




Article

Investigation of Machine Learning Model Flexibility for Automatic Application of Reverberation Effect on Audio Signal

Mantas Tamulionis [†], Tomyslav Sledevič [†] and Artūras Serackis ^{*,†}

Department of Electronic Systems, Vilnius Gediminas Technical University (VILNIUS TECH), Plytinės g. 25, LT-10105 Vilnius, Lithuania; mantas.tamulionis@vilniustech.lt (M.T.); tomyslav.sledevic@vilniustech.lt (T.S.)

* Correspondence: arturas.serackis@vilniustech.lt

† These authors contributed equally to this work.

Abstract: This paper discusses an algorithm that attempts to automatically calculate the effect of room reverberation by training a mathematical model based on a recurrent neural network on anechoic and reverberant sound samples. Modelling the room impulse response (RIR) recorded at a 44.1 kHz sampling rate using a system identification-based approach in the time domain, even with deep learning models, is prohibitively complex and it is almost impossible to automatically learn the parameters of the model for a reverberation time longer than 1 s. Therefore, this paper presents a method to model a reverberated audio signal in the frequency domain. To reduce complexity, the spectrum is analyzed on a logarithmic scale, based on the subjective characteristics of human hearing, by calculating 10 octaves in the range 20–20,000 Hz and dividing each octave by 1/3 or 1/12 of the bandwidth. This maintains equal resolution at high, mid, and low frequencies. The study examines three different recurrent network structures: LSTM, BiLSTM, and GRU, comparing the different sizes of the two hidden layers. The experimental study was carried out to compare the modelling when each octave of the spectrum is divided into a different number of bands, as well as to assess the feasibility of using a single model to predict the spectrum of a reverberated audio in adjacent frequency bands. The paper also presents and describes in detail a new RIR dataset that, although synthetic, is calibrated with recorded impulses.

Keywords: room reverberation; room impulse response; recurrent neural networks; audio signal spectrum; filter bank



Citation: Tamulionis, M.; Sledevič, T.; Serackis, A. Investigation of Machine Learning Model Flexibility for Automatic Application of Reverberation Effect on Audio Signal. *Appl. Sci.* **2023**, *13*, 5604. <https://doi.org/10.3390/app13095604>

Academic Editors: Giovanni Costantini and Daniele Casali

Received: 28 March 2023

Revised: 24 April 2023

Accepted: 26 April 2023

Published: 1 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The room impulse response, which represents the acoustic properties of the room, is widely used in a broad range of audio signal-processing tasks. RIR can be useful for sound source localisation [1], speech recognition [2], or speech signal separation [3]. If the room being analyzed is characterized by unwanted acoustic phenomena, the measured RIR can show spectral changes [4]. These changes can be eliminated by an equalization scheme. The influence of room acoustic characteristics on the RIR spectrum can vary depending on the location of the measurement, so the equalization scheme must be adaptive. Such adaptive equalization schemes typically use an FIR filter whose attenuation coefficients are continuously updated to reduce the difference between the spectrum actually obtained at the measurement position and the desired spectrum [5]. However, the flexibility of the filter depends on the filter order and the coefficient estimation algorithms.

Updating the attenuation coefficients of the FIR filter is usually performed using the filtered-x least mean square (FxLMS) algorithm [6]. However, it was later discovered that this algorithm is not stable and can cause sudden interference in its error signal [7]. Subsequent studies have proposed the use of the maximum correntropy criterion (MCC) method for adaptive filtering, which has been shown to be more robust than previously popular methods [8]. Even later, the generalised maximum correntropy criterion (GMCC)

method was proposed and performed better than the standard MCC [9]. The RIR impulse is observed to be a sparse set of coefficients, i.e., some of its intermediate values are close to zero. On this basis, it was decided that the equalization process could be further improved if the adaptive algorithm took advantage of the sparseness of the RIR impulse [10].

To create the impression of realistic room acoustics as the listener's position varies in the virtual room, we need to continuously convolve an anechoic signal in real time with a different RIR filter from a large dataset. The dataset should consist of RIRs recorded in a real room, but this is time consuming, as each new RIR requires a new measurement when a new position is chosen for the sound source or receiver. This means that, for example, to capture a set of RIR data covering the entire area of a small room (up to 10 sq. m.), more than 1000 measurements may be required, with the position of the measuring microphone changing every 10 cm. In addition, a quiet environment is needed to ensure the quality of the RIR measurements. According to ISO 3382-1, the sound source must emit a sound pressure level at least 35 dB above the background noise in the room [11].

As an alternative to RIR measurements, RIR can be modeled using one of the geometric acoustics methods. The most commonly used one is the image source method (ISM) [12]. However, satisfactory results can only be achieved in this way by modelling an almost empty room with clear geometry (e.g., a rectangle). The ISM method is a simplified assumption that sound waves propagate in straight lines at a fixed speed, the energy is uniformly attenuated, and the waves are mirrored when they reach a surface. In the real world, the sound wave is not perfectly reflected; some of it is scattered in different directions, depending on the roughness of the surface. Only the early reflections are mirror-like, and later they become increasingly diffuse. Thus, in practice, a hybrid approach is often used, where the first reflections are modeled by an ISM and the later ones by the ray-tracing method. The ISM method also does not allow the modelling of objects in the room that interfere with the propagation of the sound wave and cause reflections. Tang et al. proposed improvements using a Monte Carlo path-tracing method that can model diffuse reflections, which means better simulation of existing obstacles [13]. However, the authors point out that this algorithm also has the disadvantage of not being able to model low frequencies and diffraction well. There have been attempts to solve the problem with the use of artificial neural networks which, trained on existing RIRs, can predict the desired data.

