BLIND REVERBERATION TIME ESTIMATION IN DYNAMIC ACOUSTIC CONDITIONS

Philipp Götz[1], Cagdas Tuna[2], Andreas Walther[2], Emanuël A. P. Habets[1]

[1]International Audio Laboratories Erlangen[†], Germany.
[2]Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany.

ABSTRACT

The estimation of reverberation time from real-world signals plays a central role in a wide range of applications. In many scenarios, acoustic conditions change over time which in turn requires the estimate to be updated continuously. Previously proposed methods involving deep neural networks were mostly designed and tested under the assumption of static acoustic conditions. In this work, we show that these approaches can perform poorly in dynamically evolving acoustic environments. Motivated by a recent trend towards data-centric approaches in machine learning, we propose a novel way of generating training data and demonstrate, using an existing deep neural network architecture, the considerable improvement in the ability to follow temporal changes in reverberation time.

Index Terms— Dynamic acoustic conditions, Data-centric AI, reverberation time estimation, convolutional recurrent neural networks

1. INTRODUCTION

The time it takes the acoustic energy in a steady-state reverberant sound field to decay by $60\,\mathrm{dB}$, referred to as reverberation time RT_{60} in seconds, is a parameter of central importance in the description of many acoustic environments. It is defined by the process of propagation, reflection, and absorption of acoustic waves throughout an enclosure and across its boundaries and depends on geometric and acoustic properties of the volume or space it was measured in. Though in principle not limited to, it is mainly used as a measure for architectural spaces (e.g. rooms, halls, staircases).

In theory, RT_{60} is position independent within an enclosed volume. However, especially at wavelengths comparable to the dimensions of the enclosure, variations are often observed in real-world measurements and require averaging across multiple transmission paths throughout the volume [1]. Another critical assumption of measuring the reverberation time – either based on the direct measurement of the energy decay curve or via the integrated acoustic impulse response (AIR) – is constant acoustic conditions and, traditionally, the use of dedicated measurement signals. Once conditions change, e.g., when a room door is opened or a sound source or receiver moves to a different position, a renewed RT_{60} measurement is necessary. There are scenarios in which it is desired to continuously estimate the reverberation time from real-world signals and require the possibility to adapt to changes in acoustic conditions. Such applications pose the challenge of continuously extracting patterns of temporally decaying acoustic energy from signals of great variety and estimating the decay rate of the dynamic acoustic environments in which they were acquired.

With a wide range of potential applications in areas such as acoustic scene analysis [2], speech enhancement [3], extended reality (VR/AR) [4], acoustic environment classification [5] or audio forensics [6], the blind estimation of RT_{60} from real-world signals has seen increased interest in the past. Prior to leveraging the pattern recognition and non-linear approximation capabilities of deep neural networks (DNNs) in the recent past, traditional statistical approaches employing signal decay rate distributions and maximum likelihood estimation were used to estimate the reverberation time from noisy signals blindly [7–11]. While many of these approaches produced very useful results, the non-linear approximation capabilities of DNNs pushed the state-of-the-art even further. Convolutional neural networks (CNN) were used to estimate RT_{60} and Speech Transmission Index (STI) from spectro-temporal representations of noisy, reverberant speech [12, 13]. CNNs were extended by a recurrent layer to form convolutional recurrent neural networks (CRNN) that exploit sequential dependencies in the data [14, 15] and further improve estimation accuracy [16]. The focus of other studies was on auditory aspects of reverberation, with attempts to match artificial reverberation with perceptual characteristics of a real recording [17] and the blind estimation of AIRs from reverberant speech [18].

For approaches that involve DNNs for the estimation of RT_{60} from reverberant speech, a method of generating realistic training data is essential. In recent studies [12, 16], the reverberant speech was generated by convolving short segments of anechoic speech with measured or simulated AIRs with the assumption of static acoustic conditions, meaning that the acoustic transmission path between source and receiver captured in the AIR remained constant over time. However, real-world sound fields are often subject to temporally changing conditions, and while the solution proposed by [16] is targeted towards online RT_{60} estimation, the authors did not investigate the performance in dynamic acoustic conditions. The results presented in this work show that the system's performance is limited under such conditions, at least using their proposed training scheme. In this contribution, we aim to overcome this limitation and hence improve the estimation accuracy in dynamically changing acoustic environments. Motivated by a recent shift from a model-centric towards a data-centric approach to various problems in machine learning, we propose a novel way of generating training data and investigate its effect on the ability of a DNN to follow temporal changes in acoustic conditions.

