COMPLEX-VALUED NEURAL NETWORKS FOR THE REPRODUCTION OF SINGLE FREQUENCY SOUND FIELDS

Isaac J. Lambert Institute of Sound and Vibration Research, University of Southampton, UK
Vlad S. Paul Institute of Sound and Vibration Research, University of Southampton, UK
Philip A. Nelson Institute of Sound and Vibration Research, University of Southampton, UK

1 INTRODUCTION

Sound field reproduction refers to methods in which a target acoustic field is recreated over a volume of space. In general, this process consists of recording the target field at discrete locations using sensors, and processing these recorded signals to determine optimal driving signals for an array of sources. By reproducing the target pressure with the least error possible, a listener situated within the target area will receive the same localisation cues as if they were situated within the original sound field. Such methods have obvious applications in virtual and augmented reality, as well as telecommunications. For example, this approach can be used to create a virtual auditory environment with multiple speakers in different locations[1]. Methods for sound field reproduction also have applications in sound field control, with the only difference being a reversal in sign to eliminate an acoustic field rather than recreate it.

Early methods for sound field reproduction include pressure matching using the least mean squared error (LSE) algorithm[2,3] and wave field synthesis[4]. The pressure matching method has several notable features. First, this method is most effective at reproducing virtual sources which are located close to real sources, meaning that the error is high when virtual sources are far from real sources[2]. Second, the method results in high error when reproducing frequencies above the spatial aliasing frequency, at which there are fewer than two sensors per wavelength[2]. Based on these two factors, the performance of this method improves with increased density of both sources and sensors.

Recent approaches to sound field reproduction have attempted to improve performance without the need to increase the density of sensor spacing. This can be achieved through the use of neural networks. One approach is to utilise a convolutional neural network which minimises a cost function defined in terms of the absolute error between the target and reproduced pressure[5]. When applied to cases with irregular source arrays, this method was found to result in lower error than the LSE algorithm, particularly at high frequency. A similar approach is the use of a convolutional neural network with a cost function defined in terms of the squared error between the target and reproduced pressure, and the output of sparse layers[6]. This method achieved lower error at high frequency than both the LSE algorithm and the least absolute shrinkage and selection operator (LASSO) algorithm.

Another neural network design with potential for use in reproducing sound fields is the complex-valued multilayer perceptron (cMLP)[7]. With this method the cost function, forward propagation and back propagation are defined in terms of complex values and are calculated using the Wirtinger calculus[8]. The premise of this approach is to preserve phase information, which can be beneficial for signal reconstruction[9]. With this in mind, this paper outlines the use of a cMLP for sound field reproduction and compares its performance to the LSE algorithm. The structure of the rest of the paper is as follows. The theory defining the LSE algorithm and cMLP is outlined in the next section, followed by the methodology used to evaluate the methods, and the particular parameters of the neural networks used. Next, the results of the various tests carried out are presented and discussed. Finally, the conclusions of the research are presented.

This paper will demonstrate how a neural network can overcome some of the effects of spatial aliasing. Specifically, this is accomplished by first training the network using a large number of sensors (microphones) to sample the sound field. These are used to update the neural network using a back propagation algorithm that accounts for the relationship between the network output and the microphone signals. It is then shown that good results can be produced above the spatial aliasing frequency associated with a smaller number of sensor signals, which were used as inputs to the network.

2 THEORY

2.1 Least Mean Squared Error Optimisation

In this method, a cost function is defined from the mean squared error of a vector of error signals calculated at sensor locations. Figure 1 shows the block diagram outlining the definition of these errors.

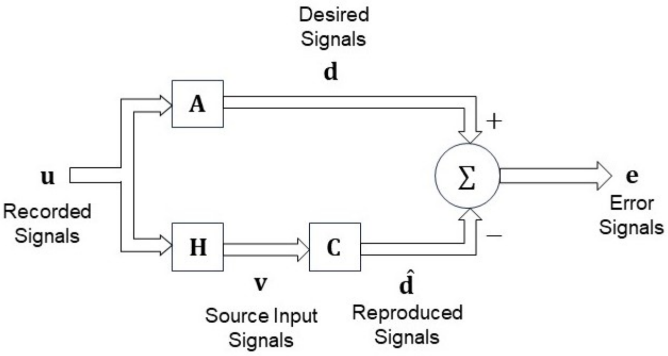

Figure 1 – Block diagram representing formulation of error signals from a target and reproduced pressure field.