The use of neural networks can be a more flexible approach and a good alternative to this task. The RIR can be estimated using its spectrogram as an image, as well as individual RIR parameters such as the geometry of the simulated room and the absorption coefficients of its surfaces. In the study by Yu and Kleijn, the RIR parameters were estimated separately, with convolutional neural networks (CNNs) used for room geometry and feedforward multilayer perceptrons (MLPs) for surface absorption coefficients [14]. The authors claim that their method works when neural networks are trained with a single RIR impulse. In fact, it should be noted that this condition is only partially fulfilled, as the algorithm is initially allowed to learn from a single simulated RIR impulse that has been generated by the ISM method using the RIR generator [15]. Afterwards, it has been shown that much better results can be achieved by increasing the number of RIRs dedicated to training. In addition, the performance of the algorithm is tested by training the networks on the recorded RIRs. The BUT ReverbDB dataset is used for this purpose [16].

Machine learning methods are applied not only to RIR generation but also to other acoustic environment analysis tasks. Classification of rooms by volume using RIR can be performed using statistical pattern recognition [17]. The authors of this paper claim that their algorithm does not require data about the distance between the sound source and the microphone. However, good results were only achieved using simulated rather than measured RIRs. Convolutional neural networks are used to perform speech recognition tasks and to build speech-to-text models. In [18], the authors used a CNN-based approach to recognise tonal speech signals. Feature extraction was performed using Mel frequency cepstral coefficients (MFCC). Machine learning can also be used to assess the competence

of psychotherapists by performing speech recognition from audio and text analysis from a report together. In [19], the possibility of determining the quality of a practitioner's performance by analysing audio recordings and transcripts of psychotherapeutic conversations and comparing the result with manual assessments of competency was explored. The best predictive performance was achieved by a Lasso regression model. In [20], the authors used time-domain features (MFCCT) in addition to MFCCs in speech emotion recognition (SER) to extract features from an audio signal. The CNN-based SER model outperformed comparable models that used non-hybrid features. Machine learning is also being applied in the field of tourism to generate additional recommendations for destinations with fewer reviews on specialised tourism portals. Missing reviews can be identified and selected from social media posts containing geolocation information. In [21], the authors used machine learning-based clustering and classification methods, namely a fine-tuned transformer neural network-based BERT model.

In this paper, we present a new dataset for RIR estimation based on the fusion of recorded and simulated RIRs. In addition, we present a study of an alternative method for modeling the spectrum of a reverberated signal. The idea of this paper is to check if a neural network can learn the effects of acoustics and replace the traditional method of using RIR filters. We train the neural network with frequency-domain data, dividing logarithmically into 1/3 or 1/12 octave. The studies test the feasibility of modeling reverberated audio for several different frequency bands by training a model for only one band, thus trying to avoid the need to train different models for each frequency band separately.

We have chosen recurrent neural networks (RNNs) for this task because they could be good for modeling reverberating audio, as they are designed to handle sequential data, allowing them to account for the time-varying nature of audio signals, and their internal memory cells can effectively capture the dependencies between successive audio samples, leading to a more accurate representation of reverberation characteristics. The bidirectional LSTM, LSTM, and GRU recurrent neural network architectures offers unique strengths and trade-offs in terms of modeling capacity, computational efficiency, and memory requirements, and a thorough evaluation can help identify the most suitable approach for capturing the complex temporal relationships present in reverberating audio signals, ultimately leading to better performance and practical applicability.

The structure of the article is as follows: Section 2 presents the dataset and the methods used in our study. The preparation of the dataset is described in detail in Section 2.1. Section 2.2 provides a detailed explanation of our method, which compared three recurrent neural network structures that attempted to predict room reverberation for each octave band. Section 3 describes our experimental setup and a comparison of the reverberation prediction results using different recurrent neural network models. Section 4 provides a discussion and concludes the results of our study.

2. Materials and Methods

2.1. Preparation of the Dataset

To train the algorithm properly, we need to create a large set of data samples, avoiding to record all RIRs as this would be time-consuming, but trying to maintain the authenticity of the RIR impulses. To achieve these goals, we decided to create a dataset of synthetic impulses, but based on the recorded RIRs. First, measurements were made in a university laboratory, choosing a small number of fixed measurement positions. Subsequently, an identical room was designed and imported into the "Odeon" acoustic design software. The acoustic parameters of the measured and modeled RIRs were compared and the absorption coefficients of the modeled room surfaces were changed accordingly. This allowed the creation of new synthetic RIRs that are authentic and correspond not only to several measured room positions but also to any selected point in the virtual room.

Measurements were taken in a small rectangular room. The main purpose of the room was to test the VR software, so it was almost empty; only three wooden tables remained after the computer screens were removed. The room has a floor area of 31.35 m² and a

ceiling height of 2.86 m. Three walls of the room are covered with large porous bricks, one wall is concrete and painted, and the ceiling is made up of small square plasterboards with aluminium gaps between them. The floor of the room is linoleum floored and access to the room is through a wide glass door.

