Audio Engineering Society

Convention Paper 10520

Presented at the 151st Convention
2021 October, Online

Automatic Loudspeaker Room Equalization Based On Sound Field Estimation with Artificial Intelligence Models

Adrian Celestinos[1], Yuan Li[1], and Victor Manuel Chin Lopez[2]

[1]*Samsung Research America, DMS Audio, Valencia CA 91355, USA*
[2]*Samsung Research Tijuana, Tijuana BC 22237, Mexico*

Correspondence should be addressed to Adrian Celestinos (a.celestinos@samsung.com)

ABSTRACT

In-room loudspeaker equalization requires a significant amount of microphone positions in order to characterize the sound field in the room. This can be a cumbersome task for the user. This paper proposes the use of artificial intelligence to automatically estimate and equalize, without user interaction, the in-room response. To learn the relationship between loudspeaker near-field response and total sound power, or energy average over the listening area, a neural network was trained using room measurement data. Loudspeaker near-field SPL at discrete frequencies was the input data to the neural network. The approach has been tested in a subwoofer, a full-range loudspeaker, and a TV. Results showed that the in-room sound field can be estimated within 1–2 dB average standard deviation.

1 Introduction

When a loudspeaker radiates sound in a room, its response gets severely altered. The frequency response at the listening position can show peaks and valleys up to 20 dB, especially in the frequency range where the wavelengths are comparable with the room dimensions. These variations can cause audible artifacts based on their width, center frequency and gain [1, 2]. This frequency response is due to the interaction of sound waves with the boundaries of the room, building distinct zones with high sound pressure level (SPL), related to the room resonances, and zones with low SPL related to zones where the sound is self-canceling. The effect in the perceived sound is boominess related to the excessive low-frequency energy that causes exaggerated sustain at some frequencies in the room.

In order to equalize the loudspeaker in-room response for a restricted listening area (LA), it is required to obtain the energy average (EA) in dB, or if one desires to equalize the entire room, the total sound power (TSP) needs to be acquired. This is normally measured with a number of microphones spaced over the target listening area in the room [3, 4, 5, 6]. This study explores the use of machine learning to solve and learn the relationship, in one case between near-field pressure of a loudspeaker and its sound field in the room, and in another case to solve and learn the relationship between an average of microphones attached to extra loudspeakers to the sound field in the listening area.

The paper proposes a computer-implemented me
automatically estimate the TSP or EA over a L
out user interaction for room equalization pu
The main focus of the paper is proper sound fi
diction by machine learning. Three audio louds
applications were tested, including a subwoofe
and a multichannel loudspeaker system. The err
ysis on the three different application cases is pre
Issues as data overfit and proper evaluation of th
els are discussed. Finally data augmentation to i
the prediction performance is detailed, followed
cussions and summary.

2 Methods

2.1 Automatic room equalization

The proposed method includes, acquiring sound pressure data at one or more discrete frequencies, obtained from a frequency response of a loudspeaker in the room, via at least one microphone. The sound pressure data is input into an artificial intelligence (AI) model which incorporates a relationship between the near-field (NF) frequency response, and the EA in a listening area or the TSP.

The loudspeaker system is connected with an equalization filter. At least one microphone is configured to acquire sound pressure data at one or more discrete frequencies obtained from a frequency response of the loudspeaker in a room. An AI model is connected with the NF microphone. The sound pressure data is input into the AI model. The AI model incorporates a relationship between the frequency response and the EA over the listening area or the TSP produced by the loudspeaker in the room. A basic block diagram for a speaker system is shown in Fig. 1, which

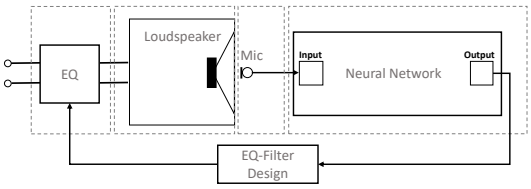

Fig. 1: Basic block diagram to estimate the in-room sound field.

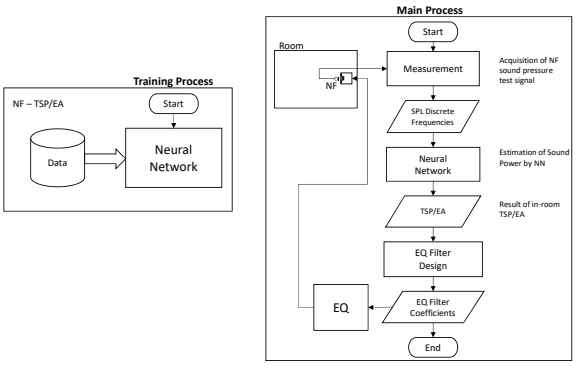

Fig. 2: Flow charts of training process and main automatic equalization process.

before the loudspeaker, that compensates the frequency response towards a desired target.

