Audio Engineering Society

Conference Paper 33

Presented at the AES 5th International Conference on Audio for Virtual and Augmented Reality 2024 August 19-21, Redmond, WA, USA

This conference paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This conference paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Leveraging GANs for a better Blind Room Identification through Energy Decay Curve Estimation

Shivam Saini[1,2] and Jürgen Peissig[1]

[1]*Leibniz Universität Hannover, Institut für Kommunikationstechnik, Hanover, Germany*
[2]*Huawei Munich Research Center, Munich, Germany*

Correspondence should be addressed to Shivam Saini (shivam.saini@stud.uni-hannover.de)

ABSTRACT

This paper presents a novel approach for estimating the octave-band Energy Decay Curve (EDC) from noisy reverberant speech signals using Generative Adversarial Networks (GANs). Traditional methods of room acoustic analysis often rely on parameters such as Reverberation Time (RT) or Clarity (C50), or direct estimation of the Room Impulse Response (RIR). However, these approaches can be limited by their dependence on controlled experimental setups and/or detailed knowledge of the acoustic environment. By contrast, the proposed method leverages the power of GANs to directly estimate the sub-band EDCs from noisy reverberant speech signals, providing a more comprehensive characterization of the room's acoustic properties compared to the single-valued parameters. The proposed method offers several advantages over traditional approaches as the EDC provides a more detailed and holistic view of the room's acoustic behaviour compared to single-parameter metrics like RT as well as overcomes the artefacts that comes with generating RIRs directly. It provides a middle-ground that is more useful in synthesizing similar reverberation as of the real RIR. To evaluate the effectiveness of our approach, we conducted a series of experiments comparing the proposed GAN-based EDC estimation to state-of-the-art models on well-known benchmarks. The results demonstrate that our method not only achieves superior performance in terms of accuracy but also shows robustness to variations sub-band reverberation characteristics. The ability to estimate the EDC directly from speech signals without requiring a priori knowledge of the room highlights the practical applicability of our approach in real-world scenarios.

1 Introduction

In the field of acoustic engineering, understanding the behaviour of sound within spaces plays the most important role [1]. Room Impulse Response (RIR) estimation, particularly in the context of room acoustics, has gathered substantial attention in recent years due to its application in various fields such as source localization [2], speech dereverberation [3], echo cancellation [4], auralization in Audio Augmented Reality(AAR) [5, 6, 7], source separation [8], etc. In the context of Audio Augmented Reality (AAR), achieving accurate acoustic matching is particularly challenging [6, 9]. Rendering must closely replicate the listening room's acoustics to maintain the plausibility of the virtual sound source, preventing any disruption to the immersive auditory experience [10].

Researchers often tackle this problem with the acoustic measurements such as measuring RIR or Binaural RIR (BRIR) [11, 12, 13]. This is usually more complex at the user end of the application.

Hence, most researchers focus on estimating it from audio [14] or audio-visual signals [15]. Estimating RIR solely from audio is generally referred to blind RIR estimation. As the blind RIR estimation problem itself is complex, it is commonly broken into smaller problems such as the analysis of the environment and the synthesis of the RIR. The analysis could involve blindly estimating the physical parameters [16, 17] such as the room geometry, room volume and the reflection coefficients, or room acoustic parameters such as RT [18-26], C50 [19, 24, 27-30], Direct-to-Reverberant Ratio (DRR) [31], etc.

While estimating RIR directly sounds appealing, it may not be the best choice due to the complex structure of RIR which may produce artefacts in predicted RIRs when tested perceptually [7, 14]. Hence, the blind room acoustic parameter estimation is majorly focused in this field and handful of publications has demonstrated the possibility of estimating these parameters solely from noisy speech [18-26] or music signals [19, 30]. Although, these parameters may give an idea about the room acoustics, an exact RIR reconstruction may still not be possible from this information [7].

One significant parameter in this domain is the Energy Decay Curve (EDC), which is pivotal for characterizing room acoustics and is often overlooked. It provides a middle ground between an RIR and the room acoustic parameters which can generally be estimated from an EDC. Estimating EDC not only allows us to estimate parameters like RT60 [32], C50, etc., it also provides us with the information about the shape of the RIR itself. Further, estimating EDC in sub-bands allows us to understand the RIR structure in frequency domain as well. The EDC is calculated as follows and generally presented in dB:

$$EDC(t) \triangleq \int_{t}^{\infty} h^2(\tau) d\tau,$$

where $h(\tau)$ is the RIR and $EDC(t)$ is the total amount of signal energy remaining in the RIR at time t. RT60 can be measured from EDC using the backward integration method that can yield an EDC from a single measurement [32].