Authentic room pulses were recorded according to ISO 3382-1 [11]. It is recommended to select and test at least two different sound source positions in the room (with a height of 1.5 m from the ground), as well as at least three to four microphone positions, which should be spaced at least 2 m (half the measured wavelength of the lowest frequency) apart, and at least 1 metre (a quarter of the wavelength of the lowest frequency) away from any reflecting surface. The different microphone positions should be chosen in such a way that the results take into account the reflections produced by all walls covered with different materials, and the height of the measuring microphone should be adjustable to 1.2 m, which corresponds to the typical height of the ear position of a seated listener. To maintain the distances specified in the standard, two positions of the sound source and three positions of the microphones were selected and tested, resulting in a total of six different combinations. The sound source and microphone positions are shown in Figure 1, as well as the grid of microphone positions used in the virtual version of this room. The standard also specifies that the sound source should be omnidirectional and should reproduce all frequencies uniformly between 125 Hz and 4000 Hz. However, to analyze the effect of room acoustics on human voice, these measurements were carried out using a directional loudspeaker, Genelec 8010A, whose directivity is compared to that of human speech in Figure 2 [22].

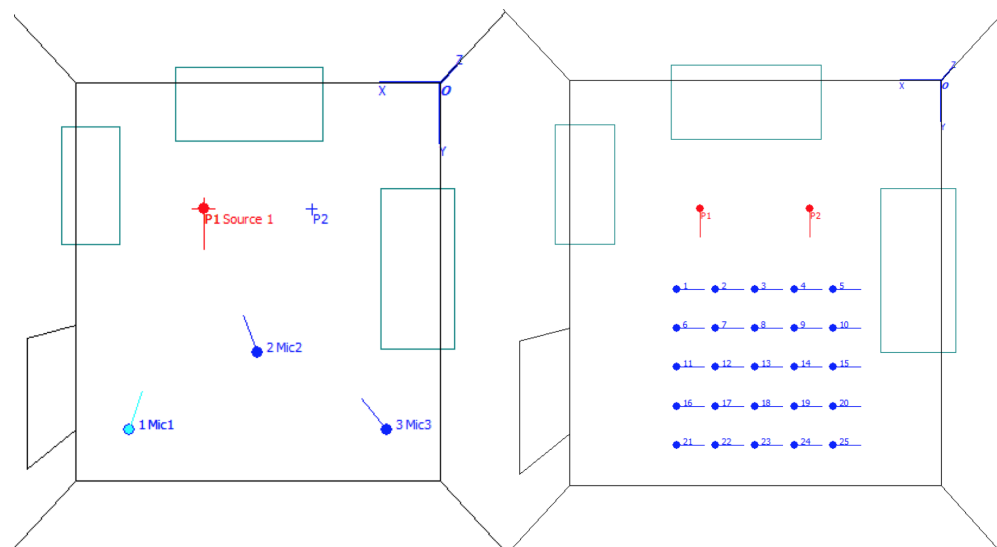


Figure 1. Left image shows the locations of the sound source and the measurement microphone in a real room, the different combinations of which were recorded as separate RIR impulses. Right image shows the selected positions of the 2 sources and 25 receivers respectively in the virtual room.

A sonarworks XREF 20 omnidirectional microphone and RME Fireface UC sound card were used as receiver and recorder. We also used the Measure impulse response tool offered by Odeon 16, which allows us to generate and transmit an exponential sweep signal and record the impulse.

The same room was then modeled in SketchUp and imported into Odeon. With the same positions for the sound sources and microphones, as well as the assumed absorption coefficients for the surfaces, the RIR simulation was performed. Odeon allows the technical characteristics of a real loudspeaker—directivity, frequency response, dynamic range, etc.—to be assigned to a virtual sound source. The user has to import and activate a CLF (common loudspeaker format) file, which can be downloaded for each specific model of almost all popular loudspeaker manufacturers.