The remainder of the paper is organized as follows. In Section 2, we review the CRNN architecture that is used in this study, Section 3 covers the different data generation methods that are proposed. In Section 4, the estimation performance resulting from the different training methods is evaluated, Section 5 concludes this paper.

[†]International Audio Laboratories Erlangen is a joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS.
Corresponding author: philipp.goetz@audiolabs-erlangen.de.

2. MODEL REVIEW

As this contribution focuses on dynamic reverberant sound fields, we choose a network architecture suitable to provide short-time estimates of the reverberation time and adopt the CRNN proposed in [16]. In the following paragraphs, the input data format, the model architecture and the relationship between temporal resolutions of the model input and output are discussed.

2.1. Model input

Using a gammatone filterbank [19] of 21 bands, with the center frequency of the lowest band set to 100 Hz, the reverberant speech signal sampled at a rate of 16 kHz is transformed into a time-frequency representation used in the training, validation, and testing process. Each gammatone spectrum is approximated by linear transformation of the short-time Fourier transform of the input signal [20], gammatone spectrograms are computed with a window length and hop length of $t_w = 4\,\text{ms}$ and $t_h = 2\,\text{ms}$ (frame overlap of $50\,\%$), respectively.

2.2. Model architecture

In the initial CNN section of the model, time-frequency features are extracted from the input gammatone spectrogram across six convolutional layers. Each convolution is followed by a rectified linear unit function (ReLU) and batch normalization across the five convolutional channels. The encoder structure is followed by dropout regularization that helps to avoid overfitting and a reorganization of the dimensions before the information flows through a recurrent layer (LSTM) with a cell state size and a hidden state size of 20 neurons each. Finally, after downsampling along the frequency dimension by a factor of two using Max Pooling, a fully connected layer followed by a ReLU function yields the estimate for each time step. The model has a total of 5611 trainable parameters, hyper-parameter optimization [21] was carried out to determine learning rate, batch size, and dropout probability. For further details, we refer the reader to [16].

2.3. Receptive field

It is crucial to investigate the temporal relationship between model input and output with a potential application in real-time scenarios. The CNN encoder structure at the beginning of the model contains strided convolutions across six layers that result in a significant amount of temporal compression of the input data. The total receptive field r_0 of this encoder structure may be computed by [22]:

$$r_0 = \sum_{l=1}^{L} \left((k_l - 1) \prod_{i=1}^{l-1} s_i \right) + 1, \tag{1}$$

where L is the number of layers that define the receptive field, k_l and s_i are the kernel size and stride of the lth and i^{th} layer, respectively. Hence, a signal length of $t_s = t_h(r_0 - 1) + t_w = 208\,\text{ms}$ is taken into account for each short-time estimate of RT_{60}, according to the parametrization of the CNN section proposed in [16] ($r_0 = 103$). Following the stride specifications of the model and the hop length t_h of the gammatone spectrogram computation, a new estimate is computed every $64\,\text{ms}$. Such a temporal resolution allows tracking RT_{60} in scenarios in which sudden changes are expected. In comparison, the previously proposed CNN-based model by Gamper and Tashev [12] produces a single estimate for a signal length of four seconds.
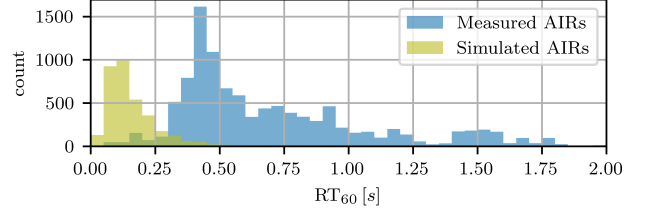

Fig. 1. Distribution of RT_{60} in measured and simulated AIRs.

3. DATA GENERATION

3.1. Reverberant data

As outlined in Sec. 2.1, the reverberant speech used in this contribution is transformed into a time-frequency representation using a gammatone filterbank. To accelerate the training process and to achieve convergence, each gammatone spectrogram is standardized [23], i.e. each sample is made zero-mean and scaled such that its standard deviation is equal to one. Furthermore, spectral whitening is applied to exclude non-linear magnitude distributions from the features learned by the neural network. In accordance with [12] and [16], the ground truth RT_{60} used during regression is obtained using the method proposed by Karjalainen et al. [24].