Traditionally, the process of estimating the EDC requires measuring an RIR hence demanding significant time, resources, and expertise. However, recent advancements in machine learning, particularly GANs, offer a promising avenue for revolutionizing this aspect of acoustic analysis. Building upon the success of GANs in various domains such as image synthesis [33, 34], natural language processing [35], and audio generation [36], our research aims to extend the utility of these generative models to the domain of room acoustics. By training GANs on a diverse dataset of audio recordings from various room configurations, reverberation times, and acoustic conditions, we aim to equip these models with the ability to generalize across a wide range of acoustic scenarios, thereby facilitating robust and accurate EDC estimation in real-world settings.

In summary, this research seeks to push the boundaries of acoustic analysis by utilizing the capabilities of GANs for Blind EDC Estimation. Through a systematic investigation and experimental validation, we aim to demonstrate the efficacy and practical utility of our proposed methodology in accurately inferring the EDCs from single-channel noisy reverberant speech signals, paving the way for an efficient room characterization in diverse acoustic environments.

2 Background

Historically, room identification majorly relied on RT estimation through direct measurement techniques, such as impulse responses. Early methods, like the Sabine [37] and Eyring [38] formulas, provided foundational insights into room acoustics based on volume and absorption coefficients. However, the limitations of these traditional methods led researchers to explore data-driven approaches.

With the rise of machine learning, large datasets enabled models to estimate reverberation parameters from audio recordings without prior environmental knowledge. Early models used regression techniques and neural networks to predict reverberation times from audio signal features. Convolutional Neural Networks (CNNs) [19, 20, 22] and Recurrent Neural Networks (RNNs) [18] showed significant promise in capturing the temporal and spatial characteristics of reverberated audio. CNNs effectively processed spectrograms and time-frequency representations, while RNNs excelled in modelling the sequential nature of audio data.

The ACE challenge [21, 39] in 2015 established a benchmark for evaluating RT and DRR estimation approaches, allowing fair comparison of models. Following this, several studies demonstrated the use of machine learning to estimate RT60 from noisy reverberant speech signals. Notably, Gamper and Tashev [22] used CNNs to predict RT60 using

Gammatone filtered spectra, outperforming previous methods. Looney and Gaubitch [29] achieved promising results in the joint blind estimation of RT60, DRR, and signal-to-noise ratio (SNR). Bryan [20] improved RT60 and DRR predictions by generating augmented training datasets from real RIRs. Ick et al. [40] introduced phase-related features, improving reverberation fingerprint estimation in unseen rooms. Götz et al. [30] extended previous work [23] to estimate sub-band RT60 and C50 under dynamic conditions using convolutional recurrent neural networks (CRNNs), showing improved performance even in static conditions. However, the model struggled with accurately predicting C50 above 15 dB and was only evaluated on noiseless data.

In acoustic characterization, CNNs are commonly used to solve blind acoustic parameter estimation problems as regression tasks, learning two-dimensional time-frequency signal patterns for end-to-end modelling. To capture long-range global context better, hybrid models combining CNNs with self-attention mechanisms have been proposed, achieving state-of-the-art results in various audio pattern recognition tasks. Recently, purely attention-based models like the Audio Spectrogram Transformer (AST) have been proposed for audio classification, attaining benchmark-leading results [41]. However, these models require substantial computational power due to their size. To address this, AudMobNet [24], a mobile audio transformer, was proposed, offering robustness against noise and much reduced computational expense.

Despite these advancements, many approaches only provide broadband parameter values, rather than frequency-dependent ones. Using a single T60 value to define room characteristics oversimplifies the room acoustics model, potentially limiting rendering realism in the context of AAR.

GANs, introduced by Goodfellow et al. [42], revolutionized the field of generative modelling by framing the problem as a game between two neural networks: the generator and the discriminator. The generator aims to produce realistic data samples, while the discriminator attempts to distinguish between real and generated samples. This adversarial process results in the generator producing increasingly realistic outputs.

Recently GANs have also been utilized in the field of room acoustics. Researchers have leveraged GANs for Augmenting RIRs [43], Speech Dereverberation [44, 45], and even generating RIRs from speech directly [14, 46]. However, these approaches often come with highly complex models and artefacts in generated RIR due to complex structure of reverberation patterns.

Blind EDC estimation presents unique challenges, yet a simpler solution than generating RIR itself. GANs, with their generative capabilities, offer a robust framework for blind EDC estimation, addressing the limitations of prior methods and paving the way for more accurate RIR analysis and synthesis.