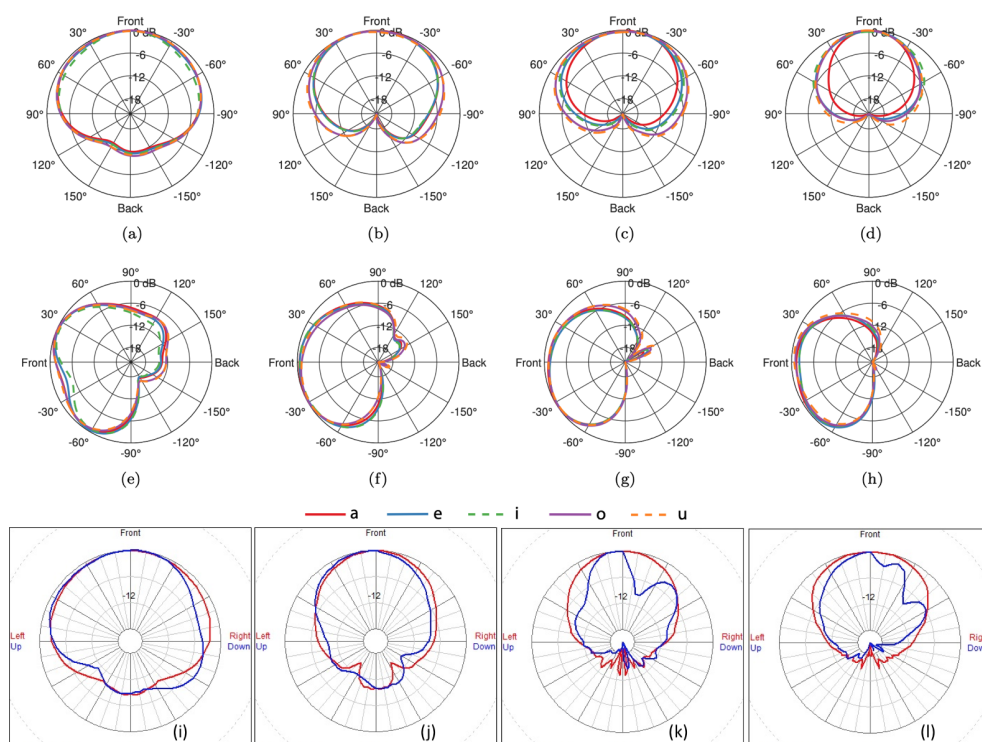


Figure 2. Comparison of the directionality of the human voice in different vowels (a–h) with the directionality of the Genelec 8010A loudspeaker used in the measurements (i–l), in the horizontal ((a–d) and red line in (i–l)) and in the vertical planes ((e–h) and blue line in (i–l)) in different bands: 1 kHz (a,e,i), 2 kHz (b,f,j), 4 kHz (c,g,k), and 8 kHz (d,h,l).

Odeon has the ability to import recorded pulses and compare them with simulated ones. The accuracy of the results depends on the precise choice of the surface absorption coefficients, and initially the results varied considerably. Another Odeon tool, “Genetic Material Optimizer”, was then used [23]. It compares the characteristics of the recorded and simulated pulses and tries to recalculate the possible absorption coefficients. Before running the algorithm, it is necessary to select the permissible limits of variation of the absorption coefficient for each material. For porous bricks and plasterboard, we have set a higher modification limit. These materials cover 3 walls and the ceiling; in general, most of the room surface. We can see that the algorithm only slightly changed the absorption coefficient of the materials with a modification limit of 50%, whereas the absorption coefficient of the materials with a higher modification limit was changed in detail.

The differences between the recorded and simulated impulses are evaluated by the JND (just-noticeable difference) value [24], which is also described in the ISO standard and corresponds to 1 dB for most acoustic evaluation parameters. This means that if the difference between the impulses is less than 1 JND, it can be assumed to be negligible and can be ignored. Before the algorithm was run, this value ranged from 13 to 15 JND in the individual frequency bands; after optimization it ranged from 0.7 to 3. Only in the lower frequency bands did the differences remain larger, but the developers of Odeon warn the user that the algorithm is not able to reduce the differences to below 1 JND in the lower frequency bands. Once the absorption coefficients have been optimized and the differences between the simulated and recorded impulse parameters have been verified to be within acceptable limits, it can be said that we have simulated the acoustics of a virtual room that closely matches the acoustics of a real room. In this case, we can create RIRs not only for the 3 fixed measurement locations but also for any point in the virtual room. In Figure 3 a comparison of the measured and simulated RIRs can be seen in terms of the early decay time before and after optimisation of the absorption coefficients. Figure 4 shows

the similarity of the spectrum of the human voice when such a signal is convolved with a measured or simulated RIR impulse.

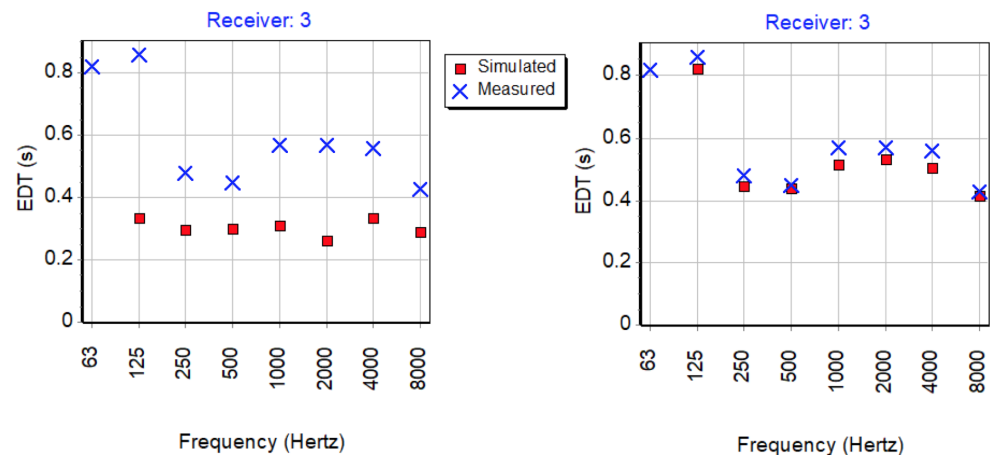


Figure 3. The recorded and simulated RIRs were compared with early decay time (EDT) values in different frequency bands before optimization (left) and after optimization (right).