Three anechoic speech data sets were used to generate the data in this study: LibriSpeech ASR Corpus [25], TIMIT Acoustic-Phonetic Continuous Speech Corpus [26] and PAVOQUE Text-to-speech Corpus [27]. Measured AIRs were taken from the ACE challenge data set [5], the IKS Aachen Impulse Response database [28], the OpenAir database [29] and the EchoThief database [30]. AIRs with a reverberation time exceeding two seconds were excluded, leaving a total of 930 measured responses with RT_{60} ranging between 0.09 s and 1.95 s. In addition to measured data, the image method [31] was used to simulate 2000 AIRs in randomly generated enclosures with RT_{60} ranging between 0.01 s and 0.83 s. The entire data was split into training and validation data by 80% and 20%, test data for the evaluation of all four models were generated separately, as outlined in Sec. 3.4. The ratio between measured and simulated AIRs in training and validation data sets was $4 : 1$. The distribution of RT_{60} in the training data is shown in Fig. 1.

3.2. Static and dynamic training data

Samples of fixed duration are generated that represent reverberant speech under both static and dynamic acoustic conditions. In the static case, a single AIR is used to reverberate the anechoic speech sample, whereas in the dynamic case, two different AIRs are used, with a change occurring after a predefined duration. We investigate four different methods of generating reverberant speech samples, which are used to train, validate and test the model:

- *Static conditions (4 sec.)*: Segments of anechoic speech with a duration of four seconds, convolved with a single AIR. This method was used in [12] and [16] and serves as a baseline.

- *Static conditions (2 sec.)*: Segments of anechoic speech with a duration of two seconds, convolved with a single AIR. The aim is to prevent the model from learning temporal dependencies of more than two seconds.

- *Dynamic conditions (deterministic)*: Samples of reverberant speech with a total duration of four seconds are generated, the AIR used to reverberate the anechoic speech is exchanged for a different one after two seconds.
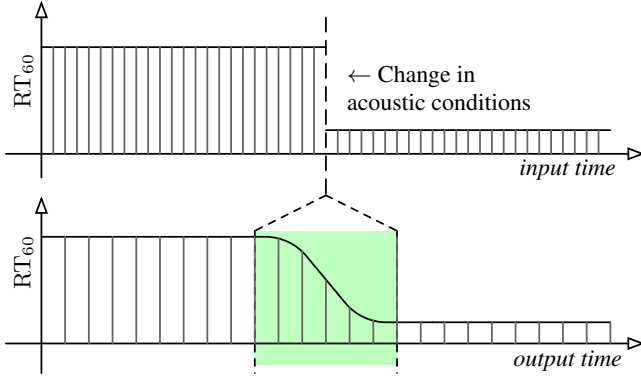
Fig. 2. Illustration of the ground truth interpolation: At the top, the ground truth at the input time scale is shown where the dashed line marks the moment of change in acoustic conditions. The bottom plot shows the ground truth at the output time scale. The green area indicates the segment of output frames, in which the moment of transition is located within the input frames. The smooth transition is a result of the six consecutive convolutional layers in the model.

- *Dynamic conditions (random)*: Samples of reverberant speech with a total duration of four seconds are generated, the AIR is switched at a random moment chosen between 0.8 and 3.2 seconds. This method is used to rule out the possibility that the model expects the change in conditions always after two seconds, as could be an effect of the deterministic dynamic training condition.

An efficient way of increasing training data is by recombining anechoic speech segments and AIRs to generate new, essentially unseen data. In this study, four-fold recombination was used, resulting in approximately 5252 hrs. of reverberant speech for training, validation, and test data for each of the four conditions. The loss function used to determine the weight update (RMSprop, adaptive learning rate optimization algorithm [32]) is the mean squared error (MSE) across all time steps in each mini-batch. The change in acoustic conditions occurs between neighboring time frames in each input sample. As a result of this, and due to the temporal compression of the convolutional layers of the model, the ground truth vector is smoothed following the kernel size and stride specifications along the temporal dimension of the CNN encoder of the model. An illustrative example is shown in Fig. 2.

3.3. Example

An example is shown in Fig. 3 for a single instance to illustrate the temporal behavior of the model resulting from the four different training methods. It is observed that the statically trained models cannot track the transition accurately and only react slowly to the changed acoustic conditions, with the statically trained model using two-second segments exhibiting the most stagnant update. The temporal interpolation of the ground truth RT_{60} on the output time scale, mentioned in Sec. 3.2, can also be seen for this example.

