Violin plots for the listening test results. The top image shows the results for the listeners' ratings in room A and the bottom one demonstrates ratings in room B. The respective coloured arrow means no significant difference was found in the listeners' ratings between the pair of the listening condition

A, but the anchor was mostly rated as smaller than the reference. This reason could be due to presence of reverb in the music piece itself that adds up to the room reverb.

5 Discussion

This study demonstrates the possibility of rendering a virtual sound source blindly from noisy reverberant speech signals. The presented results show that the proposed method is able to render a sound source which sounds similar to a physical sound source in the room. In Section 4, objective and subjective evaluations are presented. A thorough evaluation of each proposed method is given which shows the effectiveness of the method. From the objective perspective, our Estimator module shows improvements against the state-of-the-art models in overall estimation accuracy (wide-band and sub-band), inference speed and robustness against noise. Further, it was confirmed that predicting sub-band parameters along with wide-band parameters helps the network to understand the data better. We also discussed the impact of different window sizes used to calculate feature inputs. Finally, the overall end-to-end setup was evaluated objectively against the relevant baseline which shows the advantages and disadvantages of the proposed method. From the listening test results, we discovered that in most cases, especially when rendering speech, the proposed method is able to produce such results that the listeners perceived the best matching room to be the same size as (or similar to) the actual listening room. However, the same cannot be said for the music signals. There are reasons involved in each stage.

Input signal: Although the proposed network outperforms the state-of-the-art estimation techniques, its dependence solely on speech signal input affects the sub-band RT_{60} estimation. The absence of low frequencies (< 85 Hz) in the speech signals makes it difficult for the network to estimate the parameters precisely in this frequency range. This can also be seen in Fig. 7. Although the proposed method improves the estimation accuracy in low frequencies, it may still result in incorrect estimation leading to an incorrect choice of BRIR. One solution for this could be estimating the parameters using signals with a wider frequency spectrum [27, 28] however it comes with a drawback of overall lower estimation accuracy as described in Section 2.

