SOUND FIELD ANALYSIS USING A COMPLEX-VALUED NEURAL NETWORK

Vlad S. Paul Institute of Sound and Vibration Research, University of Southampton, UK
Nara Hahn 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

Virtual sensing refers to the estimation of a sound field at locations where physical sensors are not present. For example, the sound field can be predicted around the listeners head, where direct measurements might be impractical or cumbersome. This is usually achieved using data from existing microphones and mathematical models that describe the acoustic wave propagation. Related signal processing algorithms have been developed for various applications such as active noise control (ANC)[1,2], sound field reconstruction[3,4] or beamforming[5,6]. More recently, machine learning approaches have been developed to estimate the sound field at virtual sensors[7,8]. In this paper[7], the authors use generative adversial networks (GANs) to reconstruct spatial room impulses in conventional rooms and show that they can use GANs to enhance the sound field reconstruction performance by recovering some of the energy at high frequencies. This information at high frequencies is usually lost or distorted, due to the Nyquist criterion that requires a fine spatial sampling of microphones to represent the rapid variations of the sound field that occur at high frequencies. Traditional methods based for example on spherical harmonic expansion struggle to accurately reconstruct sound fields at frequencies above this aliasing threshold due to spatial aliasing and truncation errors.

Various approaches have been developed throughout the years to increase the aliasing-free bandwidth using optimal spatial sampling strategies[9,10] or signal processing techniques such as compressive sensing[11]. More recently, neural network architectures have been shown to partly overcome spatial aliasing at high frequencies for different applications[7,12]. Compared to traditional methods, these authors show that machine learning models can achieve a higher sound field reconstruction performance by improving the recovery of high frequency information, that would otherwise be lost.

In this work, a simple complex-valued multilayer perceptron (cMLP) is used to learn statistically the spatio-spectral properties of single frequency plane waves from various directions recorded by a microphone array on the surface of an open sphere. The network model is trained to predict these sound fields at control points in a sphere of a smaller radius across a range of frequencies, including those above the aliasing frequency.

2 SOUND FIELD RECONSTRUCTION USING SPHERICAL HARMONICS

2.1 Problem description

In this work, the task considered is the prediction of the sound pressure at a set of control points densely distributed throughout a sphere of radius $r_{\mathrm{ctl}} = 0.1$ m, similar to the size of the human head. The available data from which the predictions are made consists of pressure computed using a larger open spherical microphone array with radius $r_{\mathrm{m}} = 0.25$ m. The open sphere microphone array was simulated using 32 sensors on the surface of the sphere positioned at the same locations as those of an Eigenmike em32[13]. Assuming a maximum spherical harmonic order of $N = 4$, that follows from the array's spatial sampling limitations, the aliasing frequency above which the array will struggle to

capture and represent accurately the sound field is given by

$$f_{\text{alias}} = \frac{cN}{2\pi r_{\text{mic}}} \approx 873 \text{ Hz}, \tag{1}$$

where $c = 343$ m/s is the speed of sound. Above f_{alias}, spatial aliasing occurs, which means that the spatial sampling of microphones on the sphere is insufficient to capture the spatial variations of the sound field at higher frequencies.

One of the main challenges of this task is the reconstruction of the sound field at frequencies above f_{alias}, where spatial aliasing degrades the reconstruction accuracy. In addition, as discussed in the next section, issues related to the zeros of the spherical Bessel functions at specific frequencies might introduce numerical instabilities. To address these challenges, a complex-valued network model is investigated as an alternative approach for sound field reconstruction.