3.4. Evaluation data

Once all models are trained with the data generated according to the four conditions, an evaluation is performed on two separate test
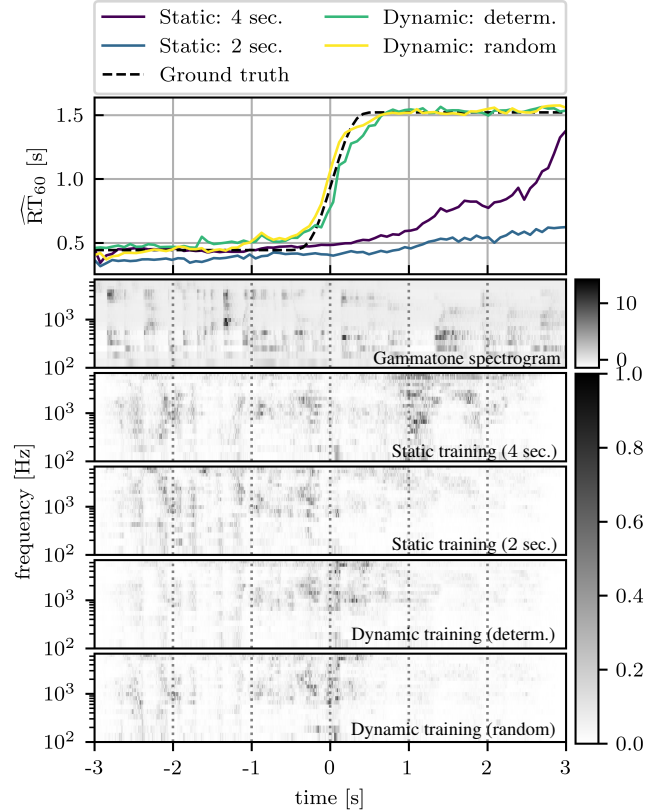
Fig. 3. The top plot shows the estimation of the four differently trained networks on a single sample. The input gammatone spectrogram and a comparison of four saliency maps [33] resulting from the four differently trained models are shown below. RT_{60} changes from $0.44\,\mathrm{s}$ in the first half of the sample to $1.52\,\mathrm{s}$ in the second half, the switch occurs at $t = 0\,\mathrm{s}$.

sets. For one, a dynamic data set is generated that contains reverberant speech segments with a duration of six seconds, including a change in acoustic conditions after three seconds. This presents all of the differently trained models with dynamic data of unseen length while the fixed change in acoustic conditions after three seconds enables convenient averaging across samples. Additionally, a second test data set representing static acoustic conditions with a duration of ten seconds is generated to investigate if the increased temporal adaptability resulting from dynamic training comes at the cost of reduced accuracy under static conditions. Both test sets use all available source data by generating unique pairings of anechoic speech and AIRs for the four individual training sets.

4. PERFORMANCE EVALUATION

4.1. Metrics

Table 1 provides an overview of various metrics that reflect the performance of the differently trained models under dynamic and static acoustic conditions. In the dynamic case, mean μ and standard deviation σ of the squared error at all time steps in each sample are averaged over the entire test set and are indicators of the overall estimation accuracy and temporal reactivity of the model. Another aspect is the estimation accuracy at the last time step in each sample,

	DYNAMIC TEST DATA						STATIC TEST DATA			
	μ	σ	MSE	ρ	Bias	MAPE	MSE	ρ	Bias	MAPE
Static training (4 sec.)	0.2185	0.1386	0.1216	0.4881	0.0815	50.89	**0.0219**	**0.9308**	−0.0369	18.86
Static training (2 sec).	0.2225	0.1457	0.1502	0.4948	0.1922	68.08	0.0347	0.8767	0.044	23.57
Dynamic training (determ.)	0.1773	0.1153	0.0637	0.7677	−0.0632	26.87	0.0278	0.8993	**−0.0197**	19.88
Dynamic training (random)	**0.1551**	**0.1018**	**0.0509**	**0.8136**	−0.0406	**23.86**	0.0268	0.9069	−0.0365	**18.69**

Table 1. Overview over various metrics describing the performance of the model trained with four different data sets, evaluated on dynamic and static test data.

after the change in acoustic conditions has occurred and the model processed the entire sample. For both static and dynamic conditions, the last step's performance is evaluated based on four different metrics: MSE, Pearson correlation coefficient ρ, bias and mean absolute percentage error (MAPE).