This approach can be extended to a system with multiple loudspeakers. The system further includes multiple equalization filters. Each equalization filter is connected to each one of the multiple loudspeakers. The system additionally includes multiple microphones. Each microphone is configured to acquire sound pressure data obtained from each of the multiple loudspeakers in the room. An AI model is connected with each of the multiple microphones. The sound pressure data is input into the AI model. The AI model incorporates a relationship between the near-field frequency response and the EA over a listening area or the TSP produced by each loudspeaker in the room.

2.2 Energy average

To obtain the TSP radiated by the loudspeaker into the room, the mean-squared sound pressure level at a number of microphones randomly distributed in the room can be computed as:

$$TSP(f) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |p_i(f)|^2}, \qquad (1)$$

where p_i is the sound pressure in dBs at the n number of microphones randomly located in the room at discrete frequencies f. Concerning the number of microphones needed to obtain a reliable measurement, Pedersen [4] has found that by using from 9 to 10 random microphone positions in the room, the RMS deviation from a reference estimate of the energy in the 3D sound field

of 20 random microphone positions gets down to 1 dB. Alternatively a moving microphone (MM) technique can be utilized to obtain the TSP [7]. To compute the EA produced by the loudspeaker over the LA, the sound pressure level at discrete frequencies is measured according to Eq. 1, but the n number of microphones are distributed over an area where the listeners normally are.

2.3 Neural Networks

Typically, the term artificial intelligence (AI) is used when a machine emulates cognitive functions that humans associate with other human minds, such as learning and problem solving. In this study we explore the use of machine learning to solve and learn the relationship between near-field pressure of a loudspeaker and the produced sound field in the room. More specifically we used neural networks in MATLAB's Deep Network Designer application [8], to automatically estimate the energy average in the room without user interactions. The sound pressure data was input into the model. Training was performed with dB SPL data measured in rooms using the NF response at discrete frequencies, the TSP, and EA over the listening area.

Feed-forward Neural Network

The Feed-forward neural network (FFNN) is one of the first successful artificial neural networks and is known for its simplicity. The information is only processed forward in the network. As the universal approximation theorem [9] describes, using one single hidden layer with enough hidden neurons can approximate any continuous function [10]. The FFNN applied here consists of one input layer, one output layer, and one hidden layer, (see Fig. 3). The input layer consists of neurons

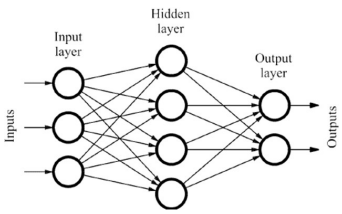

Fig. 3: Feed-forward neural network.

that receive inputs and pass them on to the other layers. The number of neurons in the input layer is equal to

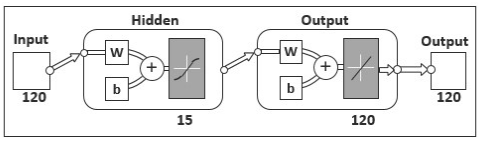

Fig. 4: Feed-forward neural network block diagram, from MATLAB.

the attributes or features in the data-set. In between the input and output layer, the hidden layer contains a number of neurons which apply transformations to the inputs before passing them to the output layer. The output layer puts out the predicted features. As the network is trained, the weights and biases are updated to learn the relationship between loudspeaker near-field response and total sound power, or energy average over the listening area.

Generalized Regression Neural Network

A generalized regression neural network (GRNN) can be categorized into the probabilistic neural networks. As described by Specht [11], a GRNN is a four-layer feed-forward neural network, consisting of an input layer, a pattern layer, a summation layer, and hidden layers, (see Fig. 5). A GRNN is often used for continuous function approximation. More specifically, the architecture for the GRNN consists of an input layer, a radial basis function (RBF) as an "activation" layer,

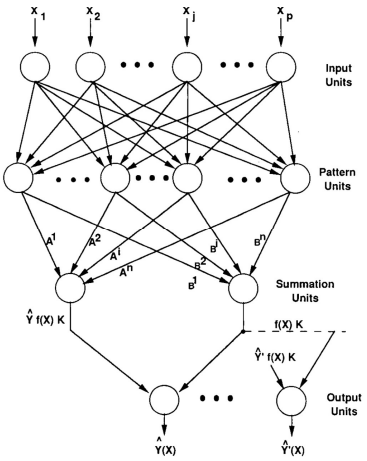

Fig. 5: Generalized regression neural network, adapted from Specht [11].