3 Methodology

Our approach involves training a GAN on a diverse dataset of audio recordings that includes various room configurations, reverberation times, and levels of background noise. The GAN is designed to learn the complex relationship between reverberant speech signals and their corresponding EDCs, enabling it to accurately predict the EDC from new, unseen audio samples.

Dataset

We utilized various publicly available datasets of measured RIRs [47, 48, 49], resulting in a total of 3000 RIRs. An anechoic speech corpus [50, 51] and an anechoic noise dataset [52] were used to generate the final samples. Each RIR was convolved with 25 random anechoic speech signals, which were then combined with ambient noise. This noise was generated using the same process but with the noise dataset and Gaussian noise instead. Babble noise was simulated by convolving 30 different speech samples from another dataset [53] with the same RIR and adding this to the generated noise. To introduce diversity in the training data, a random signal-to-noise ratio (SNR) level between -3 and 20 dB was chosen when adding the noise. This process generated over 75,000 samples for the training set.

To obtain the sub-band EDCs, respective RIR is filtered with octave-band filter with centre frequencies at 125, 250, 500, 1000, 2000 and 4000 Hz. The EDC for each sub-band RIR is then concatenated to form a tensor of size 6 x 1000. The EDC calculation for each sub-band is performed using Schroeder's backward integration method, as described previously.

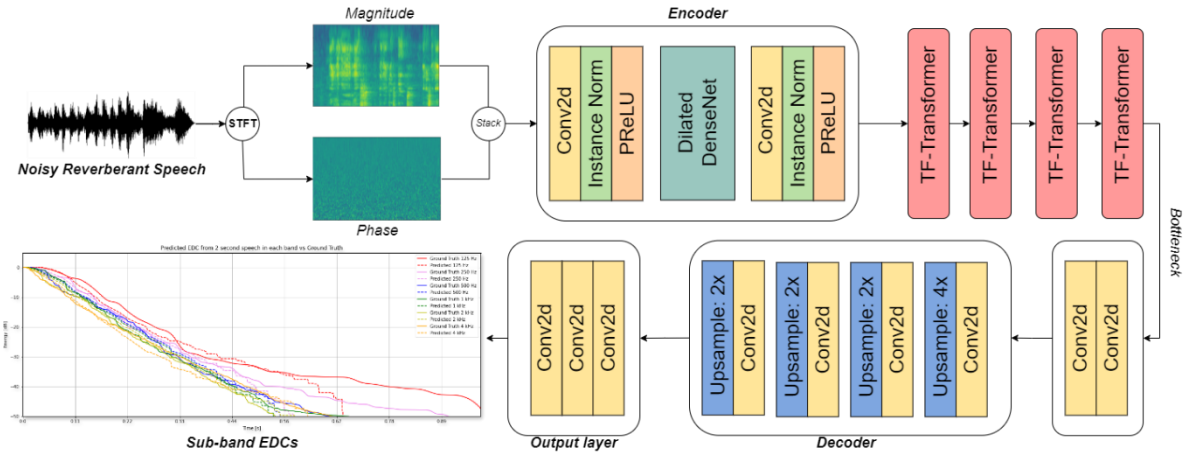


Figure 1: Proposed model architecture

Pre-processing

For the model input, magnitude and phase spectrograms were generated from the reverberant speech signal, producing a rectangular matrix of size 512 x L, where L denotes the length of the spectrum generated using a window size of 128 samples, a hop size of 64 samples, and 1024 FFT bins. To facilitate batch training, the input signals were initially trimmed or padded to a length of 4 seconds. These spectrograms were then fed into the network to produce six sub-band EDCs, each of length 1000.

Model

As mentioned earlier, we utilize a GAN-style generative model. Our generator model is based on an encoder-decoder style architecture. For the encoder, we employ the MP-SENet encoder [54] architecture (Figure 1: Proposed model architecture). The encoder transforms the input data and is followed by 4 two-stage conformer blocks also known as TS-Transformer blocks [55] (or TF-Transformer here), which are transformer-based blocks designed to process both time and frequency dimensions effectively.