2.2 Spherical harmonics-based sound field reconstruction

Assuming a source-free and homogeneous medium, the pressure $p(k, \mathbf{r})$ at a general point $\mathbf{r} = (r, \theta, \phi)$ can be computed using the spherical harmonic expansion of the solution to the Helmholtz equation[10] as:

$$p(k, r, \theta, \phi) = \sum_{n=0}^{N} \sum_{m=-n}^{n} p_{nm}\, j_n(kr) Y_n^m(\theta, \phi), \tag{2}$$

where n, m are the degree and order of the spherical harmonics, respectively, k denotes the frequency-dependent wavenumber and p_{nm} are the modal coefficients that are independent of r and represent the amplitude of each spherical harmonic mode in the sound field. The term $j_n(kr)$ is the spherical Bessel function of the first kind of order n evaluated at kr and $Y_n^m(\theta, \phi)$ are the spherical harmonic functions.

Using the pressure $p(k, r, \theta, \phi)$, one can compute the modal coefficients p_{nm} by first computing the spherical Fourier transform coefficients $\breve{p}_{nm}(k, r)$ which are radially dependent

$$\breve{p}_{nm}(k, r) = \int_0^{2\pi} \int_0^{\pi} p(k, r, \theta, \phi) Y_n^{m*}(\theta, \phi) \sin(\theta) d\theta d\phi \tag{3}$$

and then remove the radial dependency to isolate the modal coefficients $p_{nm} = \frac{\breve{p}_{nm}(k,r)}{j_n(kr)}$. Once the modal coefficients are known, the pressure at any other point $\mathbf{r}' = (r', \theta', \phi')$ can be reconstructed from

$$p(k, r', \theta', \phi') = \sum_{n=0}^{N} \sum_{m=-n}^{n} p_{nm}\, j_n(kr') Y_n^m(\theta', \phi'). \tag{4}$$

2.3 Limitations of the use of spherical harmonics

Using this approach, one can estimate the pressure at different points in the sound field, however several limitations occur. First, when the radial dependency is removed, the division by $j_n(kr)$ can lead to numerical instabilities when $j_n(kr) = 0$. The zeros in the spherical Bessel functions are frequency- and radius-dependent and can therefore impact the reconstruction of the pressure at certain frequencies for a fixed radius. A second limitation has been discussed above and is related to spatial aliasing. For example, if $p(k, r, \theta, \phi)$ contains distortions in the pressure at high frequencies, these distortions will affect the computation of $p(k, r', \theta', \phi')$ using the above approach.

2.4 Proposed solution using cMLP network models

The use of a cMLP is proposed to overcome some of the limitations discussed above. The implementation of the network model is discussed in detail in recent work[14] and will not be repeated here.

It is believed that the nonlinear behaviour of the MLP due to the activation functions has the potential to capture complex patterns and distortions introduced by spatial aliasing. Using a data-driven approach, the MLP model learns to predict the pressure at control points from the pressure at the microphones, without explicitly relying on the spherical harmonic expansion. Thus the network model avoids the numerical instabilities associated with the zeros of the spherical Bessel functions. In addition, the use of complex inputs and outputs allows the network to model both amplitude and phase information more efficiently, which is crucial for accurate sound field representation.

3 METHODOLOGY

3.1 Dataset generation

Single frequency plane waves are generated from 2000 uniformly distributed positions around the open sphere and their pressure is simulated at the microphones on the surface of the open sphere. For a general microphone position $\mathbf{r}_\mathrm{m}$ and a plane wave direction $\hat{\mathbf{r}}_\mathrm{s}$, the pressure at the microphone for a single frequency plane wave is computed from

$$p(k, \mathbf{r}_\mathrm{m}) = Ae^{-j(k\hat{\mathbf{r}}_\mathrm{s}.\mathbf{r}_\mathrm{m}+\alpha)}, \tag{5}$$

where A is the unit amplitude of the wave, $k = 2\pi f/c$ is the wavenumber that depends on the frequency f and $\hat{\mathbf{r}}_\mathrm{s}.\mathbf{r}_\mathrm{m}$ is the dot product between the two vectors. It should be noted that the vector $\hat{\mathbf{r}}_\mathrm{s}$ describes the direction of the plane wave propagation and has unit length. The additional phase term α denotes a random phase that was added to each plane wave in order to ensure that each plane wave arrives at a slightly different time at the microphone array. This way, each plane wave has a phase shift that does not depend on the distance to the microphone array and so the network will have to learn patterns that are not related to this distance.