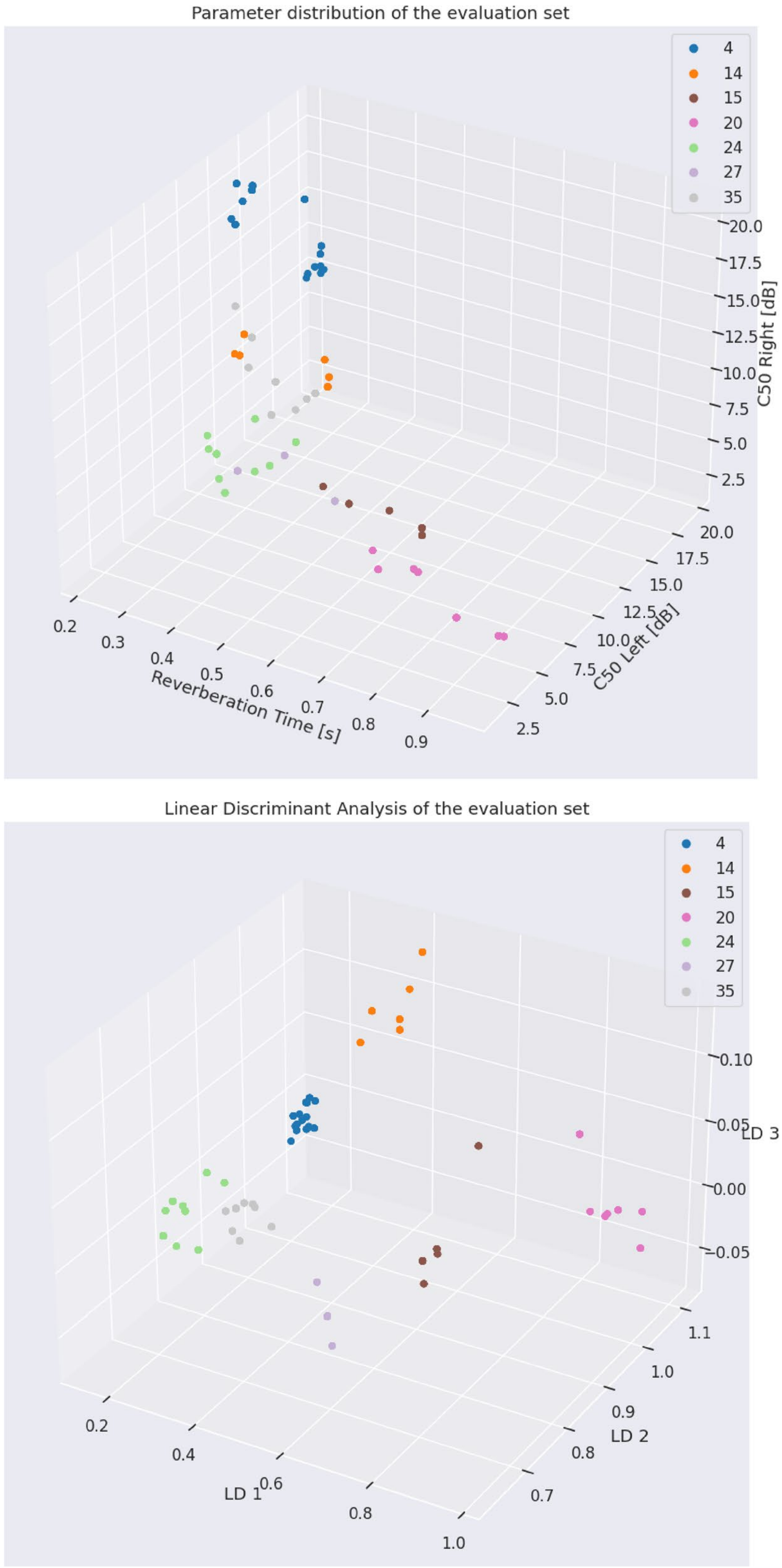
Estimated parameters: As seen from the results (Fig. 14), the convergent case, i.e. where the listening room is the same as the BRIR set, has mostly been rated most similar. Next to the convergent case, the best-match BRIR sets obtained the most similar ratings to the reference room. While the estimated parameters may relate closely to the perceptual differences, they still might not be enough to accurately describe a room. For example, even though the estimated parameters are similar when listening in room B, significant differences with the reference were found in listener ratings with the best matching rooms. This could be due to the early reflections coming from the meeting table which may have influenced the C_{50} but since more weight is given to sub-band RT_{60} in LDA [11], the matched room results in a C_{50} smaller than the JND. As a result, the matched scenarios sound distant, unreal, or bigger. Hence, the perceptual relevance of other parameters such as interaural cross correlation (IACC) and/or Initial Time Delay Gap (ITDG) shall also be investigated.

Lack of real BRIR dataset: Another possibility of how these perceptual differences could have been avoided is by the inclusion of more BRIR sets. Currently, the dataset consists of only 45 real rooms (with only a single or a few positions per room) that may or may not be closely related to the listening room as also discussed in Section 4.3. In the future, the estimation and augmentation techniques could be improved by considering more perceptually relevant acoustic features. However, further research is still needed on this topic, to better understand the mapping between objective acoustic metrics and the perception of room similarity. An alternative approach to this problem could be to modify/optimize the selected dataset to fill the gap between the predicted parameters and the selected rooms's parameters.

6 Conclusions

In this work, we propose a novel technique to blindly render any given listening environment from a speech signal. This is done in two steps. Firstly, the parameters are estimated from a noisy reverberant binaural speech signal using a mobile audio transformer. We propose improvements to the previously presented model AudMobNet for the estimation of the parameters such as by using additional features (phase and continuity) and supporting binaural signals. We demonstrate how using input features such as inter-channel phase difference (IPD) and its second-order derivative can effectively improve the overall performance of the network, especially in low frequencies RT estimation. To further improve the performance of the estimation technique, we propose a BRIR augmentation technique which can be further used to augment any multi-channel RIRs. Our augmentation approach demonstrates major improvement when compared to the state of the art in the sub-band RT_{60} estimation as well as the full-band RT_{60} and C_{50} estimation. Secondly, the estimated parameters are then utilized to select a circular set of BRIR from a given dataset using LDA that is used for rendering a virtual sound source. A perceptual evaluation was also carried out during the study and the results demonstrate the selected BRIRs to be as plausible as one based on a BRIR recorded in the actual listening room. Finally, we discuss the gaps in the presented method.