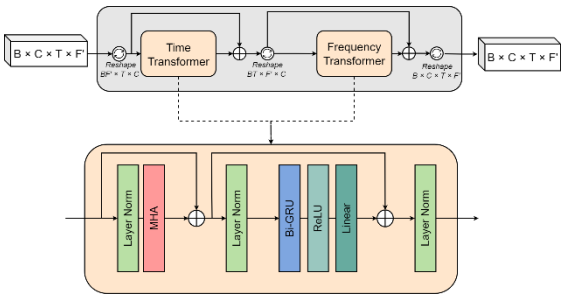


Figure 2: TF-Transformer block

The TF-Transformer block architecture (Figure 2), processes input tensors of shape Batch × Channel × Time × Frequency [B×C×T×F] by sequentially applying transformations in both time and frequency dimensions. Initially, the input tensor is reshaped to [B×F×T×C] for the time transformer. Within the time transformer, the input undergoes a series of operations including layer normalization to stabilize and accelerate the training process, followed by multi-head attention mechanisms that capture long-range dependencies along the temporal dimension. Additionally, a bidirectional GRU (Bi-GRU) layer processes the input in both forward and backward directions, enabling the model to understand context from both past and future time steps. Residual connections are integrated to add the input of each sub-layer to its output, facilitating the flow of gradients during training and improving convergence.

Subsequently, the tensor is reshaped again to [B×T×F×C] for the frequency transformer, which applies similar operations as the time transformer but focuses on the spectral dimension. This includes another round of layer normalization, multi-head attention, and Bi-GRU processing, with the goal of capturing dependencies across different frequency bands. The frequency transformer also employs residual connections to ensure efficient gradient flow and model convergence. After processing through both the time and frequency transformers, the output tensor is reshaped back to its original form, [B×C×T×F], ready for subsequent layers in the model. This dual transformer approach allows the TF-Transformer block to effectively model complex dependencies in both time and frequency domains, enhancing its ability to process and generate high-quality EDC features.

The extracted features are then passed to the bottleneck layers. The bottleneck layer of the encoder-decoder architecture has a specific size which serves as the latent representation of the input data. This latent representation is then passed through the decoder. The decoder is responsible for reconstructing the data from the latent space, gradually refining the details.

To further enhance the output resolution, the model incorporates 4 up-sampling layers. Each up-sampling layer doubles the resolution of the input through bilinear or bicubic interpolation, followed by a Conv2D layer to refine the features at each stage. Specifically, the first up-sampling layer scales the input by a factor of 4, while the subsequent layers scale by a factor of 2, each followed by a Conv2D operation to maintain and enhance the feature quality.

We try to avoid the checkerboard artefacts produced by deconvolution operation by using an up-sampling layer followed by a Conv2D layer instead. The final output layer consists of three Conv2D layers, which refine the up-sampled features to produce the final output. This final output layer is designed to generate the 6-sub band EDCs, each of the size 1000. This detailed architecture ensures that the model can accurately generate the EDCs across multiple sub-bands, capturing the intricate details required for accurate EDC synthesis. The description of each layer is further given in Table 1.

Generator	*Kernel Size*	*# channels*
Enc_Conv2D_1	1	2
DenseNet	3	64
Enc_Conv2D_2	3	64
Bottleneck_1	3	64
Bottleneck_2	3	32
Dec_Conv2D_1	3	64
Upsampling_1	4x	'bilinear'
Dec_Conv2D_2	3	32
Upsampling_2	2x	'bilinear'
Dec_Conv2D_3	3	28
Upsampling_3	2x	'bicubic'
Dec_Conv2D_4	3	28
Upsampling_4	2x	'bicubic'
Out_Conv2d_1	7	28
Out_Conv2d_2	5	12
Out_Conv2d_3	3	6

Table 1: Model layers description

We also utilize a simple discriminator model to distinguish between real and predicted EDCs. Our discriminator model is inspired by SE-Net and it provides us the difference between the pairs of real-real EDCs and the real-predicted EDCs.

Loss Function

In our proposed method, we employ multiple loss functions to address the complexity of the EDC. Specifically, we use an energy-decay loss to capture the energy-based acoustic properties of the EDC and a discriminator loss to enhance training. The energy-decay loss further consists of two components: MSE in the logarithmic scale and a monotonicity penalty. The MSE loss focuses on fine details, particularly in the lower energy regions, where precise matching is crucial for accurate acoustic modelling. The monotonicity penalty ensures that the decaying trend of the EDC remains consistent. The energy-decay loss is given as:

$$\mathcal{L}_{EDC} = \mathbb{E}\left\{\left(10 \cdot log_{10}\left(\frac{\hat{h}_{EDC,f}}{h_{EDC,f}}\right)\right)^2\right\} + \mathcal{L}_M,$$

where $\mathbb{E}\{\cdot\}$ denotes the expectation, $\hat{h}_{EDC,f}$ denotes the predicted EDC and $h_{EDC,f}$ denotes the ground truth EDC estimated from the RIR using Schroeder's backward integration method for each frequency band f. The monotonicity penalty ($\mathcal{L}_M$) is further defined as following:

$$\mathcal{L}_M = \sum_{f=0}^{F}\sum_{i=0}^{T-1} \max(\hat{h}_{EDC,f,i+1} - \hat{h}_{EDC,f,i}, 0),$$

where i is each sample and T is the length of the EDC. In our case, F is 6 and T is 1000. By combining these two losses, we achieve a balance between capturing the overall energy dynamics and the detailed structure of the EDC.