Using the methodology described above, 50 RIRs were created for this study from 2 source positions and 25 receiver positions spaced 0.5 m apart. Using our calibrated virtual room model, we can create a larger dataset if necessary. Most importantly, the simulation is realistic, validated by real records. This makes any new study more valuable, as newly implemented models can be trained and tested on real acoustic behaviour, rather than on a dataset that is usually built using simplified models in an environment that will never be close to a real room. The latest version of the described dataset and more detailed technical information can be found in <https://github.com/tamulionism/Room-Impulse-Response-dataset>, accessed on 30 April 2023.

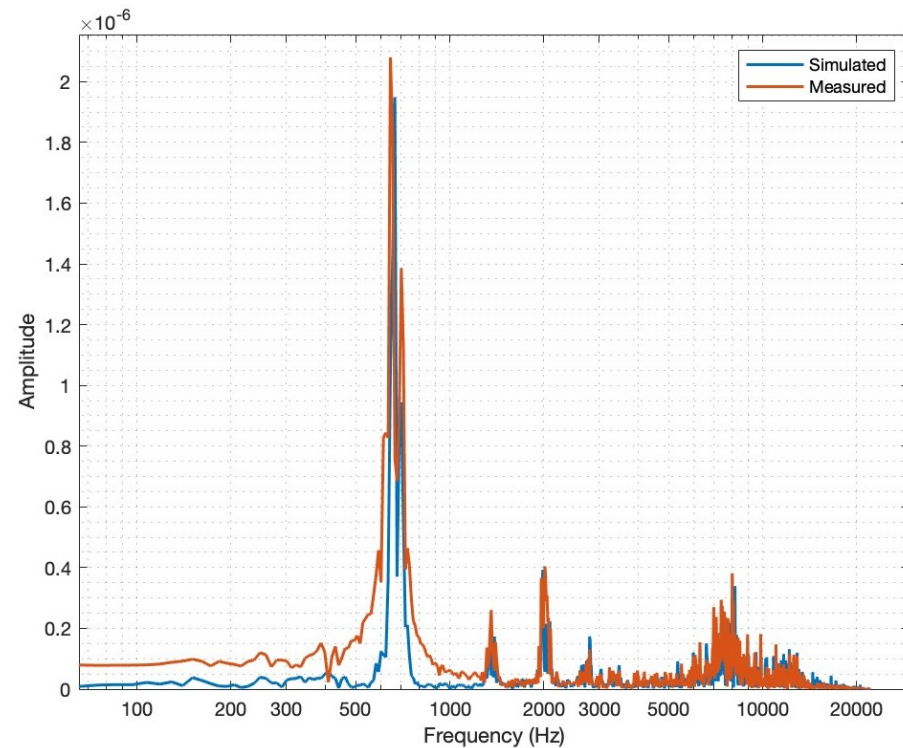


Figure 4. Comparison of the spectra of the anechoic voice signal of a singing woman convolved with an Odeon simulated RIR and an RIR impulse measured in a real room.

2.2. Deep Recurrent Neural Networks for Reverberated Signal Modeling

Three slightly different recurrent neural network structures were compared, which could be used as candidates for a reverberation prediction model:

- Long short-term memory (LSTM) [25];
- Bi-directional long short-term memory (BiLSTM) [26];
- Gated recurrent units (GRU) [27].

The architecture of the recurrent neural networks (RNN) includes feedback connections, making them more suitable for modeling acoustic effects than feed-forward network structures.

We try to predict the spectrum of the reverberated signal separately for each octave. In our study, we test all three neural network structures by training them on different frequency bands of the reverberated signal. Each prediction model consists of an input layer to which a sequence of time-varying spectral node values is sent, as well as two layers of recurrent neural network cells, one fully connected layer, and a regression layer, which generate a predicted sequence of changes in the spectral nodes over time.

To investigate the relationship between the number of RNA cells in a layer and the accuracy of the predicted spectral band, we tested the performance of networks with three different combinations of cell numbers. We first selected 10 cells in the first layer and 20 cells in the second layer, then repeated the experiments by equalling the number of cells in the two layers to 20, and finally we performed another series of experiments by increasing the number of cells in the second layer to 40.

In the experiments, we try to evaluate the ability of different network structures to predict a reverberated signal:

1. In the same frequency band used in the training, but replacing the input samples with previously unknown ones;
2. In two adjacent frequency bands, when the model was trained on the middle band and tested on adjacent bands
3. In all frequency bands when the octave is divided into 12 parts. Firstly, when a separate model was trained to predict each frequency band, and secondly, when the prediction was performed by taking input data from each frequency band separately, and the reverberated signal was predicted using a model trained on only one frequency band.

Variations of the experimental set-up were carried out to determine how flexible the prediction model can be to predict a specific band of the reverberated signal. In addition, it was necessary to see how different the model should be for neighboring frequency bands when the octave is divided into 3 or 12 parts.

The audio used for model training and experimental testing was divided into 250 ms analysis frames. This is the maximum delay time that can be accepted in real-time auralization systems [28]. The conversion from time to frequency domain was performed using a window of 512 sample width with 256 sample overlaps. The data used to train the models were divided into three parts: training (70%) validation (15%) and testing (15%). All models were trained using the same training options: the ADAM optimizer, a constant learning rate of 0.001, shuffle of the data after each epoch from 10,000, and a small batch size of 50.