Appendix


Parameter distribution of the evaluation set

Linear Discriminant Analysis of the evaluation set

Fig. 15 Acoustic parameter distribution of all the BRIRs used in evaluation set (top) and the LDA using all the parameters (bottom)

Abbreviations

AAR	Audio augmented reality
AR	Augmented reality
AST	Audio spectrogram transformer
BRIR	Binaural room impulse response
CNN	Convolutional neural network
CRNN	Convolutional recurrent neural network
DOV	Direction of view
DRR	Direct-to-reververant ratio
DS	Direct sound
DTW	Dynamic time warping
ELR	Early-to-late reverberant ratio
EDC	Energy decay curve
EQ	Equalization
ER	Early reflections
HATS	Head and torso simulator
HRTF	Head-related transfer function
IACC	Inter-channel cross correlation
ICD	Inter-channel continuity difference
ILD	Inter-channel level difference
IPD	Inter-channel phase difference
ITDG	Initial time delay gap
JND	Just noticeable difference
LDA	Linear discriminant analysis
LR	Late reverberations
MAE	Mean absolute error
MFLOPS	Million floating point operations per second
ML	Machine learning
MSE	Mean squared error
PCA	Principal component analysis
RIR	Room impulse response
RM-ANOVA	Repeated measures analysis of variance
RMSE	Root mean squared error
RT	Reverberation time
SNR	Signal-to-noise ratio
STFT	Short-term Fourier transform

Acknowledgements

We would like to extend our gratitude to all the colleagues from Audiovisual Tech Lab at Huawei Munich Research Center who helped us by participating in the listening tests.

Authors' contributions

With constant feedback on the manuscript from IE and his major contribution to the perceptual evaluation, SS was able to finish this research in time. JP backed the research with his valuable input and support.

Authors' information

SS, born in 1995, is currently pursuing his Ph.D. at Huawei Munich Research Center in collaboration with Leibniz University, Hanover, Germany. He received his M.Sc. from Technische Universität Ilmenau in 2020. In 2019–2020, he worked on his Master's Thesis in Binaural Reproduction at Mercedes-Benz Research and Development in Stuttgart, Germany. Since after, he has been working on his Ph.D. in Spatial Audio Rendering applications.

IE received a B.Sc. degree in Electronic Systems Engineering and Master's degree on Telematics and Telecommunication Networks from the University of Malaga in 2015 and 2016, respectively. He was a doctoral (and briefly, postdoctoral) researcher at Imperial College London between 2016 and 2021, investigating spatial audio perception and signal processing methods for binaural audio rendering, obtaining his Ph.D. in 2021. In 2018 and 2019, he worked as a Research Intern at Facebook Reality Labs on the topics of headphone equalization and binaural perception. As of 2022, he is a Senior Research Engineer at the Audiovisual Technology Lab at Huawei Munich Research Center.

JP received his Ph.D. in Physics from the University of Göttingen, Germany, in the fields of Acoustics, Psychoacoustics, and Digital Signal Processing. In 1991 he worked at Bell Laboratories, Murray Hill, in the group of Gary Elko. In 1995 he joined Sennheiser electronics R&D in Germany. Since 2004 he has been lecturing on Acoustics and Signal Processing at Leibniz University of Hannover. Beginning in 2005, he was responsible for the Sennheiser Research facility in Palo Alto, CA, focusing on Digital Signal Processing for Audio. Since 2014 Dr. Peissig has been heading the communications systems group at the Institute of Communications Technology of Leibniz University of Hannover. Today he lectures on Digital Signal Processing, Communication Schemes, Signal Processing for Acoustics, and Psychoacoustics and Electroacoustics. Dr. Peissig is a member of the IEEE Communications Society, Audio Engineering Society, and German Acoustical Society. His actual fields of research in audio and acoustics are: signal processing for acoustics sensor and actuator arrays, acoustics noise cancellation processing, and Audio-Signal Processing and Psychoacoustics for Immersive, 3D Virtual and Augmented Reality Audio.