In addition to the energy-decay loss, we incorporate a discriminator loss within our generative adversarial framework. During training, the discriminator is fed pairs of real EDCs and pairs of real and predicted EDCs. The real EDC pairs are labelled with ones, indicating authenticity, while the real and predicted EDC pairs are labelled with zeros, indicating generated data. The discriminator's objective is to distinguish between these two types of pairs accurately. The loss for the discriminator is computed by comparing its output with the provided labels. Specifically, the loss for real EDC pairs is calculated using the MSE loss between the discriminator's output for these pairs and a tensor of ones. This approach allows the discriminator to learn to identify

the subtle differences between real and predicted EDCs, thereby improving the generative model's ability to produce realistic EDCs.

Post processing

For post-processing of the predicted EDC, we employ a two-step approach to ensure a smooth and physically accurate representation of the cumulative energy decay. Initially, cubic spline interpolation is applied to the EDC. This technique involves fitting a smooth, continuous curve through the discrete EDC data points, resulting in a full-scale, interpolated EDC that captures the underlying trend with high fidelity. Following the interpolation, we enforce the monotonic non-increasing property of the decay curve. The EDC, by definition, is a cumulative sum of energy, which should never increase over time; it either remains constant or decreases as energy dissipates. To ensure this, we adjust the interpolated EDC to correct any segments where the values might inadvertently increase due to neural network artefacts or interpolation errors. Similar to monotonicity loss, this adjustment involves scanning through the interpolated EDC and ensuring that each value is no greater than the preceding value, thereby preserving the physically realistic decay of energy.

4 Evaluation

In this chapter, we present a comprehensive assessment of our GAN-based model for blind EDC estimation, comparing its performance against state-of-the-art models in the similar field.

EDC Similarity: We assess the similarity between the estimated and true EDC curves using the Log EDR loss, defined as the logarithm of the mean squared error between the predicted and ground truth EDC [46]. Table 2 compares the Log-EDR loss of the proposed model with that of the SOTA models. Both the models are trained to estimate/generate RIR from reverberant speech signals, and the EDC is calculated from the estimated RIR. To ensure a fair comparison, a new unseen dataset was generated using the VCTK Speech Corpus [57] and ACE evaluation set RIRs.

Method	Frequency Band					
	125	*250*	*500*	*1k*	*2k*	*4k*
Ours	**-5.32**	**-6.01**	**-5.92**	**-5.70**	**-5.48**	**-5.23**
FiNS	-4.17	-3.00	-3.64	-3.16	-3.00	-1.48
S2IR	-4.70	-3.42	-4.23	-3.85	-3.66	-2.08

Table 2: Performance analysis of proposed model, FiNS [14] and S2IR [46] on *ACE * VCTK* set.

The results presented in Table 2 demonstrate that the proposed method significantly outperforms the SOTA methods, especially in high-frequency regions. We believe this is due to the less complex structure of the EDC compared to the RIR in high frequencies, making it easier for the model to capture. Furthermore, since the model is trained to predict the logarithmic EDC, it accurately predicts the low-energy regions of the RIR, which are dominant in high-frequencies. In addition to that, predicting RIR first and calculating EDC from the predictions may have introduced further errors in the SOTA models.

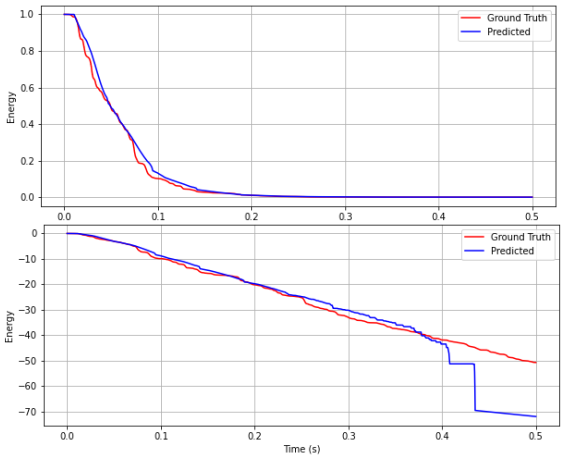


Figure 3: Predicted EDC and Ground Truth in linear (top) and dB (bottom) scale.