In this instance $\mathbf{A}$, $\mathbf{H}$ and $\mathbf{C}$ are matrices of transfer functions. $\mathbf{C}$ represents acoustic propagation between the sources and receivers. $\mathbf{H}$ represents processing which is applied to recorded signals to determine the source strengths. For single frequency sound fields $\mathbf{A}$ is an identity matrix. From the block diagram, the cost function $J(\omega)$ is defined as

$$J(\omega) = \mathbf{e}^{\mathsf{H}}(\omega)\mathbf{e}(\omega) + \beta\mathbf{v}^{\mathsf{H}}(\omega)\mathbf{v}(\omega) \tag{1}$$

where β is a regularisation parameter and H denotes the conjugate transpose. The minimum of the cost function $J_0(\omega)$ is defined as

$$J_0(\omega) = \mathbf{d}^{\mathsf{H}}[1 - \mathbf{C}(\omega)[\mathbf{C}^{\mathsf{H}}(\omega)\mathbf{C}(\omega) + \beta\mathbf{I}]^{-1}\mathbf{C}^{\mathsf{H}}(\omega)]\mathbf{d}(\omega), \tag{2}$$

and the optimal source strengths $\mathbf{v}_0(\omega)$ are defined as

$$\mathbf{v}_0(\omega) = [\mathbf{C}^{\mathsf{H}}(\omega)\mathbf{C}(\omega) + \beta\mathbf{I}]^{-1}\mathbf{C}^{\mathsf{H}}(\omega)\mathbf{d}(\omega). \tag{3}$$

Sources driven with these source strengths will therefore result in the minimum mean squared error at the sensor locations.

2.2 Complex-Valued Multilayer Perceptron

The neural network used was the complex-valued multilayer perceptron developed by Paul and Nelson[10]. Figure 2 shows the architecture of the cMLP. In the figure, I is the number of sensors used to generate input signals, J is the size of the hidden layer, K is the number of sources and M is the number of sensors used to generate the output error signals. The layers are numbered with (2) being the input layer and (1) being the hidden layer. The general steps involved in calculating forward and back propagation are the same as those described by Paul and Nelson[10], with the difference being that the acoustic propagation matrix $\mathbf{C}$ must be included in the calculations.

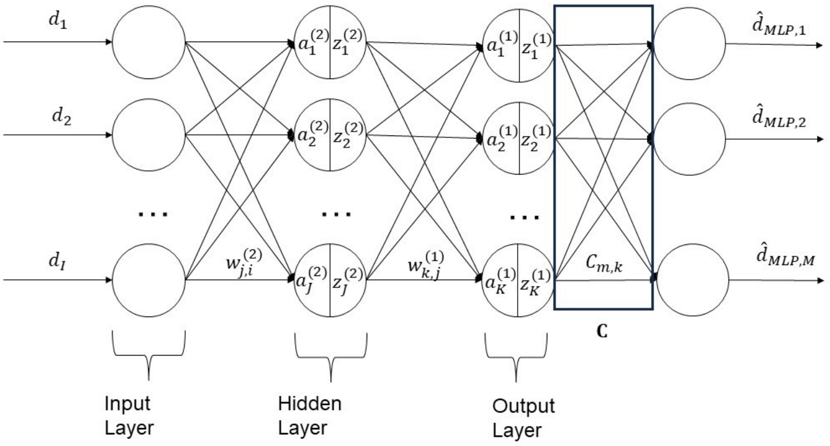

Figure 2 – Architecture of multilayer perceptron. Block $\mathbf{C}$ represents acoustic propagation.