Funding

This work is part of the PhD research of Shivam Saini at Leibniz Universität Hannover commissioned by Huawei Technologies Düsseldorf GmbH.

Availability of data and materials

The data generated and the model used during the study cannot be made public as it is part of ongoing research. However, it is believed that the information given in this paper is sufficient for the reproduction of the methodology. Furthermore, the collection of BRIR dataset used for the training of the model is publically available under ASH-IR github repository. The data augmentation technique presented is described in Section 3.2.1 and easily reproducible with a few lines of code. The AudMobNet model used for training could be reproduced using the instructions given in MobileViTv3 [56], followed by Saini and Peissig [13] and finally Section 3.1.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 9 November 2023 Accepted: 6 March 2024
Published online: 27 March 2024

References

1. R. Gupta, J. He, R. Ranjan, W.S. Gan, F. Klein, C. Schneiderwind, A. Neidhardt, K. Brandenburg, V. Välimäki, Augmented/mixed reality audio for hearables: Sensing, control, and rendering. IEEE Signal Proc. Mag. **39**(3), 63–89 (2022). https://doi.org/10.1109/MSP.2021.3110108
2. J.E. Summers, Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms, and Acoustic Virtual Reality. J. Acoust. Soc. Am. **123**(6), 4028–4029 (2008). https://doi.org/10.1121/1.2908264
3. H. Møller, Fundamentals of binaural technology. Appl. Acoust. **36**(3), 171–218 (1992). https://doi.org/10.1016/0003-682X(92)90046-U. https://www.sciencedirect.com/science/article/pii/0003682X9290046U
4. E. Wenzel, M. Arruda, D. Kistler, F. Wightman, Localization using nonindividualized head-related transfer functions. J. Acoust. Soc. Am. **94**, 111–23 (1993). https://doi.org/10.1121/1.407089
5. W. O. Brimijoin, A. W. Boyd, M. A. Akeroyd, The contribution of head movement to the externalization and internalization of sounds. PloS one. **8**(12), e83068 (2013). https://doi.org/10.1371/journal.pone.0083068
6. D.R. Begault, E.M. Wenzel, M.R. Anderson, Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. J. Audio Eng. Soc. **49**(10), 904–916 (2001)
7. S. Werner, F. Klein, T. Mayenfels, K. Brandenburg, in *2016 IEEE Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, A summary on acoustic room divergence and its effect on externalization of auditory events (IEEE, 2016)
8. 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** (2022). https://doi.org/10.1177/23312165221092919
9. T.J. Cox, F. Li, P. Darlington, Extracting room reverberation time from speech using artificial neural networks. J. Audio Eng. Soc. **49**(4), 219–230 (2001)