It shall be noted as shown in Figure 3, extreme errors are introduced in low-energy regions below -60 dB. We attribute this to the noise floor of the input spectrum, which the model cannot accurately learn. Additionally, this region is not as critical for synthesizing an RIR. Therefore, we trim all energy below -60 dB when calculating the EDR loss. The energy in this region can also be extrapolated from the available information, but we leave this for future work.

Model	$RMSE_{RT}$	ρ_{RT}
QAReverb [31]	0.255	0.77
JROPE [28]	0.288	0.76
FiNS [14]	0.237	0.32
SG-RIR [56]	0.176	0.84
Ours	**0.010**	**0.89**

Table 3: RT60 estimation performance comparison of models presented in ACE challenge, RIR estimators and proposed model.

RT60 Estimation: To understand the model better we further predict RT60 from the EDC and evaluate using the ACE challenge's evaluation metrics: ρ, MSE, and Bias, between the estimations derived from the predicted and the real EDCs. These metrics provide a robust framework for benchmarking the performance of acoustic models against the SOTA models. The RT is derived by calculating the slope of EDC over different ranges and finding the best fit. This best fit is then chosen to extrapolate the value of T60. It is to be noted, in order to avoid any conflict in ground truth and predictions when evaluating, the ground truth for ACE evaluation set was also generated using the same method.

Table 3 shows the comparison of full-band RT estimation against state of the art models. The models on top of the table are the best performing models from the ACE challenge. The models in lower half focus of estimating the RIR directly from reverberant speech. The results show us that the proposed model outperform all existing SOTA models in RT estimation affirming to our hypothesis of a better understanding of room acoustics.

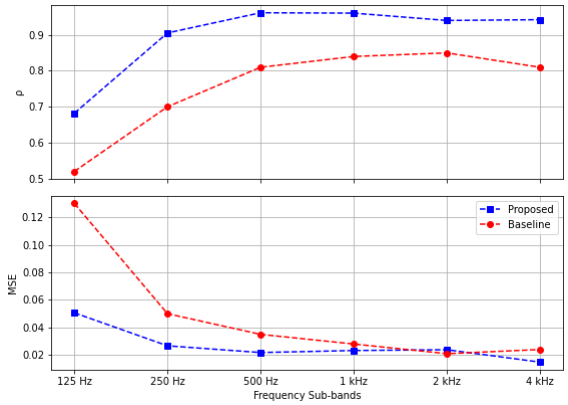


Figure 4: Sub-band RT60 estimation performance comparison against baseline model [24].

For a comprehensive evaluation of sub-band RT60 estimation accuracy, we compare our model against the state-of-the-art sub-band RT estimation model [24] on the ACE challenge evaluation set. As shown in Figure 4, the proposed outperforms the recently proposed AudMobNet in terms of sub-band RT60 estimation. Our GAN model excels in predicting lower frequencies as it learns to predict the EDC directly, rather than a single RT60 value, which can fluctuate significantly even in the ground truth due to the inherent computation involved. Additionally, the model input includes phase features which further assist the model in predicting low frequencies as shown before in [7].

Figure 5 shows us the proposed GAN often tends to underestimate the RT in each sub-band. This underestimation may be attributed to the monotonicity penalty loss and the enforcement of monotonicity during post-processing, which do not account for fluctuations. Implementing smoothing techniques in future, such as moving averages, may improve overall estimation accuracy.

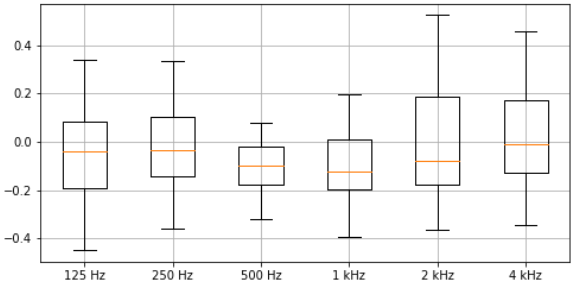


Figure 5: Estimation bias in RT calculation.

It is important to note that although both the baseline and the proposed models have the same number of parameters, the proposed model exhibits a lower real-time factor (RTF) due to a higher number of floating-point operations (FLOPs), resulting in a substantially slower training process (at least 100x slower). This increased computational complexity is due to the multiple Bi-directional GRU and Multi-Headed Attention layers present in the TF-Transformers.