Three different datasets of single frequency plane waves were generated based on f_alias. The first dataset contained plane waves of a frequency below the aliasing frequency ($f_\mathrm{low} = 436$ Hz), corresponding to one octave lower than f_alias. The second dataset used plane waves at the aliasing frequency and the third dataset was built with plane waves at $f_\mathrm{high} = 1746$ Hz, corresponding to one octave above f_alias.

The control points are generated inside a sphere of radius $r_\mathrm{ctl} = 0.1$ m, which is similar to the radius of a human head. A number of 958 sensors were used as control points and were uniformly distributed inside the sphere, spaced at a distance $d_\mathrm{ctl} \ll \lambda_\mathrm{min}/2$ from each other. Here, λ_min corresponds to the wavelength of the highest frequency f_high used to generate one of the datasets. Using such a spacing, the control points were close enough to avoid spatial aliasing, ensuring that the sound field at f_high can be reconstructed accurately. The spatial aliasing at the control points is different from that at the microphones on the open sphere and would occur during training in the target outputs. This is not desired, since the aim is to train the network to overcome the spatial aliasing of the microphone array by accurately reconstructing the sound field at the control points. Figure 1 shows the simulated setup with the control points, the microphones on the open sphere and a small number of plane wave directions of arrival (DOA) for illustration purposes. It should be noted that from the total number of 2000 plane wave locations, 20% of the data was used for testing and 80% for training.

The cMLP model was trained using the pressure of each individual plane wave recorded by the microphone array as input and the target output was the pressure recorded by the control points. The network model was trained to learn how to predict the sound field at control points, in order to be able to predict the sound field from plane wave directions that have not been seen during training. Figure 2 shows an example of the input and output data for one of the plane wave directions.

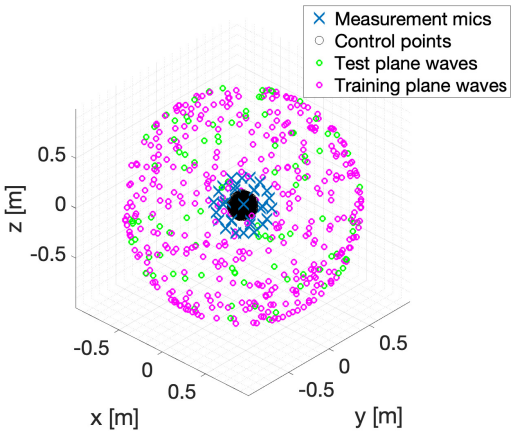

Figure 1 – Sound field reconstruction simulation setup.

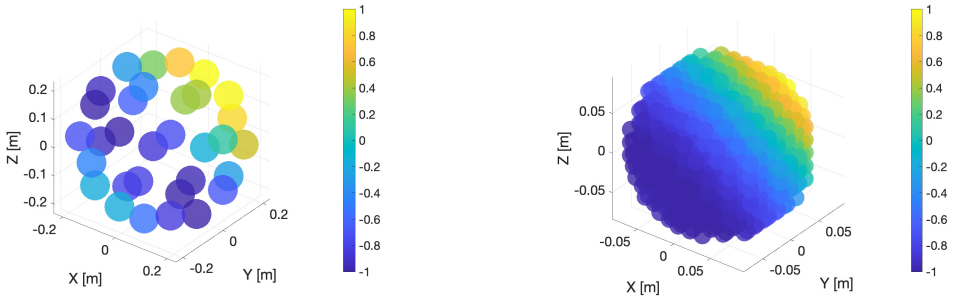

Figure 2 – Example of input data (left) and output data (right) for training a cMLP model, showing the real part of the pressures at the microphone array and at the control points for a plane wave at $f_{\text{low}} = 436$ Hz.