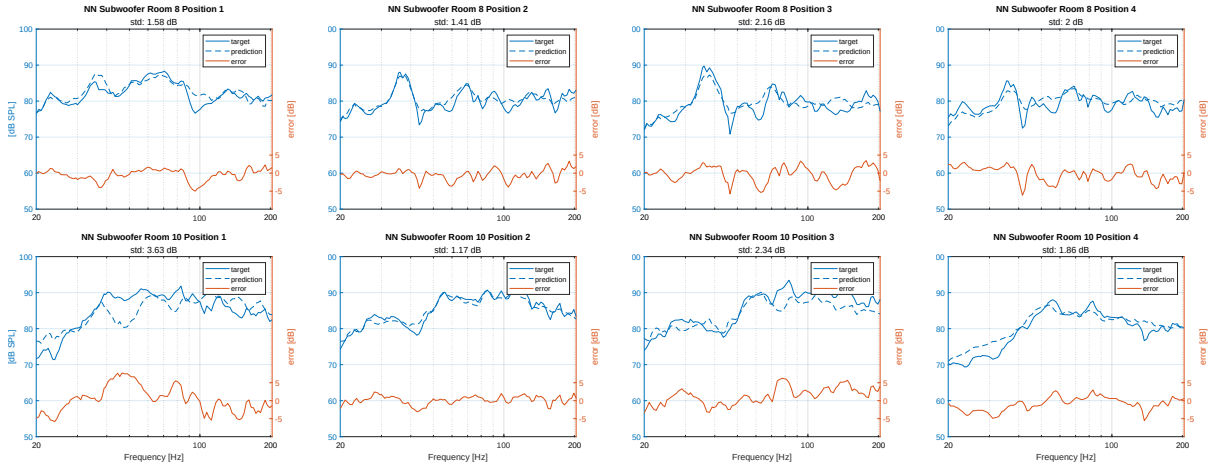

Fig. 6: FFNN subwoofer in-room prediction. Upper row, room 8, lower row, room 10.

and an output linear layer [12]. A GRNN can be utilized in different applications as system identification, modeling, and control of online dynamic systems. The training of a GRNN consists of a single-pass learning with no back propagation involved. At the same time, a GRNN requires high memory capacity and can be computationally expensive; the number of the hidden neurons is usually equal to the number of the training samples, since the hidden layer consists of at least one neuron for each pattern in the training set [11]. In MATLAB the function `newgrnn` was used to design and test a GRNN [8]. The parameter value "spread" of the radial basis functions was set to 1 in the GRNN tests.

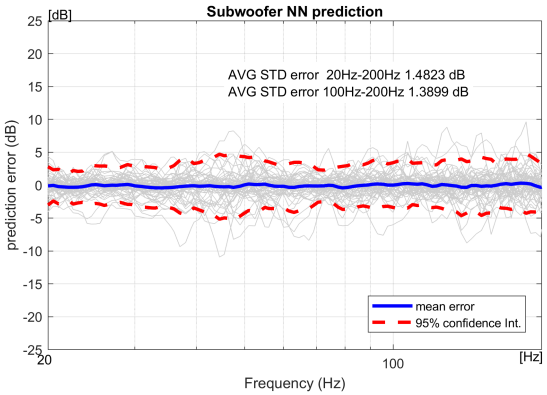

Fig. 7: FFNN subwoofer in-room prediction error.

2.4 Application Cases

In this section, the audio application cases studied in this paper are presented. In all cases the TSP or the EA over a listening area was the target output. Bayesian regularization was employed in the training process to acquire better generalization. The overall dataset was divided randomly into three subsets: training set (70%), test set (15%), and validation set (15%). The mean squared error was computed to evaluate the training performance.

Subwoofer

The first study case was a subwoofer prototype built to test a former method presented in [13]. The data obtained in the preceding study was utilized to train the network. A measurement microphone was attached to the prototype in front of the driver for the experiments.

The subwoofer was measured in 11 typical US living rooms. In each room, the subwoofer was placed at 4 to 6 positions completing a data-set of 60 typical subwoofer positions. On each room at least one corner position was included.

The near-field pressure was obtained via the attached microphone and the actual total sound power in the room was obtained using nine microphones randomly positioned in the room as detailed in Eq. 1. A multi-tone test signal was utilized to compute the complex frequency response on each microphone [14]. A FFNN, configured with one hidden layer and one output layer

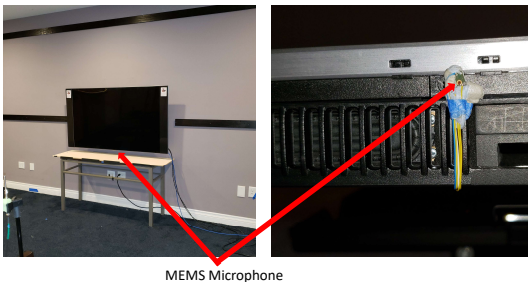

Fig. 8: TV with MEMS microphone attached for room measurements.

was utilized. In this experiment, 15 neurons were employed in the hidden layer, (see Fig. 4). The NF frequency response was the input to the FFNN, and the expected output is the TSP in the room. The training process was performed in dB scale; the produced result was in dB SPL. Figure 7, shows the prediction error for the subwoofer case. An averaged standard deviation of 1.4 dB (20–200 Hz) was accomplished, a significantly better result compared to a previous study presented by the author [13]. In Fig. 6, the TSP predictions of four subwoofer positions in rooms 8 and 10 are presented.