10. H. Löllmann, E. Yilmaz, M. Jeub, P. Vary, in *2010 IEEE Proceedings of international workshop on acoustic echo and noise control (IWAENC)*, An improved algorithm for blind reverberation time estimation (IEEE, 2010)
11. L. Treybig, S. Saini, S. Werner, U. Sloma, J. Peissig, in *Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference*, Room acoustic analysis and brir matching based on room acoustic measurements (Audio Engineering Society, 2022)
12. J. Eaton, N.D. Gaubitch, A.H. Moore, P.A. Naylor, in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, The ace challenge — corpus description and performance evaluation (IEEE, 2015)
13. S. Saini, J. Peissig, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Blind room acoustic parameters estimation using mobile audio transformer (2023). https://doi.org/10.1109/WASPAA58266.2023.10248186
14. M. Lee, J.H. Chang, in *2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, Blind estimation of reverberation time using deep neural network. https://doi.org/10.1109/ICNIDC.2016.7974586
15. H. Gamper, I.J. Tashev, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Blind reverberation time estimation using a convolutional neural network. pp. 136–140. https://doi.org/10.1109/IWAENC.2018.8521241
16. F. Xiong, S. Goetze, B. Kollmeier, B.T. Meyer, Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(2), 255–267 (2019). https://doi.org/10.1109/TASLP.2018.2877894
17. D. Looney, N.D. Gaubitch, in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Joint estimation of acoustic parameters from single-microphone speech observations. https://doi.org/10.1109/ICASSP40776.2020.9054532
18. N.J. Bryan, in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation. https://doi.org/10.1109/ICASSP40776.2020.9052970
19. P. Götz, C. Tuna, A. Walther, E.A.P. Habets, in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Blind reverberation time estimation in dynamic acoustic conditions. https://doi.org/10.1109/ICASSP43922.2022.9746457
20. S. Deng, W. Mack, E.A. Habets, in *Proc. Interspeech 2020*, Online Blind Reverberation Time Estimation Using CRNNs (2020), pp. 5061–5065. https://doi.org/10.21437/Interspeech.2020-2156
21. C. Ick, A. Mehrabi, W. Jin, in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Blind acoustic room parameter estimation using phase features. https://doi.org/10.1109/ICASSP49357.2023.10094848
22. P. Srivastava, A. Deleforge, E. Vincent, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Blind room parameter estimation using multiple multichannel speech recordings (2021). https://doi.org/10.1109/WASPAA52581.2021.9632778
23. EN ISO 3382-2:2008 - Acoustics - Measurement of room acoustic parameters - Part 2: Reverberation time in ordinary rooms (ISO 3382-2:2008)
24. 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
25. L.G. Marshall, An acoustics measurement program for evaluating auditoriums based on the early/late sound energy ratio. J. Acoust. Soc. Am. **96**(4), 2251–2261 (1994)
26. H. Gamper, in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, Blind c50 estimation from single-channel speech using a convolutional neural network. https://doi.org/10.1109/MMSP48831.2020.9287158
27. P. Callens, M. Cernak, Joint blind room acoustic characterization from speech and music signals using convolutional recurrent neural networks. (2020). https://arxiv.org/abs/2010.11167
28. P. Götz, C. Tuna, A. Walther, E.A.P. Habets, Online reverberation time and clarity estimation in dynamic acoustic conditions. J. Acoust. Soc. Am. **153**(6), 3532–3542 (2023). https://doi.org/10.1121/10.0019804
29. F. Klein, A. Neidhardt, M. Seipel, Real-time estimation of reverberation time for selection of suitable binaural room impulse responses (2019). https://doi.org/10.22032/dbt.39968
30. Z. Tang, N.J. Bryan, D. Li, T.R. Langlois, D. Manocha, Scene-aware audio rendering via deep acoustic analysis. IEEE Trans. Vis. Comput. Graph. **26**(5), 1991–2001 (2020). https://doi.org/10.1109/TVCG.2020.2973058
31. A. Ratnarajah, S. Ghosh, S. Kumar, P. Chiniya, D. Manocha, Av-rir: Audio-visual room impulse response estimation (2023). arXiv preprint arXiv:2312.00834
32. C.J. Steinmetz, V.K. Ithapu, P. Calamia, in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Filtered noise shaping for time domain room impulse response estimation from reverberant speech (IEEE, 2021)