3.2 Network parameters

The cMLP model has 32 neurons as input layer, corresponding to the complex pressure recorded by the microphone array on the surface of the open sphere. Two hidden layers of 400 neurons each are used to design the model and the output layer consists of 958 neurons corresponding to the complex pressure recorded by the control points. Since all the data used by the cMLP model is complex, the complex cardioid[15] activation function was used in the hidden layers and the tanh function was used in the output layer. The training was stopped after 50 iterations and a learning rate of 0.001 was used to update the weights based on the ADAM optimizer[16].

4 RESULTS

The performance of the cMLP is evaluated for the three different datasets using three different test scenarios. For the first case, the model trained on single plane waves will be evaluated on how well it can predict the reconstruction of plane waves from the test dataset. For the second scenario, the same trained model will be tested on how well it can reconstruct a plane wave, if noise is added to the pressure recorded by the open sphere microphone array. White noise signals having an SNR value of 20 dB are added to the input plane wave and the estimated output is compared to the target clean plane wave. The objective is to evaluate if the model trained only on reconstructing clean plane waves is able to work with noisy plane waves. For the last scenario, two plane waves are added together as input into the trained model and its estimated output is compared to the actual sum of the two plane waves at the control points. This way, an initial attempt is made to understand if the network model can interpret the superposition of two plane waves, even if it was trained on single plane waves. The performance of the cMLP model will be compared to the benchmark technique explained in Section 2.2 for the first and

second scenario. For visualisation purposes, the plots showing the reconstructed pressures will be shown only for the horizontal plane of control points, corresponding to 115 microphones.

4.1 Scenario 1: Noise-free plane waves

Figure 3 shows the comparison between the target and the estimated pressures of one plane wave arriving from $(\theta, \phi) = (112°, 43°)$ at all three frequencies. The location of the plane wave was part of the test dataset. As expected, the lower the frequency of the plane wave is, the higher the reconstruction