5 Conclusions

In this study, we presented a novel approach for estimating the EDC from noisy speech signals using GANs. Traditional methods for room acoustic analysis, which rely on detailed measurements, are often impractical for end-users. Blind RIR estimation techniques offer a solution but are typically complex and prone to artefacts when directly estimating RIRs. By focusing on EDC estimation, our method provides a balanced alternative, capturing essential room acoustic parameters such as RT without the drawbacks of direct RIR synthesis.

Our GAN-based model processes 4-second noisy speech inputs, transforming them into down-sampled EDCs across 6 sub-bands. With an encoder-decoder architecture and 2.3 million trainable parameters, the model efficiently extracts and refines acoustic features. The results demonstrated high accuracy for RT estimation, with an MSE of 0.01 and a correlation

of 0.89 on the validation set, showcasing the model's effectiveness in diverse real-world scenarios. Future works may include exploring different inputs such as mel-spectrums, a faster generator model, implementation of EDC smoothening techniques, extrapolation of low energy regions and the synthesis of RIR from the predicted EDC.

6 References

[1] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, pp. 171-218, 1992.

[2] C. Evers, H. W. Löllmann, H. Mellmann, A. Schmidt, H. Barfuss, P. A. Naylor and W. Kellermann, "The LOCATA challenge: Acoustic source localization and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, p. 1620–1643, 2020.

[3] P. A. Naylor and N. D. Gaubitch, Speech dereverberation, Springer Science & Business Media, 2010.

[4] Y. Lin and D. D. Lee, "Bayesian regularization and nonnegative deconvolution for room impulse response estimation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 839-847, 2006.

[5] J.-M. Jot and K. S. Lee, "Augmented reality headphone environment rendering," in *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*, 2016.

[6] R. Gupta, J. He, R. Ranjan, W.-S. Gan, F. Klein, C. Schneiderwind, A. Neidhardt, K. Brandenburg and V. Välimäki, "Augmented/Mixed Reality Audio for Hearables: Sensing, control, and rendering," *IEEE Signal Processing Magazine*, vol. 39, pp. 63-89, 2022.

[7] S. Saini, I. Engel and J. Peissig, "An end-to-end approach for blindly rendering a virtual sound source in an audio augmented reality environment," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, p. 16, 2024.

[8] S. Gannot, E. Vincent, S. Markovich-Golan and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, p. 692–730, 2017.

[9] A. Neidhardt, C. Schneiderwind and F. Klein, "Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical Framework," *Trends in Hearing*, vol. 26, 2022.

[10] S. Werner, F. Klein, T. Mayenfels and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[11] J. Blauert, The technology of binaural listening, Berlin: Springer, 2013.

[12] B. G. Shinn-Cunningham, N. Kopco and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *The Journal of the Acoustical Society of America*, vol. 117, p. 3100–3115, 2005.

[13] L. Treybig, S. Saini, S. Werner, U. Sloma and J. Peissig, "Room acoustic analysis and BRIR matching based on room acoustic measurements," in *Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference*.

[14] C. J. Steinmetz, V. K. Ithapu and P. Calamia, "Filtered Noise Shaping for Time Domain Room Impulse Response Estimation From Reverberant Speech," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

[15] A. Ratnarajah, S. Ghosh, S. Kumar, P. Chiniya and D. Manocha, "AV-RIR: Audio-Visual Room Impulse Response Estimation," *arXiv preprint arXiv:2312.00834*, 2023.

[16] P. Srivastava, A. Deleforge and E. Vincent, "Blind Room Parameter Estimation Using Multiple Multichannel Speech Recordings," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

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

[18] S. Deng, W. Mack and E. A. P. Habets, "Online Blind Reverberation Time Estimation Using CRNNs," in *Proc. Interspeech 2020*, 2020.

[19] P. Callens and M. Cernak, "Joint blind room acoustic characterization from speech and music signals using convolutional recurrent neural networks," *CoRR*, vol. abs/2010.11167, 2020.

[20] N. J. Bryan, "Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[21] J. Eaton, N. D. Gaubitch, A. H. Moore and P. A. Naylor, "Estimation of Room Acoustic Parameters: The ACE Challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 1681-1693, 2016.

[22] H. Gamper and I. J. Tashev, "Blind Reverberation Time Estimation Using a Convolutional Neural Network," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*.