3. Results

Table 1 presents the results of an experimental study where we used different RNN types (bidirectional LSTM, LSTM and GRU) architectures to simulate the reverberated signal in a single frequency band. The aim of this study was to investigate which RNN architecture can be used for reverberant signal modeling and how the size of the recurrent layers affects the results.

As can be seen from Table 1, RNN structures with more parameters, such as LSTM and BiLSTM, show a stable increase in R-squared as we increase the size of the hidden layer (the number of recurrent unit cells in the layer). The GRU-based RNN structure showed

unstable results after training, so that in some experimental studies the GRU-based model was not used at all (see Table 2).

Table 1. Comparison of different RNN structures by varying the number of cells in the layers of the recurrent neural network.

RNN	Layer Size	SSE (10^3)	RSS [min max]	R-Squared	RMSE [min max]
LSTM	10 + 20	15.85	[0.24 65.29]	−0.13	$[3.36 \times 10^{-5} \ 1.46]$
	20 + 20	3.67	[0.23 37.39]	0.74	$[2.24 \times 10^{-5} \ 0.66]$
	20 + 40	1.36	[0.19 26.03]	0.90	$[1.57 \times 10^{-5} \ 0.56]$
BiLSTM	10 + 20	6.43	[0.15 43.44]	0.54	$[1.18 \times 10^{-4} \ 0.81]$
	20 + 20	2.71	[0.17 22.16]	0.81	$[7.09 \times 10^{-6} \ 0.37]$
	20 + 40	2.16	[0.17 20.38]	0.85	$[2.06 \times 10^{-5} \ 0.38]$
GRU	10 + 20	5.05	[0.32 29.97]	0.64	$[1.30 \times 10^{-4} \ 0.60]$
	20 + 20	2.45	[0.21 34.03]	0.83	$[6.37 \times 10^{-5} \ 0.60]$
	20 + 40	3.31	[0.18 32.59]	0.77	$[7.99 \times 10^{-5} \ 0.66]$

Table 2. Comparison of different RNN structures with fixed layer sizes, trained on a single bin, covering 1/12 of the octave band width. Tested on 12 neighboring bins.

RNN	Bin Number	SSE (10^4)	RSS (Mean)	R-Squared	RMSE (Mean)
LSTM	1	30.5	9.59	0.63	0.1621
	2	52.6	12.54	0.37	0.2272
	3	75.6	14.45	0.27	0.2703
	4	84.3	14.74	0.27	0.2793
	5	65.4	14.10	0.32	0.2545
	6	17.6	9.16	0.57	0.1337
	7	1.72	2.95	0.92	0.0191
	8	12.8	4.78	0.70	0.0523
	9	106	5.69	0.33	0.0881
	10	129	6.92	0.27	0.1177
	11	117	6.55	0.35	0.1060
	12	147	6.99	0.39	0.1091
BiLSTM	1	15.9	8.48	0.81	0.1788
	2	13.0	9.71	0.84	0.2240
	3	15.1	10.55	0.85	0.2614
	4	18.3	10.71	0.84	0.2749
	5	15.7	10.77	0.84	0.2709
	6	8.1	7.28	0.80	0.1507
	7	1.10	1.94	0.95	0.0145
	8	10.4	4.22	0.76	0.0489
	9	86.8	5.09	0.45	0.0821
	10	104	4.76	0.41	0.1321
	11	103	4.84	0.43	0.1218
	12	114	6.28	0.52	0.0817

To compare the flexibility of the selected RNNs in learning individual frequency bands of the reverberated signal, we trained 30 structures (each of the 10 octaves of the human audible frequency spectrum was divided into three bands). We used RNN models with 20 cells in the first hidden layer and 40 cells in the second hidden layer, which is the largest structure studied in the first experiment and which showed the best fitting results.

Tables 3 and 4 show the results of an experimental study to test whether a model trained to predict the central band of an octave divided into three parts is good enough to predict adjacent frequency bands. We compared the results for 8 different octaves, ignoring only the first and last octaves—frequencies below 40 Hz and above 10 kHz. A noticeable reduction of fit was observed. We can also see from the results that the use of the central frequency band model to predict neighboring frequency bands also depends on the octave

chosen. This is an expected result, as we cannot normally achieve a uniform distribution of sound content across all octaves in any real recording dataset.

Table 3. R-squared fitting estimate comparison of LSTM trained on a single bin, tested on neighboring ones, using resolution of 1/3 of the octave band width.

R-Squared	5 bin	8 bin	11 bin	14 bin	17 bin	20 bin	23 bin	26 bin
Bin at the Left	0.72	0.64	0.68	0.90	0.61	0.13	0.76	0.85
Bin at the Center	0.78	0.84	0.87	0.98	0.95	0.95	0.87	0.88
Bin at the Right	0.65	0.74	0.71	0.79	0.30	0.62	0.75	0.87

Table 4. RMSE fitting estimate comparison of LSTM trained on a single bin, tested on neighboring ones, using resolution of 1/3 of the octave band width.