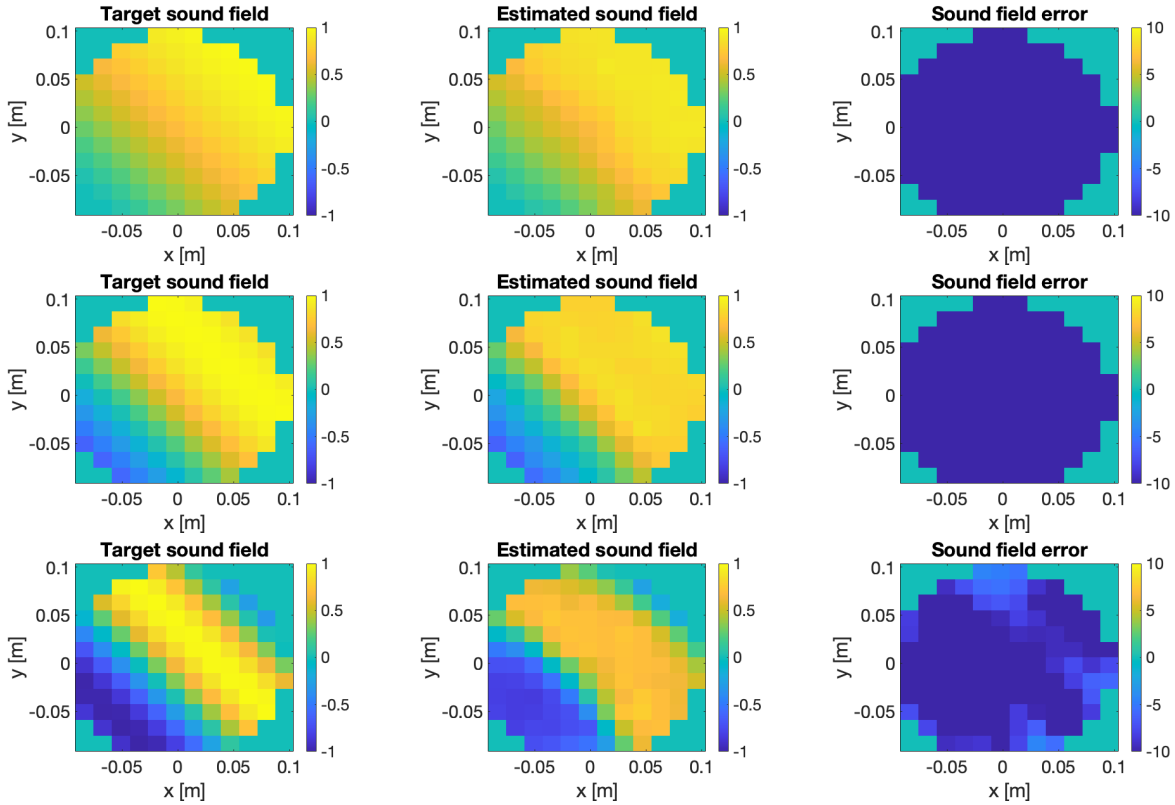

Figure 3 – Example of reconstructed sound fields using the cMLP model. Top figures: $f_{\text{low}} = 436$ Hz; middle figures: $f_{\text{alias}} = 873$ Hz; bottom figures: $f_{\text{high}} = 1746$ Hz. The errors are shown in dB.

accuracy becomes. In terms of averaged reconstruction error for the horizontal plane of control points, the cMLP achieves a reconstruction error of -25.8 dB for the lowest frequency, -17.9 dB for the aliasing frequency and -10 dB for the highest frequency. The comparison with the other two benchmark techniques is shown in the next section in Figure 5, where the averaged reconstruction error over all 958 control points is compared to that resulting from the noisy plane waves.

4.2 Scenario 2: Noisy plane waves

Using the same plane wave direction $(\theta, \phi) = (112°, 43°)$, Figure 4 shows the estimated sound field of the cMLP model if a noisy plane wave having a frequency of $f_{\text{high}} = 1748$ Hz is used as input. Compared to the noise-free case, the averaged reconstruction error in this case increases by approximately 1 dB, suggesting that even if the input plane wave is slightly altered by adding 20 dB SNR white noise, the network is able to estimate a sound field that is close to the target sound field, the reconstruction error increasing by only 1 dB.

Figure 5 shows a comparison between the cMLP model and the benchmark technique in terms of reconstruction error for all three frequencies investigated when estimating clean plane waves (∞ dB SNR) and noisy plane waves (20 dB SNR). It can be observed that the performance of the benchmark

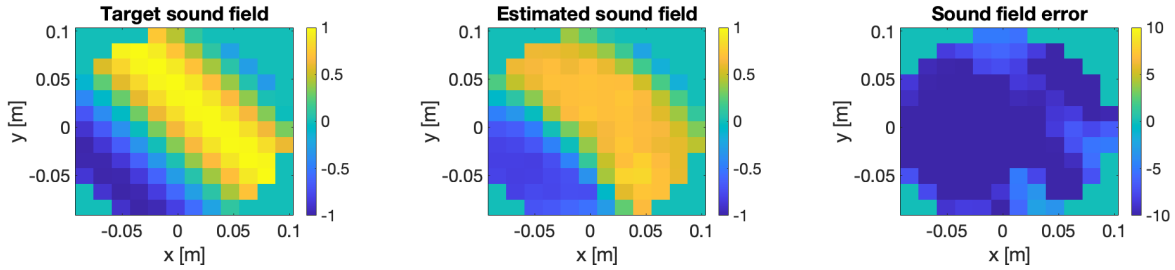

Figure 4 – Example of a reconstructed sound field at control points using the cMLP model. The input plane wave has 20 dB SNR added white noise. The errors are shown in dB.