With $\mathbf{C}$ added to the output layer of the network, the cost function J is in terms of the complex output vector from the network $\hat{\mathbf{d}}_{\boldsymbol{mlp}}$, given by

$$\hat{\mathbf{d}}_{\boldsymbol{mlp}} = \mathbf{C}\mathbf{z}^{(1)}. \tag{4}$$

In order to update the weights in the network through back propagation, it is necessary to compute the gradient of the loss function with respect to each of the weights in the neural network. Full details of this process are given by Paul and Nelson[10]. With the inclusion of the acoustic propagation matrix $\mathbf{C}$, this process requires some additional steps. For example, the gradient of the cost function with respect to the weights in the matrix $\mathbf{W}^{(1)}$ is calculated from

$$\frac{\partial J}{\partial \mathbf{w}^{(1)*}} = \frac{\partial J}{\partial \mathbf{z}^{(1)}} \cdot \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)*}} + \frac{\partial J}{\partial \mathbf{z}^{(1)*}} \cdot \frac{\partial \mathbf{z}^{(1)*}}{\partial \mathbf{w}^{(1)*}} \tag{5}$$

where $\mathbf{w}^{(1)}$ is the vector of values that includes the columns of the matrix $\mathbf{W}^{(1)}$ such that $\mathbf{w}^{(1)} =$ vec($\mathbf{W}^{(1)}$). Because the cost function J is a function of $\hat{\mathbf{d}}_{\boldsymbol{mlp}}$, Equation (5) should also be re-written in terms of $\hat{\mathbf{d}}_{\boldsymbol{mlp}}$, which results in

$$\frac{\partial J}{\partial \mathbf{w}^{(1)*}} = \frac{\partial J}{\partial \hat{\mathbf{d}}_{\boldsymbol{mlp}}} \cdot \frac{\partial \hat{\mathbf{d}}_{\boldsymbol{mlp}}}{\partial \mathbf{z}^{(1)}} \cdot \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)*}} + \left(\frac{\partial J}{\partial \hat{\mathbf{d}}_{\boldsymbol{mlp}}} \cdot \frac{\partial \hat{\mathbf{d}}_{\boldsymbol{mlp}}}{\partial \mathbf{z}^{(1)}} \right)^* \cdot \frac{\partial \mathbf{z}^{(1)*}}{\partial \mathbf{w}^{(1)*}}. \tag{6}$$

From Equation (4) it is apparent that $\frac{\partial \hat{\mathbf{d}}_{\boldsymbol{mlp}}}{\partial \mathbf{z}^{(1)}} = \mathbf{C}$, meaning that Equation (6) can be written as

$$\frac{\partial J}{\partial \mathbf{w}^{(1)*}} = \left(\frac{\partial J}{\partial \hat{\mathbf{d}}_{\boldsymbol{mlp}}} \cdot \mathbf{C} \right) \cdot \frac{\partial \mathbf{z}^{(1)}}{\partial \mathbf{w}^{(1)*}} + \left(\frac{\partial J}{\partial \hat{\mathbf{d}}^*_{\boldsymbol{mlp}}} \cdot \mathbf{C}^* \right) \cdot \frac{\partial \mathbf{z}^{(1)*}}{\partial \mathbf{w}^{(1)*}}. \tag{7}$$

Using this formulation in the derivation outlined by Paul and Nelson[10] results in

$$\frac{\partial J}{\partial \mathbf{W}^{(l)}} = [\mathbf{0} \quad \mathbf{I}] \left[\widetilde{\mathbf{d}} \quad \widetilde{\mathbf{B}}^{(l)}\right]^{\mathsf{T}} \mathbf{z}^{(l+1)\mathsf{H}}, \tag{8}$$

where the vector $\widetilde{\mathbf{d}}$ is defined in this case as

$$\widetilde{\mathbf{d}} = \left[(\hat{\mathbf{d}}_{\boldsymbol{mlp}} - \mathbf{d})^{\mathsf{H}}\mathbf{C} \quad (\hat{\mathbf{d}}_{\boldsymbol{mlp}} - \mathbf{d})^{\mathsf{T}}\mathbf{C}^{*}\right]. \tag{9}$$