RMSE (Mean)	5 bin	8 bin	11 bin	14 bin	17 bin	20 bin	23 bin	26 bin
Bin at the Left	0.0971	0.1375	0.1491	0.0680	0.0742	0.3782	0.0746	0.0678
Bin at the Center	0.0777	0.0680	0.0687	0.0182	0.0189	0.0260	0.0367	0.0484
Bin at the Right	0.1644	0.1100	0.1335	0.1206	0.1737	0.2157	0.0940	0.0531

By dividing each octave into 12 parts, we can analyze the half-tone pattern of the reverberated signal. In this part of the study, we first decided to compare the ability to learn from samples for each frequency band separately. The results are shown in Table 5. We again trained three RNN structures with layer sizes of 20 + 40 RNN cells in two hidden layers, respectively. The LSTM and BiLSTM-based models showed relatively stable results, but the GRU-based RNN was difficult to train to be close to matching all 12 frequency bands.

Table 5. Comparison of different RNN structures with fixed layer sizes, trained on a single bin, covering 1/12 of the octave band width. 12 trained models in total, for neighboring bins.

RNN	Bin Number	SSE (10^3)	RSS (Mean)	R-Squared	RMSE (Mean)
LSTM	1	30.42	3.59	0.96	0.0227
	2	70.52	4.41	0.91	0.0325
	3	59.89	4.89	0.94	0.0354
	4	110.75	5.62	0.90	0.0483
	5	117.16	5.53	0.88	0.0452
	6	38.58	4.55	0.91	0.0311
	7	17.16	2.95	0.92	0.0191
	8	19.62	2.93	0.95	0.0194
	9	23.79	2.18	0.98	0.0113
	10	245.23	1.94	0.86	0.0176
	11	315.59	8.44	0.82	0.1642
	12	16.87	2.13	0.99	0.0134
BiLSTM	1	38.78	3.05	0.95	0.0278
	2	62.71	3.30	0.92	0.0249
	3	29.26	3.46	0.97	0.0236
	4	36.86	3.94	0.97	0.0268
	5	47.86	3.82	0.95	0.0283
	6	19.48	3.19	0.95	0.0222
	7	10.99	1.94	0.95	0.0145
	8	10.95	1.92	0.97	0.0126
	9	10.47	1.55	0.99	0.0099
	10	36.03	1.14	0.98	0.0097
	11	39.34	1.25	0.98	0.0096
	12	4.25	1.57	1.00	0.0092

Table 5. Cont.

RNN	Bin Number	SSE (10^3)	RSS (Mean)	R-Squared	RMSE (Mean)
GRU	1	32.56	3.69	0.96	0.0241
	2	68.70	4.73	0.92	0.0365
	3	86.63	5.16	0.92	0.0430
	4	145.46	5.60	0.87	0.0527
	5	73.08	5.49	0.92	0.0409
	6	89.77	4.85	0.78	0.0405
	7	95.12	3.90	0.58	0.0418
	8	24.23	2.95	0.94	0.0216
	9	57.30	2.43	0.96	0.0164
	10	N/A	32.44	N/A	0.6615
	11	N/A	31.48	N/A	0.6381
	12	124.86	2.69	0.95	0.0210

For the last study, we chose the seventh band, which is in the middle of the twelve. The experimental results of the model trained for one frequency band and used to predict the reverberation of the other frequency bands are presented in Table 2. The GRU-based RNN model was not considered in this experimental study because the initial tests showed even worse fitting accuracy and the same trends as for the LSTM and BiLSTM-based RNNs.

4. Conclusions

This paper discusses the flexibility of a recurrent neural network to automatically compute a reverberated audio signal. The algorithm models the reverberation-affected signal in the frequency domain by analysing the spectrum on a logarithmic scale. The study examines three different recurrent network structures and compares the modeling of the reverberated signal when each octave of the spectrum is divided into a different number of bands. Using a model trained for one mid-octave band (No. 7) and tested as a model for applying the reverberation effect to the remaining 11 bands, it was found that even a half-tone change in the spectrum should be analyzed separately. To ensure a good prediction of the full spectrum of the reverberated signal, we need to train a separate one-dimensional RNN model for each band. This can be defined as a limitation of our proposed method.

The BiLSTM-based RNN has shown more stable results in part of the frequency spectrum. Considering that all models were trained using the same audio recordings, it can be concluded that this type of RNN is more flexible in adapting to frequency changes related to room reverberation. Future work may explore methods for consolidating these models or refining the architecture to achieve more efficient and scalable solutions for modeling reverberated audio across different frequency bands.