TV

A Samsung TV Q7 55-inch model was measured in four rectangular rooms at Samsung Audio Lab, including a gym and three listening rooms. In each room, the TV was placed at 10–15 positions acquiring 51 measurement positions. A Knowles SPM0687LR5H-1 MEMS microphone was glued in the center of the

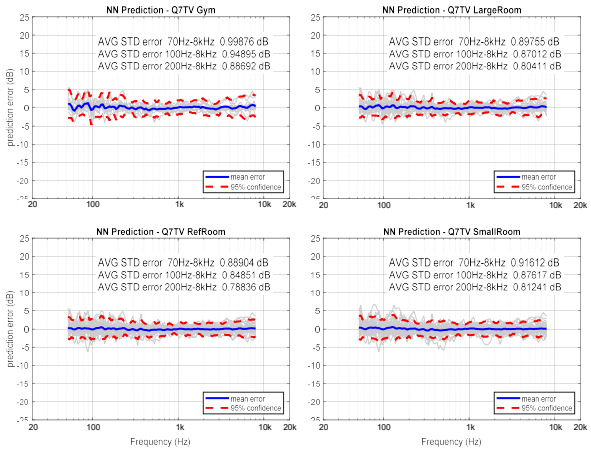

Fig. 9: TV prediction error analysis.

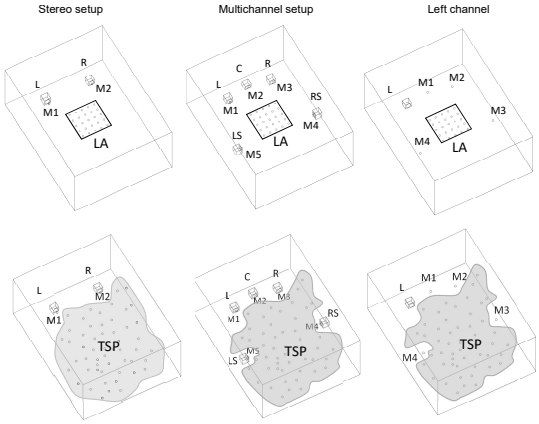

Fig. 10: LA and TSP for three loudspeaker setups.

lower part of the frame of the TV, between the two speakers, (see Fig. 8). The NF and MM measurements were conducted at each TV position in every room. The NF frequency response was the input for the FFNN, while the MM frequency response was the target output that represents the actual total sound power in the room.

For the TV case, an overall prediction error of 1.6%, and an average standard deviation of about 1 dB was achieved over the four rooms. The prediction error for the four different rooms is shown in Fig. 9.

Full-range loudspeaker

For the third case, a full-range loudspeaker was utilized to predict its sound field in the room. First, room simulations were carried out, and secondly a 7.1 multichannel loudspeaker system was set up in one of the listening rooms at Samsung Audio Lab.

FDTD Simulations

A room and loudspeaker were simulated using Finite Differences in the Time Domain (FDTD), which is a wave propagation model. A compact closed-box speaker was simulated in a rectangular room with $4.83 \times 6.36 \times 2.74$ m dimensions. The observation plane was at 1.2 m height, the considered frequency range was from 20 to 1000 Hz. More details of the simulation method can be found in [15]. A listening area defined by 5×5 virtual microphones centered at a sweet spot, at 1.2 m height was defined. To compute the TSP, 45 virtual microphones randomly spaced in

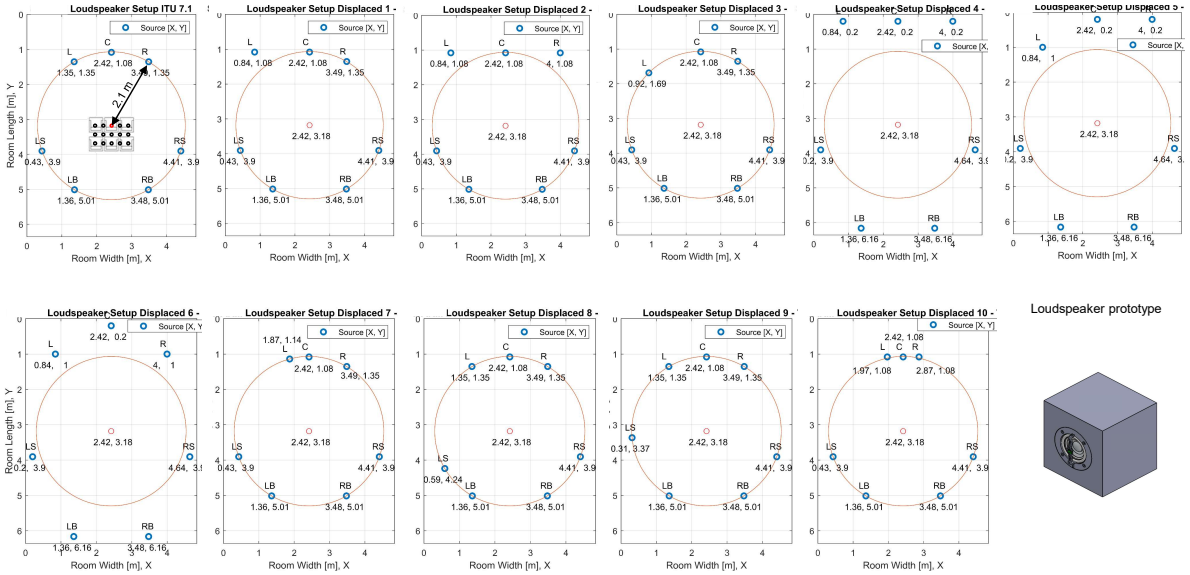

Fig. 11: Multichannel setups for training and loudspeaker prototype with NF microphone attached.

