DATA-DRIVEN APPROACH FOR ACOUSTIC SOURCE LOCALIZATION

Arnav Joshi[*1,2], **Hamid Daryan**[2], **and Jean-Pierre Hickey**[2]
[1]Indian Institute of Technology Indore, India
[2]Department of Mechanical and Mechatronics Engineering, University of Waterloo, Canada

1 Introduction

Determining the strength and location of acoustic sources is crucial for studying aeroacoustic noise generation such as in vortical flows [1]. A robust technique is needed for the accurate mapping of these sources. The conventional acoustic beamforming method [2] is reliable but has limitations, suffering from spatial aliasing and poor resolution at lower source frequencies. Deep learning algorithms have emerged as powerful tools in a growing number of disciplines due to their ability to learn patterns and extract features from limited or unstructured data. Researchers have utilised deep learning to overcome the limitations of traditional methods. Xu et al [3] have used Densely connected neural networks (DNNs) for acoustic source imagine and while this paper employs largely the same methodology, it uses a Convolutional Neural Network (CNN) which is computationally less expensive, to determine the spatial and temporal characteristics of stationary and moving sound sources [4]. A training database was developed using analytically-defined monopoles which were randomly distributed over a scanning grid of fixed size. A 64-channel microphone array was simulated in a plane parallel to the plane of the scanning grid to gather information about the sources in the form of the Cross-Spectral Matrix (CSM) which was then used as an input feature to the CNN. The results showed that the CNN model was able to identify position, strength, and velocity of the sources over a range of frequencies with far better accuracy and resolution than the traditional methods.

2 Method

The proposed method is data-driven and hence a database containing enough data samples for training the CNN has to be generated first. The CNN model is then explained in detail.

2.1 Data Generation

The scanning grid is a 1.2m x 1.2m area divided into an NxN grid. The computational power and training requirement increases as the resolution of the scanning grid increases. It contains S sources distributed randomly across the N^2 grid points. The microphone array plane contains M microphones arranged in the shape of a logarithmic spiral (M=64 in this case) and is located 1.2m below the scanning grid. The logarithmic spiral arrangement was chosen to ensure good performance over a range of frequencies. The sound sources are modelled as monopoles and are assumed to radiate spherical pressure signals. Fast fourier transform has been applied on the signal to convert it from time domain to frequency domain. The pressure signal from a source s on the scanning grid to a microphone m on the array plane [3] is given as

$$P_s(m) = \frac{e^{-j2\pi r_s/c_0}}{4\pi|r_s|} \quad (1)$$

where r_s is the distance between the particular source s and the microphone m and c_0 is the speed of sound in air which is 343 m/s. Pressure signals from every source are added at every microphone to generate the pressure vector **P** given as

$$\mathbf{P} = [\sum_{s=1}^{S} P_s(1), \sum_{s=1}^{S} P_s(2), ..., \sum_{s=1}^{S} P_s(M)] (2)$$

Vector **P** has dimensions Mx1. The Cross-Spectral Matrix (CSM) is defined as

$$\mathbf{CSM} = \mathbf{P}\mathbf{P}^H (3)$$

where P^H is the complex conjugate of the pressure vector. The Ground Truth Matrix (GTM) contains the actual source position data. Sources are positioned randomly within the matrix and the entries that have a source are assigned the source strength values. The remaining entries (where there is no source) are assigned the value zero. The CSM is an MxM matrix and will be used as an input to the CNN. The network will be trained against the GTM which has the same dimensions as that of the scanning grid. A training sample consists of the CSM obtained from the random positioning of the sources within the GTM and, the GTM.

2.2 Convolutional Neural Network

An artificial neural network is a simulation of the biological brain composed of artificial neurons or nodes. These nodes make up layers which are interconnected to progressively extract features and learn from the data being fed to the network. The inputs to a neuron are assigned weights by the network. The activation function associated with the node calculates the output of the node based on the weighted sum of the inputs. During training, the network compares its prediction with the actual output through a loss function and modifies the weights accordingly until they reach the optimal values. Typically, an artificial neural network has an input layer, multiple hidden layers, and an output layer.