[23] P. Götz, C. Tuna, A. Walther and E. A. P. Habets, "Blind Reverberation Time Estimation in Dynamic Acoustic Conditions," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[24] S. Saini and J. Peissig, "Blind Room Acoustic Parameters Estimation Using Mobile Audio Transformer," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

[25] M. Lee and J.-H. Chang, "Blind estimation of reverberation time using deep neural network," in *2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*.

[26] H. Löllmann, E. Yilmaz, M. Jeub and P. Vary, "An improved algorithm for blind reverberation time estimation," in *2010 IEEE Proceedings of international workshop on acoustic echo and noise control (IWAENC)*.

[27] H. Gamper, "Blind C50 estimation from single-channel speech using a convolutional neural network," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*.

[28] F. Xiong, S. Goetze, B. Kollmeier and B. T. Meyer, "Joint Estimation of Reverberation Time and Early-To-Late Reverberation Ratio From Single-Channel Speech Signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 255-267, 2019.

[29] D. Looney and N. D. Gaubitch, "Joint Estimation Of Acoustic Parameters From Single-Microphone Speech Observations," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[30] P. Götz, C. Tuna, A. Walther and E. A. P. Habets, "Online reverberation time and clarity estimation in dynamic acoustic conditions," *The Journal of the Acoustical Society of America*, vol. 153, pp. 3532-3542, June 2023.

[31] T. d. M. Prego, A. A. de Lima, R. Zambrano-López and S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *2015 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, 2015.

[32] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, p. 1187–1188, 1965.

[33] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*, 2016.

[34] A. Brock, J. Donahue and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[35] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen and L. Carin, "Adversarial feature matching for text generation," in *International conference on machine learning*, 2017.

[36] J. Kong, J. Kim and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, p. 17022–17033, 2020.

[37] W. C. Sabine, "Architectural acoustics," in *Proceedings of the American Academy of Arts and Sciences*, 1906.

[38] C. F. Eyring, "Reverberation time in “dead” rooms," *The Journal of the Acoustical Society of America*, vol. 1, p. 217–241, 1930.

[39] J. Eaton, N. D. Gaubitch, A. H. Moore and P. A. Naylor, "The ACE challenge — Corpus description and performance evaluation," in *2015 IEEE Workshop on Applications of*

Signal Processing to Audio and Acoustics (WASPAA).

[40] C. Ick, A. Mehrabi and W. Jin, "Blind Acoustic Room Parameter Estimation Using Phase Features," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[41] Y. Gong, Y.-A. Chung and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021.

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, p. 139–144, 2020.

[43] A. Ratnarajah, Z. Tang and D. Manocha, "IR-GAN: Room impulse response generator for far-field speech recognition," *arXiv preprint arXiv:2010.13219*, 2020.

[44] J. Su, Z. Jin and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *arXiv preprint arXiv:2006.05694*, 2020.

[45] J. Su, Z. Jin and A. Finkelstein, "HiFi-GAN-2: Studio-Quality Speech Enhancement via Generative Adversarial Networks Conditioned on Acoustic Features," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

[46] A. Ratnarajah, I. Ananthabhotla, V. K. Ithapu, P. Hoffmann, D. Manocha and P. Calamia, "Towards improved room impulse response estimation for speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[47] I. Szöke, M. Skácel, L. Mošner, J. Paliesek and J. Černockỳ, "Building and evaluation of a real room impulse response dataset," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, p. 863–876, 2019.

[48] D. T. Murphy and S. Shelley, "OpenAIR: An Interactive Auralization Web Resource and Database," in *Audio Engineering Society Convention 129*, 2010.

[49] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, p. E7856–E7865, 2016.

[50] C. Hopkins, S. Graetzer and G. Seiffert, *ARU speech corpus (University of Liverpool)*, University of Liverpool, 2019.

[51] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*.

[52] P. Götz, C. Tuna, A. Walther and E. A. P. Habets, "Aid: Open-Source Anechoic Interferer Dataset," in *2022 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*.

[53] 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 Processing Letters*, vol. 22, p. 1006–1010, 2014.

[54] Y.-X. Lu, Y. Ai and Z.-H. Ling, "MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra," *arXiv preprint arXiv:2305.13686*, 2023.

[55] R. Cao, S. Abdulatif and B. Yang, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Interspeech 2022*, 2022.

[56] Z. Liao, F. Xiong, J. Luo, M. Cai, E. S. Chng, J. Feng and X. Zhong, "Blind Estimation of Room Impulse Response from Monaural Reverberant Speech with Segmental Generative Neural Network," 2023.

[57] J. Yamagishi, C. Veaux and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," 2019.