33. A. Ratnarajah, S.X. Zhang, Y. Luo, D. Yu, M3-audiodec: Multi-channel multi-speaker multi-spatial audio codec (2023). arXiv preprint arXiv:2309.07416
34. P. Li, Y. Song, I. McLoughlin, W. Guo, L. Dai, An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition. Proc. Interspeech **2018**, 3087–3091 (2018). https://doi.org/10.21437/Interspeech.2018-1242
35. Y. Gong, Y.A. Chung, J. Glass, in *Proc. Interspeech 2021*, AST: Audio Spectrogram Transformer (2021), p. 571–575. https://doi.org/10.21437/Interspeech.2021-698
36. S. Werner, F. Klein, T. Mayenfels, K. Brandenburg, in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, A summary on acoustic room divergence and its effect on externalization of auditory events (2016), p. 1–6. https://doi.org/10.1109/QoMEX.2016.7498973
37. S. Werner, G. Götz, F. Klein, in *Audio Engineering Society Convention 142*, Influence of head tracking on the externalization of auditory events at divergence between synthesized and listening room using a binaural headphone system (Audio Engineering Society, 2017)
38. J. Blauert, *The technology of binaural listening* (Springer, Berlin, 2013)
39. D.T. Murphy, S. Shelley, *in Audio Engineering Society Convention 129,* Openair: an interactive auralization web resource and database (Audio Engineering Society, 2010)
40. I. Szöke, M. Skácel, L. Mošner, J. Paliesek, J. Černockỳ, Building and evaluation of a real room impulse response dataset. IEEE J. Sel. Top. Signal Process. **13**(4), 863–876 (2019)
41. G.J. Mysore, Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?–a dataset, insights, and challenges. IEEE Signal Process. Lett. **22**(8), 1006–1010 (2014)
42. C. Hopkins, S. Graetzer, G. Seiffert. *Aru speech corpus* (University of Liverpool, 2019). https://doi.org/10.17638/datacat.liverpool.ac.uk/681. https://datacat.liverpool.ac.uk/681/. Principal Investigator: Professor Carl Hopkins
43. P. Götz, C. Tuna, A. Walther, E.A. Habets, in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aid: Open-source anechoic interferer dataset (2022)
44. V. Panayotov, G. Chen, D. Povey, S. Khudanpur, in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Librispeech: an asr corpus based on public domain audio books (IEEE, 2015)
45. T. Hidaka, Y. Yamada, T. Nakagawa, A new definition of boundary point between early reflections and late reverberation in room impulse responses. J. Acoust. Soc. Am. **122**, 326–32 (2007). https://doi.org/10.1121/1.2743161
46. V. Garcia-Gomez, J.J. Lopez, in *Audio Engineering Society Convention 144*, Binaural room impulse responses interpolation for multimedia real-time applications (Audio Engineering Society, 2018)
47. V. Bruschi, S. Nobili, A. Terenzi, S. Cecchi, in *Audio Engineering Society Convention 152*, An improved approach for binaural room impulse responses interpolation in real environments (Audio Engineering Society, 2022)
48. F. Wefers, *Partitioned convolution algorithms for real-time auralization*, vol. 20 (Logos Verlag Berlin GmbH, Berlin, 2015)
49. T.d.M. Prego, A.A. de 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 (IEEE, 2015)
50. J. Yamagishi, C. Veaux, K. MacDonald, CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), University of Edinburgh. The Centre for Speech Technology Research (CSTR), (2019) Available: https://datashare.ed.ac.uk/handle/10283/3443
51. H.P. Seraphim, Untersuchungen über die unterschiedsschwelle exponentiellen abklingens von rauschbandimpulsen. Acta Acustica U. Acustica. **8**(4), 280–284 (1958)

52. J.S. Bradley, R. Reich, S. Norcross, A just noticeable difference in c50 for speech. Appl. Acoust. **58**(2), 99–108 (1999)
53. M. Blevins, A.T. Buck, Z. Peng, L. M. Wang, Quantifying the just noticeable difference of reverberation time with band-limited noise centered around 1000 Hz using a transformed up-down adaptive method (Proceedings of the International Symposium on Room Acoustics, Toronto, 2013).
54. International Telecom Union, Rec. ITU-R BS. 1534-1. Method for the subjective assessment of intermediate quality level of coding systems (2003)
55. International Telecom Union, Rec. ITU-R BS. 1534-3. Method for the subjective assessment of intermediate quality levels of coding systems (2015). https://www.itu.int/rec/R-REC-BS.1534
56. S.N. Wadekar, A. Chaurasia, Mobilevitv3: mobile-friendly vision transformer with simple and effective fusion of local, global and input features (2022). arXiv preprint arXiv:2209.15159

Publisher's Note