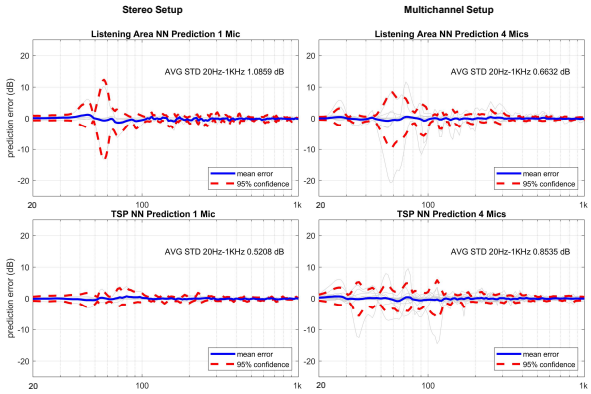

Fig. 12: Full-range loudspeaker room prediction from FDTD simulations.

the room were set up. A NF virtual microphone was included on each loudspeaker. For each loudspeaker position the EA over the LA was calculated. Each magnitude response was 1/12-octave smoothed before input to the FFNN.

Two variations were selected for this test, a stereo setup with one NF microphone on each loudspeaker, (see left graph in Fig. 10), and a multi-channel setup with one NF microphone on each loudspeaker respectively (middle graph in Fig. 10). For the first variant, to predict the EA or the TSP of Left loudspeaker, the input to the FFNN was the SPL at discrete frequencies of the NF microphone of the Right loudspeaker. For the second variant to predict the EA or the TSP of Left loudspeaker the input to the FFNN is the EA of all NF microphones excluding the active loudspeaker (Left speaker). Interpolation to 120 frequencies from the simulation data was performed for the FFNN training. The FFNN contained 10 hidden neurons, and was trained for 500 epochs. Bayesian regularization was applied for the training process.

The prediction error for the full-range loudspeaker in the FDTD room simulations are presented in Fig. 12. In left column, the prediction errors of utilizing only one microphone to predict the EA over the listening area (top graph) and the TSP (lower graph) are shown, (see Fig. 10). In the right column, the error of utilizing four microphones to predict the EA over the listening area (top graph), and the TSP (lower graph) are presented, (see loudspeaker setups in Fig. 10). As can be observed, an overall prediction standard deviation ranging from 0.5–0.8 dB was obtained.

2.4.1 Multichannel loudspeaker setup

Seven 135 × 125 × 150 mm sealed boxes with a 51 mm full-range driver each, were set up with individual multichannel amplification in one of the listening rooms at Samsung Audio Lab. Each loudspeaker prototype included a MEMS microphone mounted with

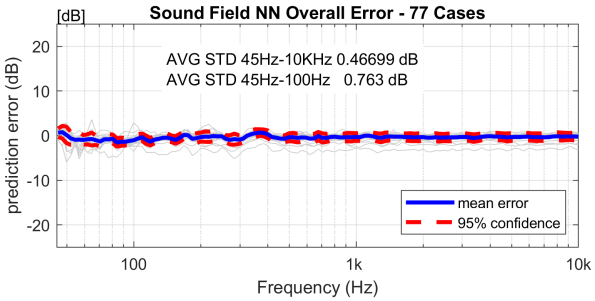

Fig. 13: Room prediction error analysis of full-range loudspeakers in multichannel setups.

a fixture in proximity of the driver as seen in Fig. 11. An ITU 7.1 loudspeaker layout was established as an initial setup for testing.

Eleven different loudspeaker setups variants, starting from the ITU 7.1 reference setup were measured for 77 cases, (see Fig. 11). A transfer function measurement from each loudspeaker terminals to each near-field microphone was obtained utilizing the logarithmic sweep method [16]. At the same time, the transfer functions from each loudspeaker to 3×5 microphones laid over a 0.65×1.3 m listening area at ear height were acquired, completing 1694 impulse responses. Each magnitude response was 1/12-octave smoothed before input to the FFNN. The data was split into 80% for training (62 cases), and 20% for validation (15 cases). Batch normalization was utilized to improve the results. The FFNN for this study had the following attributes:

- Sequence: 1×125 vector (dB) magnitude loudspeaker sound field response
- Batchnorm: batch Normalization layer
- Bilstm: bidirectional LSTM layer with input size = 125 and 512 hidden units
- Batchnorm: another Batch Normalization layer
- Output: a fully connected layer with Input Size = 1024
- Size = 125, 125×1024 weights , and 125×1 bias
- MSE: a regression layer with a MSE loss function

Figure 13, presents the prediction error for the multichannel loudspeaker setup in the real listening room. An average standard deviation of ± 0.46 dB (45–10k Hz) was achieved over the full data.