The matrix $\widetilde{\mathbf{B}}^{(l)}$ is the product of the compound matrix of gradients ($\widetilde{\mathbf{H}}^{(l)}$) and the compound matrices of weights ($\widetilde{\mathbf{W}}^{(l)}$), which are defined in detail by Paul and Nelson[10]. Importantly, $\widetilde{\mathbf{B}}^{(l)}$ does not include the matrix $\mathbf{C}$, meaning that the derivation of the equations for back propagation is identical to that outlined by Paul and Nelson[10], and so is not included here. Note that the derivation outlined above relies on the holomorphic nature of Equation (4). For a non-holomorphic function, additional cross terms would need to be computed, as discussed by Paul and Nelson[10].

3 METHODOLOGY

3.1 Sound Field Simulation

The simulation consisted of an array of 4 sources located at the corners of a 4m x 4m square and an array of sensors forming a regular grid over the central 1m x 1m square (target area). The number of sensors in the grid varied depending on the numerical experiment. The acoustic pressure at each sensor was calculated from the source strengths using the monopole source model, assuming free field conditions. The complex pressure at each sensor $p(\boldsymbol{r})$ was therefore given by

$$p(\mathbf{r}) = \sum_{k=1}^{K} \frac{\rho_0 v_k}{4\pi|\mathbf{r}_k - \mathbf{r}|} e^{-jk|\mathbf{r}_k - \mathbf{r}|} \tag{10}$$

where v_k and $\mathbf{r}_k$ are the source strength (volume acceleration) and position of each source. In addition to the sensor locations, the acoustic pressure was calculated at 2500 evenly spaced points within the target area. The average mean squared error over these calculation points was used as a measure of performance of the reproduction methods. Figure 3 shows the location of the 4 sources and the target area over which the sensors and error calculation points were distributed.

At each frequency, target sound fields consisted of 100 single frequency plane waves, with angles of incidence equally spaced between $-180°$ and $180°$.

Optimal source strengths for each target plane wave, for the source and sensor geometry outlined above, were calculated using Equation (3). A regularisation value of $\beta = 0.1$ was used to prevent large values of the condition number of the inverse term $(\mathbf{C}^{\mathsf{H}}\mathbf{C} + \beta\mathbf{I})^{-1}$. This was particularly important at multiples of the spatial aliasing frequency. The source strengths derived with this method were then included in the simulation for comparison with the cMLP.

3.2 Complex-Valued Multilayer Perceptron

The neural network consisted of a hidden layer size of 200, a batch size of 4 and was trained over 100 iterations. The batch size was selected to prevent convergence to a local minimum. A complex cardioid activation function was used in the hidden layer and a hyperbolic tangent activation function was used in the output layer. Training was carried out using 80% of the target sound fields, selected at random from the target plane waves. The remaining 20% of the data was used for validation. The

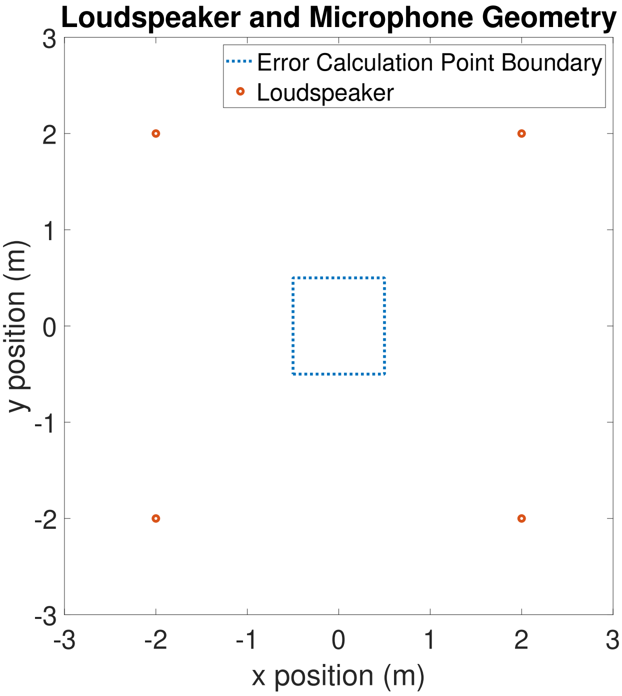

Figure 3 – Arrangement of sources and area covered by error calculation points.

neural network was trained using one-third octave centre frequencies from 125 Hz to 4 kHz, in order to test performance at frequencies ranging from below to above the spatial aliasing frequency. From this training a vector of source strengths was extracted from the network. These could then be compared to the source strengths derived from the LSE algorithm using the sound field simulation.