Author Contributions: Conceptualization and methodology, all authors; validation, M.T.; analysis, A.S.; writing—original draft preparation, M.T. and T.S.; supervision, A.S.; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mane, S.S.; Mali, S.G.; Mahajan, S.P. Localization of Steady Sound Source and Direction Detection of Moving Sound Source Using CNN. In Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 6–8 July 2019.
2. Tang, Z.; Meng, H.Y.; Manocha, D. Low-Frequency Compensated Synthetic Impulse Responses for Improved Far-Field Speech Recognition. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6974–6978.
3. Jenrungrot, T.; Jayaram, V.; Seitz, S.; Kemelmacher-Shlizerman, I. The Cone of Silence: Speech Separation by Localization. 2020. Available online: <https://arxiv.org/abs/2010.06007> (accessed on 30 April 2023)
4. Bergner, J.; Preihs, S.; Hupke, R.; Peissig, J. A System for Room Response Equalization of Listening Areas Using Parametric Peak Filters. In Proceedings of the 2019 AES International Conference on Immersive and Interactive Audio (March 2019), York, UK, 27–29 March 2019.
5. Cecchi, S.; Carini, A.; Spors, S. Room Response Equalization—A Review. *Appl. Sci.* **2018**, *8*, 16. [[CrossRef](#)]
6. Fuster, L.; De Diego, M.; Azpicueta-Ruiz, L.A.; Ferrer, M. Adaptive Filtered-x Algorithms for Room Equalization Based on Block-Based Combination Schemes. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 1732–1745. [[CrossRef](#)]
7. Kurian, N.C.; Patel, K.; George, N.V. Robust Active Noise Control: An Information Theoretic Learning Approach. *Appl. Acoust.* **2017**, *117*, 180–184. [[CrossRef](#)]
8. He, Z.C.; Ye, H.H.; Li, E. An Efficient Algorithm for Nonlinear Active Noise Control of Impulsive Noise. *Appl. Acoust.* **2019**, *148*, 366–374. [[CrossRef](#)]
9. Zhao, J.; Zhang, H.; Wang, G. Fixed-Point Generalized Maximum Correntropy: Convergence Analysis and Convex Combination Algorithms. *Signal Process.* **2019**, *154*, 64–73. [[CrossRef](#)]
10. Kumar, K.; George, N.V. A Generalized Maximum Correntropy Criterion Based Robust Sparse Adaptive Room Equalization. *Appl. Acoust.* **2020**, *158*, 107036. [[CrossRef](#)]
11. ISO 3382-1; Acoustics—Measurement of Room Acoustic Parameters—Part 1: Performance Spaces. International Organization for Standardization: Geneva, Switzerland, 2009.
12. Allen, J.B.; Berkley, D.A. Image Method for Efficiently Simulating Small-Room Acoustics. *J. Acoust. Soc. Am.* **1979**, *65*, 943–950. [[CrossRef](#)]
13. Tang, Z.; Chen, L.; Wu, B.; Yu, D.; Manocha, D. Improving Reverberant Speech Training Using Diffuse Acoustic Simulation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6969–6973.
14. Yu, W.; Kleijn, W.B. Room Acoustical Parameter Estimation from Room Impulse Responses Using Deep Neural Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 436–447. [[CrossRef](#)]
15. Habets, E. RIR Generator. 2010. Available online: <https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator> (accessed on 30 April 2023).
16. Szoke, I.; Skacel, M.; Mosner, L.; Paliesek, J.; Cernocky, J.H. Building and Evaluation of a Real Room Impulse Response Dataset. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 863–876. [[CrossRef](#)]
17. Shabtai, N.R.; Zigel, Y.; Rafaely, B. Room Volume Classification from Room Impulse Response Using Statistical Pattern Recognition and Feature Selection. *J. Acoust. Soc. Am.* **2010**, *128*, 1155–1162. [[CrossRef](#)] [[PubMed](#)]
18. Dua, S.; Kumar, S.S.; Albagory, Y.; Ramalingam, R.; Dumka, A.; Singh, R.; Rashid, M.; Gehlot, A.; Alshamrani, S.S.; Alghamdi, A.S. Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network. *Appl. Sci.* **2022**, *12*, 6223. [[CrossRef](#)]
19. Attas, D.; Power, N.; Smithies, J.; Bee, C.; Aadahl, V.; Kellett, S.; Blackmore, C.; Christensen, H. Automated Detection of the Competency of Delivering Guided Self-Help for Anxiety via Speech and Language Processing. *Appl. Sci.* **2022**, *12*, 8608. [[CrossRef](#)]
20. Alluhaidan, A.S.; Saidani, O.; Jahangir, R.; Nauman, M.A. Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Appl. Sci.* **2023**, *13*, 4750. [[CrossRef](#)]
21. Silaa, V.; Masui, F.; Ptaszynski, M. A Method of Supplementing Reviews to Less-Known Tourist Spots Using Geotagged Tweets. *Appl. Sci.* **2022**, *12*, 2321. [[CrossRef](#)]
22. Pörschmann, C.; Arend, J.M. Analyzing the Directivity Patterns of Human Speakers. In Proceedings of the 46th DAGA, Hannover, Germany, 16–19 March 2020; pp. 1141–1144.
23. ODEON Room Acoustics Software User’s Manual. Version 16. Available online: <https://odeon.dk/download/Version17/OdeonManual.pdf> (accessed on 30 April 2023).
24. Bradley, J.S. Review of Objective Room Acoustics Measures and Future Needs. *Appl. Acoust.* **2011**, *72*, 713–720. [[CrossRef](#)]
25. Irie, K.; Tüske, Z.; Alkhouli, T.; Schlüter, R.; Ney, H. LSTM, GRU, Highway and a Bit of Attention: An Empirical Overview for Language Modeling in Speech Recognition. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 3519–3523.
26. Kurata, G.; Audhkhasi, K. Improved Knowledge Distillation from Bi-Directional to Uni-Directional LSTM CTC for End-to-End Speech Recognition. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018.

27. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A. S.; Asari, V.K. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics* **2019**, *8*, 292. [[CrossRef](#)]
28. Wenzel, E.M. Effect of increasing system latency on localization of virtual sounds. In Proceedings of the 16th International Conference: Spatial Sound Reproduction (March 1999), Arktikum, Finland, 10–12 April 1999.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.