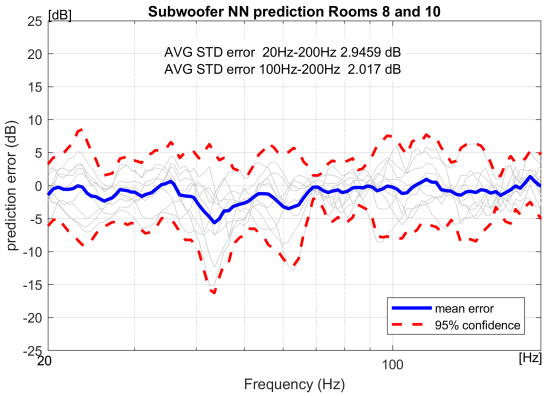

Fig. 14: FFNN subwoofer prediction error for rooms 8 and 10, network re-trained without these rooms.

3 Results

Among the three applications, the subwoofer seems to be the most challenging case in terms of predicting the TSP. In the TV application and in the multichannel loudspeaker setup, the prediction error seems to be very promising. But if one wanted to deploy this technology in a product, the key question would be: how would this technology work in a room that was not in the training data?

To answer this question for the subwoofer application, a new network with the same structure was trained, but excluding room 8 and 10 in the training set, then the network model was reevaluated with these two rooms. The result of the prediction error of this experiment is shown in Fig. 14. As it can be seen, the 95% confidence interval is about ± 4–5 dB which is not acceptable.

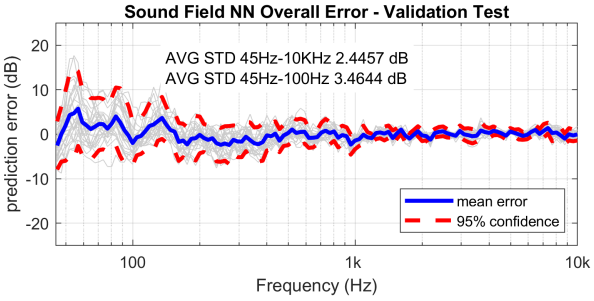

Fig. 15: FFNN full-range loudspeaker in-room prediction error, 28 different evaluation cases which were not present in training.

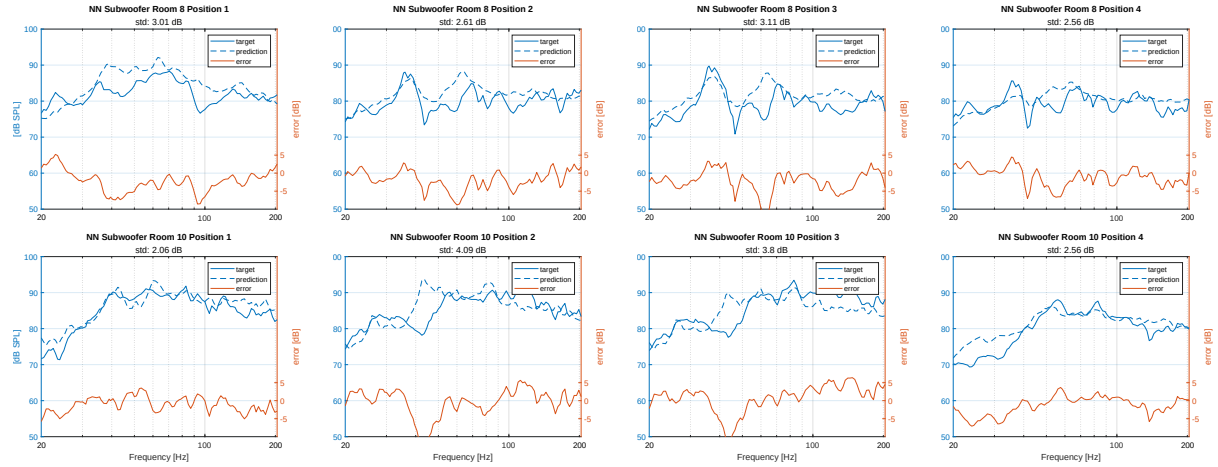

Fig. 16: FFNN subwoofer in-room prediction. Upper row, room 8, lower row, room 10. Rooms 8 and 10 excluded from training.

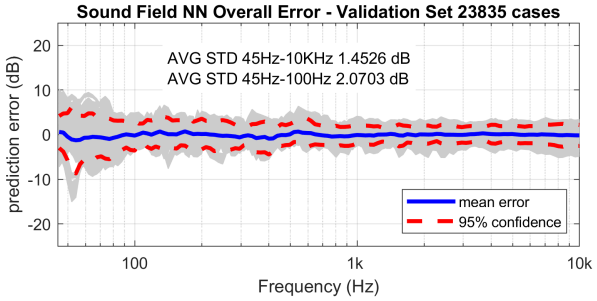

Fig. 17: Augmented FFNN data evaluation, full-range loudspeaker in-room prediction error, the 23.8k cases were not included in training.