Different input and output layer sizes were utilized throughout testing. In the first test, the neural network consisted of an input size of 16 and an output size of 16. In this case, the neural network was supplied with the same information as the LSE algorithm using 16 sensors. In the second test, the neural network had an input size of 4 and an output size of 16. Therefore, the network updated based on error signals using 16 sensors, but was only supplied with 4 sensor signals from which to calculate source strengths. A final test was carried out with an input size of 4 and an output size of 36.

4 RESULTS

4.1 Neural Network with 16 Inputs and 16 Outputs

In this case the neural network was supplied with the same 16 microphone signals as the LSE algorithm, allowing a direct comparison of the two methods. Figure 4 shows the mean squared training and reproduction error of the neural network and LSE algorithm. The error is shown across frequency and at each frequency is averaged across all 100 angles of incidence and across the 2500 error calculation points.

The similar performance of the neural network and LSE algorithm is evident from Figure 4, with only a small variation in mean squared error between the methods occurring at low frequencies (<150 Hz). Note that the spatial aliasing frequency of this sensor arrangement was 514.5 Hz, and above this frequency the error of both methods is large and relatively consistent. Furthermore, the validation error of the neural network was close to the training error, demonstrating that the network was able to reproduce plane waves from directions that were not within the training set.

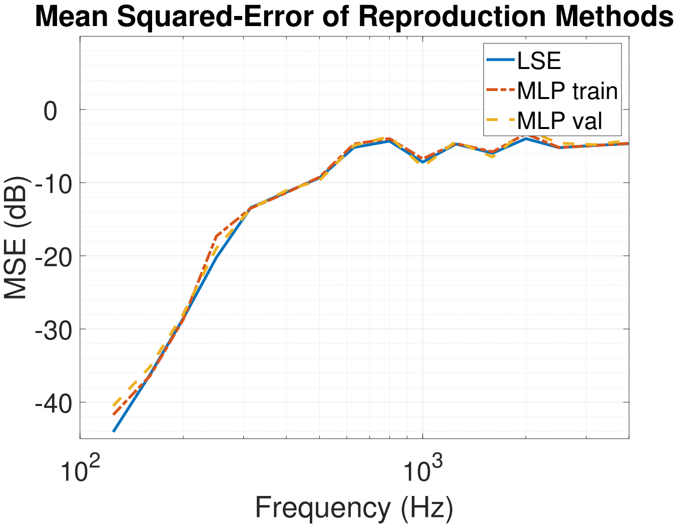

Figure 4 – Averaged mean squared error across all angles of incidence and error calculation points of cMLP with 16 inputs and 16 outputs (training and validation data) and LSE with 16 inputs.

4.2 Neural Network with 4 Inputs and 16 Outputs

In this test, the neural network was supplied with 16 microphone signals in the output layer and 4 signals in the input layer. This meant that the network was trained using 16 error signals at each frequency, but calculated source strengths based on only the 4 microphone signals corresponding to the corners of the sensor grid. The performance of this method can be demonstrated by comparison to the mean squared error of the LSE algorithm using all 16 sensors and using only the 4 corner sensors. The mean squared error of the two methods, across angles of incidence and the 2500 error calculation points, is shown in Figure 5

Figure 5 demonstrates how the performance of the neural network in this case is closer to the LSE with 16 sensors than the LSE with 4 sensors. In particular, the error of the neural network is largest at the spatial aliasing frequency of the 16 sensor arrangement, rather than that of the 4 sensors.