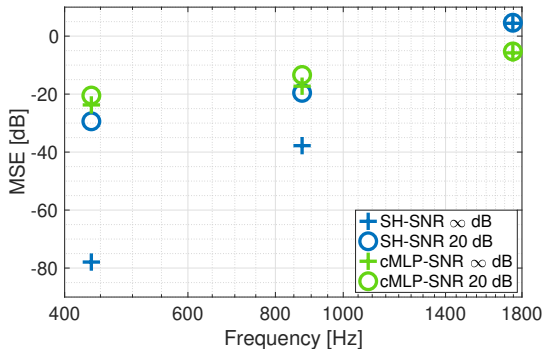

Figure 5 – Comparison of sound field reconstruction error between the cMLP performance and the benchmark technique (SH) when reconstructing noise-free and noisy plane waves at frequencies $f_{\text{low}} = 436$ Hz, $f_{\text{alias}} = 873$ and $f_{\text{high}} = 1746$ Hz.

technique (SH) is much better than that of the cMLP model at f_{low} and f_{alias} for the noise-free scenario. However, the cMLP model outperforms the benchmark technique at the frequency above the aliasing frequency. Also, it is interesting to see that the benchmark technique is more sensitive to noisy plane waves, while the cMLP model decreases the reconstruction performance by only few decibels.

4.3 Scenario 3: Superposition of plane waves

For the final scenario, two single frequency plane waves from different directions ($\theta_1 = 112°$,$\phi_1 = 214°$) and ($\theta_2 = 143°$,$\phi_2 = 181°$) are added together and sent as input through the trained cMLP model. The target sound field is the superposition of the two plane waves at the control points. Figure 6 shows the estimated sound field of the cMLP model for the superposed plane waves at $f_{\text{high}} = 1746$ Hz. Compared to the previous two cases, here the sound field reconstruction error is much higher, which

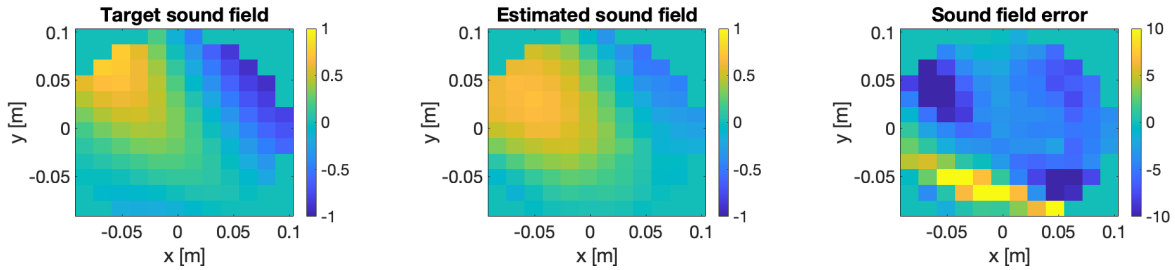

Figure 6 – Example of a reconstructed sound field at control points using the cMLP model. The input into the trained network is the superposition of two plane waves. The errors are shown in dB.

is not surprising. The network model has been trained only using single plane waves, however when

a superposition of two plane waves is sent as input into the network, the estimated output is not completely different compared to the target sound field. Some correct patterns can be observed in the networks' estimate, but further investigations are needed in order to have conclusive results.

4.4 Discussion

The results above show that a relatively small network model (approx. 550,000 parameters) can be trained to reconstruct the sound field of single frequency plane waves at some defined control points. At frequencies below the aliasing frequency, the benchmark technique outperforms the cMLP model by a significant amount if the plane waves are noise-free. One of the reasons for this large difference at f_{low} and f_{alias} is the training behaviour of the network. As shown in Figure 7, the network was still improving after 50 iterations, suggesting that if trained further, the performance of the models could have been improved.

Another interesting observation is the fact that the trained network is less sensitive to noisy plane waves compared to the benchmark technique, even if it was trained on clean noisy-free data. Looking at the benchmark technique, the difference in reconstruction error between the clean and noisy plane waves at $f_{\mathrm{low}}, f_{\mathrm{alias}}$ is very large, while in the cMLP case, the difference is only a few decibels.