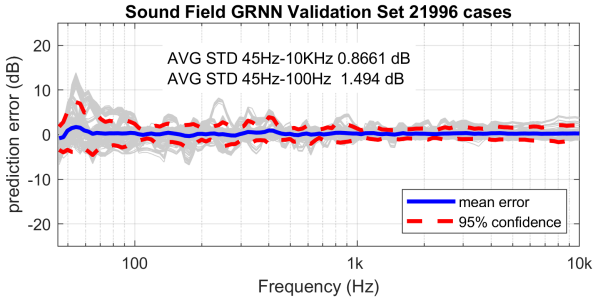

Fig. 18: Augmented GRNN data evaluation, full-range loudspeaker in-room prediction error, the 22k cases were not included in training.

Figure 16, shows the TSP prediction errors in rooms 8 and 10, evaluated with the trained network model. These rooms were not included in the training set. The detriment on the TSP estimation is obvious compared with results in Fig. 7. As for the multichannel setup, four extra loudspeaker positions setups were measured, then the FFNN model was evaluated with the data from the 28 extra loudspeaker positions. The resulting prediction on each of the 28 cases presented a variable dB offset compared with the expected EA over the listening area. To analyze the prediction error on each case, the resulting frequency response was normalized with the expected EA over the listening area. In Fig. 15 the prediction error of this experiment is shown. As it can be observed, the confidence intervals at low frequencies increased, and the standard deviation grew from 0.4 dB to 2.4 dB (45 –10 kHz), and from 0.7 dB to 3.4 dB (45–100 Hz).

These not so encouraging results may indicate that the FFNN has overfit the data. To overcome the issue of overfitting, the multichannel loudspeaker data was augmented by adding cases with a variety of high- and low-pass filters to the input and corresponding target. Also different dB offsets (up to $\pm$10 dB in 1 dB steps) were added to selected data cases, completing 95340 cases from a database of 84 cases. The result of this experiment is shown in Fig. 17 where a validation set of 23k cases was evaluated with the network model. It is worth mentioning that these cases were not included in the training. As seen in Fig. 17 the standard deviation improved from 2.4 dB to 1.4 dB (45–10 kHz), and from 3.4 dB to 2 dB (45–100 Hz).

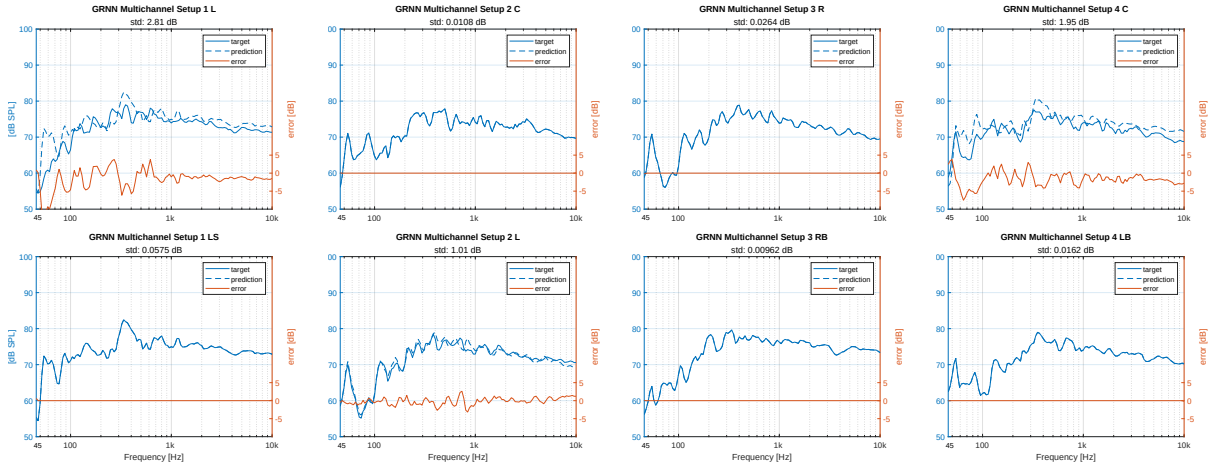

Fig. 19: GRNN Multichannel loudspeaker in-room predictions.

Finally the same augmented data was utilized to train a GRNN network. The network was trained with 73318 cases and the rest of the data was left for validation. In Fig. 18 the prediction error is presented. A significant improvement in prediction error was observed, the standard deviation reached 0.8 dB (45–10k Hz), and about 1.5 dB (45–100 Hz). Figure 19, presents frequency response results of eighth cases evaluated in Fig. 18, and predicted with the GRNN.

4 Discussion

Neural networks can be used to predict the sound field produced by a loudspeaker in a room, but it seems that the network can overfit the data if the amount of cases is insufficient.

In the subwoofer application data, consisting of 60 cases does not seem to be enough to predict the sound field in rooms that were not used in the training set. It appears that there is less relationship between the near-field response and the sound field in the room. This issue seems to also be true for the TV application case.

On the other hand, for the multichannel loudspeaker setup, the prediction of the sound field benefits because more information about the sound field in the room is gathered by the NF microphones attached to the other loudspeakers. However, it seems that the amount of data to predict the sound field in the room needs to be in the order of thousands of cases. Another discussion point is the data distribution or variety. By looking at the training of the multichannel setups (Fig. 11) there does not seem to be enough variety in terms of loudspeaker positions within the room.