The improved performance of the neural network when compared to the LSE algorithm with 4 sensors can also be demonstrated by plotting the acoustic pressure across the target area of space. Figure 6 shows the real part of the complex pressure across the 1 x 1 m target area. This pressure is shown for the target pressure field, the cMLP using 4 inputs and 16 outputs, and the LSE using 4 sensors. The performance of the methods is comparable at 125 Hz and at 1 kHz because the frequencies are respectively below and above the spatial aliasing frequencies of both sensor arrays. However, at 315 Hz the neural network produces markedly better results.

4.3 Neural Network with 4 Inputs and 36 Outputs

This final case demonstrates how the neural network performs when there is a large difference in the density of sensors between the input and output layer. Figure 7 shows the mean squared error of the LSE algorithm with 4 sensors and with 16 sensors, and the error of the neural network with 4 inputs and 36 outputs. As before, the error is averaged across the 100 angles of incidence and the 2500 error calculation points.

Figure 7 demonstrates that the performance of the neural network is limited by the small number of input signals, as the error is not substantially reduced when compared to Figure 5, despite the increased number of sensors in the output layer.

5 CONCLUSION

This paper has shown how the performance of a cMLP compares to that of the LSE algorithm in sound field reproduction. When using identical error signals, the performance of the neural network is close to that of the LSE algorithm. The neural network can be trained using a larger output layer than input layer, which allows effective reproduction at frequencies above the spatial aliasing frequency of the input layer. This neural network builds on that outlined by Paul and Nelson[10] through the inclusion of the matrix $\mathbf{C}$, representing acoustic propagation beyond the output layer of the network.

Further work should investigate whether the neural network is capable of obeying linear superposition. This would involve training the neural network on sound fields containing multiple plane waves, and comparing the result to the corresponding sum of reproduced single plane wave sound fields. The ability to follow linear superposition would allow reproduction of more complicated sound fields using the cMLP.

This work was carried out in simulation and not in real-time. Further work could investigate both the effectiveness of the neural network with real sources and receivers, and whether the approach can be adapted to real-time processing. The matrix $\mathbf{C}$ included in the neural network was identical to the source model used in the simulation, which would not be the case in practical applications. This was also true of the source model used in the LSE calculations.

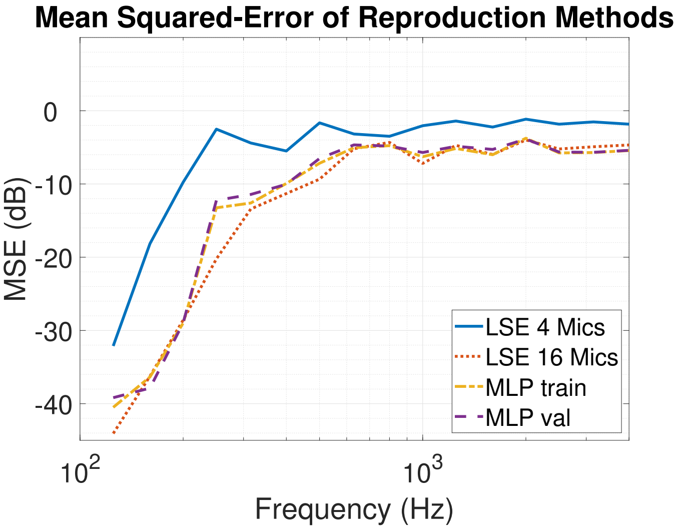

Figure 5 – Averaged mean squared error across all angles of incidence and error calculation points of cMLP with 4 inputs and 16 outputs (training and validation data) and LSE with 4 inputs and with 16 inputs.