Under dynamic conditions, the two networks trained with dynamic samples exhibit consistently better metrics with a reduction in MSE by over 50% and an improvement of the correlation coefficient by approximately 65% when compared to the static training using four seconds. At the same time, mean and standard deviation are also reduced by approximately 30%. Under static acoustic conditions, the dynamically trained models perform only slightly less accurately.

4.2. Temporal evaluation

In addition to the analysis of performance metrics, three aspects that describe the temporal behaviour of the model are investigated qualitatively. The temporal characteristics of the estimation error response to changing acoustic conditions, the degree of renewed network activation based on frequency-averaged saliency maps [33] and the importance attribution of new information following the acoustic change based on frequency-averaged integrated gradients [34]. Fig. 4, top, shows the MSE response for the four training conditions, averaged over the entire test data set and separated by the direction of change in RT_{60}. The difference in temporal behaviour is noticeable: when a transition from dry to reverberant acoustic conditions occurs, the statically trained models are only able to gradually update their estimation, while the dynamically trained models are quick to react. Furthermore, the dynamically trained models show improved accuracy with a lower average MSE before the change in conditions for reverberant environments.

Another aspect is the magnitude of the frequency averaged, normalized gradients resulting from the four differently trained networks. The statically trained models do not significantly react to the change in conditions and can only update the estimation gradually. The dynamically trained network's ability to track the acoustic change is indicated by the renewed activation of neurons concentrated around the moment of transition in Fig. 3, with the *dynamic (random)* training condition exhibiting the largest gradients. Reflected by the average integrated gradients in Fig. 4, the two statically trained models only attribute importance to the beginning of each sample while later parts are taken into account to a lesser degree. In contrast, the models trained with dynamic samples continuously consider new information with a slight emphasis on the moment of changing acoustic conditions.

One interesting insight arises from the comparison of training conditions *Static (2 sec.)* and *Dynamic (deterministic)*. While the total number of pairings of anechoic speech and AIRs seen by the model during training is equal in both conditions, the dynamically
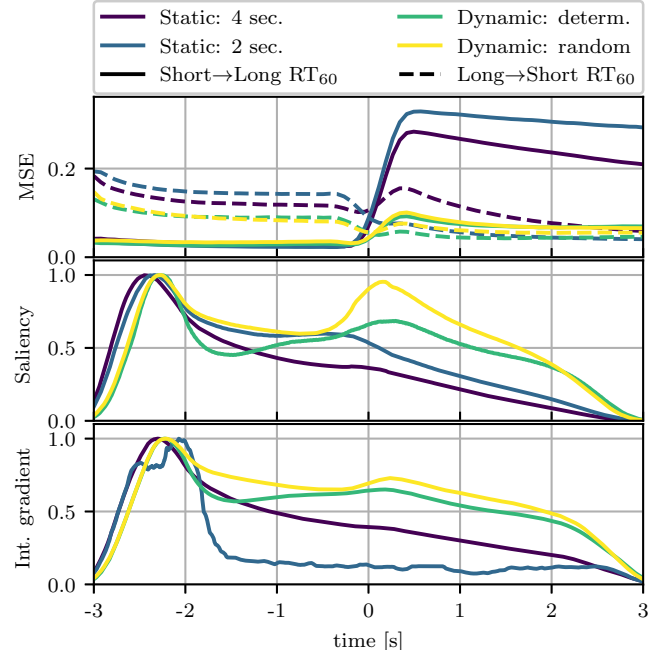

Fig. 4. Three plots describing the temporal behaviour of the four trained networks, the change in acoustic conditions occurs at $t = 0$. Top: Average squared error separated by the direction of change in RT_{60}; Middle: frequency-averaged saliency; Bottom: feature importance attribution based on frequency-averaged integrated gradients.

trained model performs considerably more accurately when acoustic conditions change. This observation supports the hypothesis that a portion of the convolutional features and recurrent modeling learned by the dynamically trained network relates to the detection of and reaction to changing acoustic conditions.

5. CONCLUSION

We investigated a previously proposed neural network [16] for blind RT_{60} estimation under dynamic acoustic conditions and identified an insufficient ability to track temporal changes in the acoustic properties of the environment. We proposed a novel way of generating training data that results in a considerable improvement in the model's capability of estimating changing RT_{60} while incurring only a slight loss of estimation accuracy under static acoustic conditions. The findings presented in this study illustrate the great potential of data-centric approaches to existing problems in acoustic scene analysis in particular and machine learning in general.