The process of acquiring the right amount of data for a loudspeaker setup can be overwhelming due to the amount of loudspeaker positions and variety of rooms needed. With our measurement data of 60–80 cases the network overfit the data. Not until the data was augmented to more than 90k cases did the prediction error improved to a reasonable level for evaluation on the same room. More tests need to be done in order to verify if a network can be trained to generalize the prediction on several rooms with a limited amount of data.

Future research work may be to focus on a fast room simulation, as the image source model, to be able to simulate thousands of cases and find out what is the minimum amount of data to generalize the problem and predict the sound field with acceptable errors in several room sizes and shapes. It seems that a GRNN can give better results in this kind of audio applications, but the network has to be trained with enough variety of cases, and the amount of memory can be an issue. Other methods, such as linear regression models may be suitable for these kind of audio applications.

5 Summary

A novel method to estimate the in-room sound field produced by loudspeakers using AI models has been proposed. To automatically estimate the TSP or EA over a listening area without user interaction, neural

networks were utilized. The input to the trained model is the near-field response of the loudspeaker and the output is the predicted TSP or EA over a listening area. Three different applications have been tested, a subwoofer, a TV, and a full-range loudspeaker in a multichannel setup. Results have shown better performance in the sound field estimation, in comparison with former methods, but the neural network can easily overfit the data, if not enough cases and variety are included. Care must be taken on the amount of data used for the training. As learned from this study, data augmentation can achieve acceptable results on the sound field room prediction, but more research still needs to be done.

6 Acknowledgments

Samsung Electronics and Samsung Research America supported this work. The author would like to thank the entire staff of Samsung's US Audio Lab who helped with all aspects of this research, offered insightful suggestions, and contributed to this work, in particular: Will Saba, Eduardo Rubio, Pascal Brunet, Sunil Bharitkar and Allan Devantier.

References

[1] Toole, F. E. and Olive, S. E., "The Modification of Timbre by Resonances: Perception and Measurement," *J. Audio Eng. Soc*, 36(3), pp. 122–142, 1988.

[2] Avis, M. R., Fazenda, B. M., and Davies, W. J., "Thresholds of Detection for Changes to the Q Factor of Low-Frequency Modes in Listening Environments," *J. Audio Eng. Soc*, 55(7/8), pp. 611–622, 2007.

[3] Elliott, S. J. and Nelson, P. A., "Multiple-Point Equalization in a Room Using Adaptive Digital Filters," *J. Audio Eng. Soc*, 37(11), pp. 899–907, 1989.

[4] Pedersen, J. A., "Sampling the Energy in a 3-D Sound Field," in *Audio Engineering Society Convention 123*, 2007.

[5] Bharitkar, S. and Kyriakakis, C., "Objective Function for Automatic Multi-Position Equalization and Bass Management Filter Selection," in *Audio Engineering Society Convention 119*, 2005.

[6] Celestinos, A., Brunet, P., and Kubota, G., "Non-Linear Optimization of Sound Field Control at Low Frequencies Produced by Loudspeakers in Rooms," in *Audio Engineering Society Convention 145*, 2018.

[7] Peace, P., Nageli, S., and Sprinkle, C., "Moving Microphone Measurements for Room Response in Cinema," in *Audio Engineering Society Convention 144*, 2018.

[8] Deep Learning Toolbox version 14.1, "MATLAB," (R2020b), the MathWorks, Natick, MA, USA.

[9] Kratsios, A., "The Universal Approximation Property," *Annals of Mathematics and Artificial Intelligence*, 89(5), pp. 435–469, 2021.

[10] Hornik, K., Stinchcombe, M., and White, H., "Multilayer feedforward networks are universal approximators," *Neural Networks*, 2(5), pp. 359–366, 1989, ISSN 0893-6080.

[11] Specht, D., "A general regression neural network," *IEEE Transactions on Neural Networks*, 2(6), pp. 568–576, 1991, doi:10.1109/72.97934.

[12] Wasserman, P. D., *Advanced methods in neural computing / Philip D. Wasserman.*, Van Nostrand Reinhold, New York, 1993, ISBN 0442004613.

[13] Celestinos, A., Banka, R., and Brunet, P. M., "In-Room Low-Frequency Sound Power Optimization using Near Field Response," in *Audio Engineering Society Convention 149*, 2020.

[14] Brunet, P., Decanio, W., Banka, R., and Yuan, S., "Use of Repetitive Multi-Tone Sequences to Estimate Nonlinear Response of a Loudspeaker to Music," in *Audio Engineering Society Convention 143*, 2017.

[15] Celestinos, A. and Nielsen, S. B., "Low-Frequency Loudspeaker–Room Simulation Using Finite Differences in the Time Domain—Part 1: Analysis," *J. Audio Eng. Soc*, 56(10), pp. 772–786, 2008.

[16] Farina, A., "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *Audio Engineering Society Convention 108*, 2000.