A Convolutional Neural Network (CNN) [5] is a type of neural network that finds its application extensively in image classification and segmentation. The network takes an image for its input. The convolution layer applies a series of filters to it that help the network capture the high-level features of the image. The pooling layer then reduces the dimensions of the image, preserving the dominant features and reducing the

*arnavjoshi.iiti.me@gmail.com

number of parameters and computational requirements. Once the convolution and pooling operations are done, the final image is flattened and fed to a regular neural network.
The input image or feature in this case is the Cross-Spectral Matrix which encapsulates the pressure signal data of the sound sources obtained by the microphone array. The network is trained against the ground truth which too is flattened before training thus converting it into an N^2x1 vector. The number of hidden layers can vary and while more number of hidden layers enable the model to learn better, care should be taken to avoid overfitting the data. The activation function used is Rectified Linear Unit (ReLU) which ouputs the input value if it is greater than or equal to zero, and zero otherwise. The optimizer is ADAM (derived from *Adaptive Moment Estimation*) which is a gradient-based optimization algorithm for updating the weights, and the loss function is mean squared error (mse) which is given as

$$mse = \frac{\sum_{i=1}^{N}(y_{pred} - y_{gt})^2}{N} \tag{4}$$

where y_{pred} is the predicted vector given by the network and y_{gt} is the ground truth vector.

3 Results

Various CNN models were developed, each of them trained to detect a fixed number of uniform sources at a particular frequency. The input source strength was taken to be 1 Pa. 50000 random samples were generated for training and 10000 for validation. The models were trained for around 100 epochs. Results for a particular case- 6 sources at 8000 Hz spread randomly over a 12x12 scanning grid- are shown in this paper as a representative of the general trend.

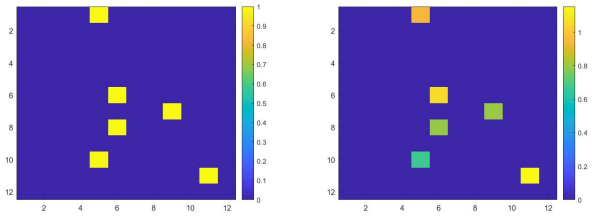

Figure 1: Ground Truth (left) and CNN prediction (right).

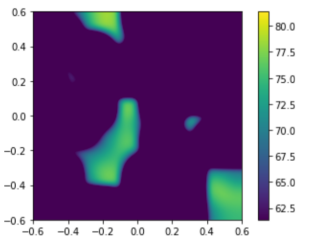

Figure 2: Beamforming Output (in decibels)

4 Discussion

Figure 1 shows that the model managed to locate all the 6 sources perfectly. The predictions were far more accurate and of much better resolution compared to the beamforming output (Figure 2) at higher frequencies, and even more so at lower frequencies where beamforming was rendered virtually unhelpful. The model can not only detect static sources but also track moving sources. Pressure data recorded at an instant through the microphone array can be used to predict the source's position at that instant from which the its velocity and acceleration can be extrapolated. Furthermore, models trained to detect random number of sources and sources with different strengths were also developed to explore the performance on more realistic scenarios.

5 Conclusions

The limitations of conventional beamforming at resolving complex source distributions, especially at lower frequencies, prompted the search for an alternative method that was more robust and accurate. A data-driven approach based on deep learning was employed. A Convolutional Neural Network trained to detect and track static and moving acoustic sources was developed. The Cross-Spectral Matrix containing the pressure data obtained by the microphone array was used an input to the network while it was trained against the Ground Truth Matrix. Multiple CNN models were developed to span a range of source frequencies and the performance was found to be far better than acoustic beamforming. To challenge the model further, it was trained for scenarios with a greater degree of randomness like detecting sources with different strengths or detecting a number of sources within a fixed range. There is still some scope for refinement but overall, the results show much promise and it is expected that with more data and training, a robust and generalized deep learning framework for detection of acoustic sources in real-life applications can be successfully built. This preliminary work will be extended to identify locations of acoustic sources in vortical and turbulent flows.

Acknowledgments

This research was enabled in part by support provided by Sharcnet and Compute Canada (www.computecanada.ca).

References

[1] Hamid Daryan, Fazle Hussain, and Jean-Pierre Hickey. Aeroacoustic noise generation due to vortex reconnection. *Phys. Rev. Fluids*, 5:062702, Jun 2020.

[2] Leandro de Santana. Fundamentals of acoustic beamforming. *Design and Operation of Aeroacoustic Wind Tunnel Tests for Group and Air Transport*, 2017.

[3] Pengwei Xu, Elias JG Arcondoulis, and Yu Liu. Acoustic source imaging using densely connected convolutional networks. *Mechanical Systems and Signal Processing*, 151:107370, 2021.

[4] Rémi Cousson, Quentin Leclere, Marie-Agnès Pallas, and Michel Berengier. Identification of acoustic moving sources using a time-domain method. In *bebec*, 2018.

[5] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. Ieee, 2017.