The most important observation that follows from these simulations is that the cMLP is able to outperform the benchmark technique above aliasing frequency. Similar observations have been made for different applications by [7,12]. In the case presented here, the cMLP model learns complex mappings between the distorted input data due to spatial aliasing and the aliasing free data at the control points. This way, the network learns to compensate for the distortions and missing information caused by aliasing. These observations are very promising, suggesting that one could use machine learning models to improve the analysis or reconstruction of sound fields above the aliasing frequency limit.

Regarding the superposition of two plane waves, the initial results presented here are promising. However, further work is needed to understand if the network model is able to learn the characteristics of the different single plane waves in such a way that it can understand the superposition of two plane waves. A more comprehensive analysis that is focused only on this topic would be useful.

Figure 7 – Training behaviour of the cMLP model for the two datasets: $f_{\mathrm{low}} = 436$ Hz (left) and $f_{\mathrm{alias}} = 873$ Hz (right).

5 CONCLUSION

This work showed that a complex-valued neural network can be trained to predict a sound field at a set of control points and can achieve a better reconstruction accuracy than the benchmark at frequencies above the aliasing frequency. The performance of the network was tested in different scenarios and results showed considerable potential for such an approach to be applied to more complex tasks. The authors plan to investigate the features the network is learning from the input pressures through deeper analysis of the weight matrices and if the trained model at a single frequency plane wave can be applied to other closely related frequencies.

6 REFERENCES

1. J. Zhang, S. J. Elliott, and J. Cheer. Robust performance of virtual sensing methods for active noise control. *Mechanical Systems and Signal Processing*, 152:107453, 2021.

2. C. Shi, R. Xie, N. Jiang, H. Li, and Y. Kajikawa. Selective virtual sensing technique for multi-channel feedforward active noise control systems. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8489–8493. IEEE, 2019.

3. T. Nishida, N. Ueno, S. Koyama, and H. Saruwatari. Region-restricted sensor placement based on gaussian process for sound field estimation. *IEEE Transactions on Signal Processing*, 70:1718–1733, 2022.

4. S. Damiano, F. Borra, A. Bernardini, F. Antonacci, and A. Sarti. Soundfield reconstruction in reverberant rooms based on compressive sensing and image-source models of early reflections. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 366–370. IEEE, 2021.

5. Y. Gu, C. Zhou, N. A. Goodman, W.-Z. Song, and Z. Shi. Coprime array adaptive beamforming based on compressive sensing virtual array signal. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2981–2985. IEEE, 2016.

6. D. Fernandez Comesana, K. R. Holland, J. Wind, and H.-E. de Bree. Adapting beamforming techniques for virtual sensor arrays. 2012.

7. E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft. Generative models for sound field reconstruction. *The Journal of the Acoustical Society of America*, 153(2):1179–1190, 2023.

8. F. Ma, T. D. Abhayapala, and P. N. Samarasinghe. Circumvent spherical bessel function nulls for open sphere microphone arrays with physics informed neural network. In *Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum*, pages 2169–2175, 2023.

9. P. Pal and P. P. Vaidyanathan. Coprime sampling and the MUSIC algorithm. In *Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, pages 289–294, 2011. doi: 10.1109/DSP-SPE.2011.5739227.

10. B. Rafaely. *Fundamentals of spherical array processing*, volume 8. Springer, 2015.

11. E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

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

13. mh acoustics. *em32 Eigenmike release notes*. 25 Summit Ave Summit, NJ 07901, April 2013. Available at https://mhacoustics.com/sites/default/files/EigenmikeReleaseNotesV15.pdf.

14. 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):1099–1110, 08 2024. ISSN 0001-4966. doi: 10.1121/10.0028230.

15. P. Virtue, S. X. Yu, and M. Lustig. Better than real: Complex-valued neural nets for mri fingerprinting. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 3953–3957, 2017. doi: 10.1109/ICIP.2017.8297024.

16. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. doi: 10.48550/arXiv.1412.6980.