6. REFERENCES

[1] "ISO 3382-1:2009(EN): Acoustics — measurement of room acoustic parameters," 2009.

[2] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[3] Jacob Benesty, Jingdong Chen, and Emanuël A.P. Habets, *Speech enhancement in the STFT domain*, Springer Science & Business Media, 2011.

[4] Luca Remaggi, Kim Hansung, Annika Neidhardt, Adrian Hilton, and J.B. Philip, "Perceived quality and spatial impression of room reverberation in VR reproduction from measured images and acoustics," in *Proc. of the 23rd Int. Congr. Acoust*, 2019.

[5] James Eaton, Nikolay D. Gaubitch, Alastair H. Moore, and Patrick A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1681–1693, 2016.

[6] Hafiz Malik and Hany Farid, "Audio forensics from acoustic reverberation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 1710–1713.

[7] Rama Ratnam, Douglas L. Jones, Bruce C. Wheeler, William D. O'Brien Jr., Charissa R. Lansing, and Albert S. Feng, "Blind estimation of reverberation time," *The Journal Acoust. Soc. of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[8] Heinrich W. Löllmann and Peter Vary, "Estimation of the reverberation time in noisy environments," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2008.

[9] James Eaton, Nikolay D. Gaubitch, and Patrick A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 161–165.

[10] Heiner Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2010, pp. 1–4.

[11] Jimi Y.C. Wen, Emanuël A.P. Habets, and Patrick A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 329–332.

[12] Hannes Gamper and Ivan J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2018, pp. 136–140.

[13] Suradej Duangpummet, Jessada Karnjana, Waree Kongprawechnon, and Masashi Unoki, "Blind estimation of room acoustic parameters and speech transmission index using MTF-based CNNs," *arXiv preprint arXiv:2103.07904*, 2021.

[14] Paul Callens and Milos Cernak, "Joint blind room acoustic characterization from speech and music signals using convolutional recurrent neural networks," *arXiv preprint arXiv:2010.11167*, 2020.

[15] Ricardo Falcon Perez, Georg Götz, and Ville Pulkki, "Machine-learning-based estimation of reverberation time using room geometry for room effect rendering," in *Proc. of the International Congress on Acoustics: integrating 4th EAA Euroregio*, 2019, vol. 9, p. 13.

[16] Shuwen Deng, Wolfgang Mack, and Emanuël A.P. Habets, "Online blind reverberation time estimation using CRNNs," in *Proc. Interspeech Conf.*, 2020, pp. 5061–5065.

[17] Andy Sarroff and Roth Michaels, "Blind arbitrary reverb matching," in *Proc. Conf. on Digital Audio Effects*, 2020.

[18] Christian J. Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia, "Filtered noise shaping for time domain room impulse response estimation from reverberant speech," *arXiv preprint arXiv:2107.07503*, 2021.

[19] Roy D. Patterson, Ian Nimmo-Smith, John Holdsworth, and Peter Rice, "An efficient auditory filterbank based on the gammatone function," in *A meeting of the IOC Speech Group on Auditory Modelling at RSRE*, 1987, vol. 2.

[20] Ben Gold, Nelson Morgan, and Dan Ellis, *Speech and audio signal processing: processing and perception of speech and music*, John Wiley & Sons, 2011.

[21] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[22] André Araujo, Wade Norris, and Jack Sim, "Computing receptive fields of convolutional neural networks," *Distill*, vol. 4, no. 11, pp. e21, 2019.

[23] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, pp. 9–48. Springer, 2012.

[24] Poju Antsalo, Aki Makivirta, Vesa Valimaki, Timo Peltonen, and Matti Karjalainen, "Estimation of modal decay parameters from noisy response measurements," in *Audio Engineering Society Convention 110*, 2001.

[25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[26] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[27] Marc Schröder, Marcela Charfuelan, Sathish Pammi, and Ingmar Steiner, "Open source voice creation toolkit for the MARY TTS Platform," in *Annual Conference of the International Speech Communication Association*, 2011.

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

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

[30] "Echothief impulse response library," http://www.echothief.com/, Accessed: 2021-10-05.

[31] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal Acoust. Soc. of America*, vol. 65, no. 4, pp. 943–950, 1979.

[32] Tijmen Tieleman, Geoffrey Hinton, et al., "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[34] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," in *Proc. Intl. Conf. Machine Learning (ICML)*, 2017, pp. 3319–3328.