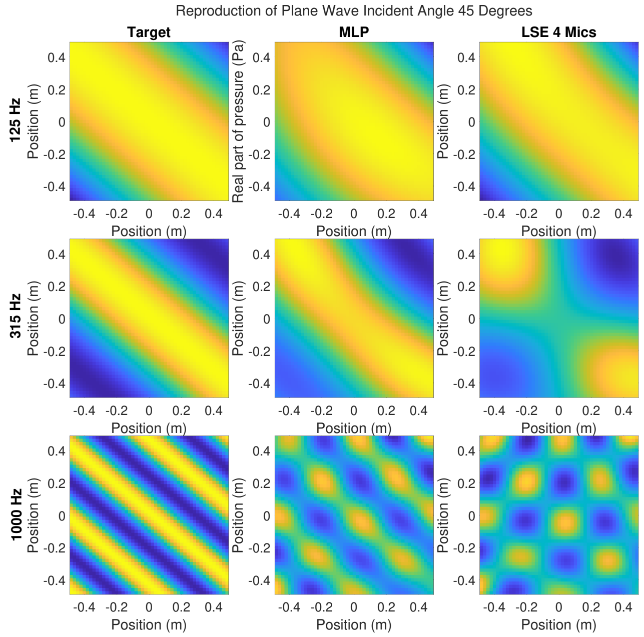

Figure 6 – Real part of complex pressure in measurement area given by target plane wave (first column), cMLP with 4 inputs and 16 outputs (second column), and LSE with 4 inputs (third column) when reproducing an incident wave at $45°$.

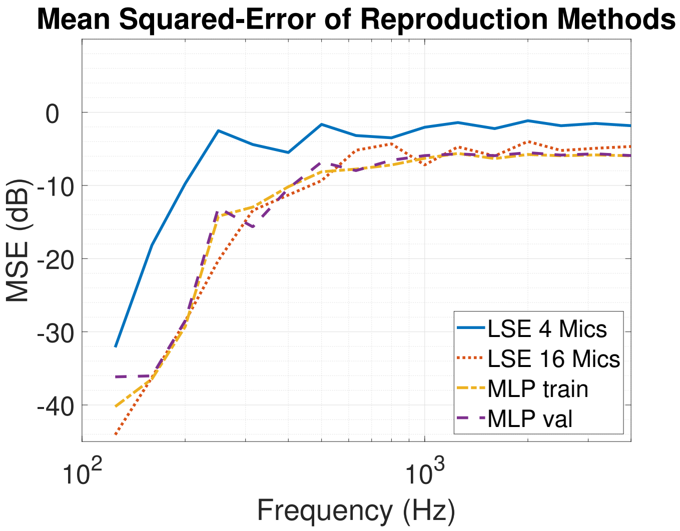

Figure 7 – Averaged mean squared error across all angles of incidence and error calculation points of cMLP with 4 inputs and 36 outputs (training and validation data) and LSE with 4 inputs and with 16 inputs.

6 REFERENCES

1. H. Khalilian, I. V. O. Bajić, and R. Vaughan. A simulation of a three-dimensional sound field reproduction system for immersive communication. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 93(5):980–995, 2017.

2. P. A. Nelson. Active control of acoustic fields and the reproduction of sound. *Journal of Sound and Vibration*, 177(4):447–477, 1994.

3. O. Kirkeby, P. A. Nelson, F. Orduna-Bustamente, and H. Hamafa. Local sound field reproduction using digital signal processing. *The Journal of the Acoustical SOciety of America*, 100(3):1584–1593, 1996.

4. A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *Journal of the Acoustic Society of America*, 177:2764–2778, 1993.

5. L. Comanducci, F. Antonacci, and A. Sarti. Synthesis of soundfields through irregular loudspeaker arrays based on convolutional neural networks. *Journal on Speech and Music Processing*, (1):17, 2024.

6. X. Hong, B. Du, S. Yang, and X. Lei, M. Zeng. End-to-end sound field reproduction based on deep learning. *The Journal of the Acoustical SOciety of America*, 153(5):3055, 2023.

7. V. S. Paul and P. A. Nelson. Complex valued neural networks for audio signal processing. *Reproduced Sound 2021*, 43, 2021.

8. W. Wirtinger. Zur formalen theorie der funktionen von mehr komplexen veränderlichen. *Mathematische Annalen*, 97(1), 1927.

9. A. V. Oppenheim and J. S. Lim. The importance of phase in signals. *Proceedings of the IEE*, 69(5):529–541, 1981.

10. V. S. Paul and P. A. Nelson. Efficient design of complex-valued neural networks with application to the classification of transient acoustic signals. *The Journal of the Acoustical Society of America*, 156(2), 2024.